TESIS DOCTORAL

Reconocimiento robusto de voz con datos perdidos o inciertos

Universidad de Granada

Autor:
José Andrés González López

Directores:
Antonio Miguel Peinado Herreros
Ángel Manuel Gómez García

Doctorado en Tecnologías Multimedia

Departamento de Teoría de la Señal, Telemática y Comunicaciones

Granada, Febrero 2013

El doctorando **D. José Andrés González López** y los directores de la tesis **D. Antonio Miguel Peinado Herreros** y **D. Ángel Manuel Gómez García** garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de los directores de la tesis y, hasta donde nuestro conocimiento alcanza, en la realización del trabajo se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Granada, a 23 de enero de 2013

Director/es de la Tesis Doctorando

Fdo.: Antonio M. Peinado Fdo.: Ángel M. Gómez Fdo. José A. González

A mi padres, por todo su esfuerzo y sacrificio,
y a Nuria por su amor y comprensión.

Agradecimientos

Parece que fue ayer cuando empecé a recorrer, paso a paso, este sendero que me ha traído hasta donde hoy nos encontramos. Durante el camino muchas han sido las personas con las que me he cruzado, dejando su impronta en este trabajo. Las siguientes líneas son para todos vosotros.

En primer lugar, quisiera agradecer a Victoria Sánchez la confianza depositada en mí, porque sin aquel correo que me escribió al acabar Procesamiento de la Voz, hoy día no estaría aquí. Gracias a José Luis Carmona y Juan Andrés, por su amistad y compañerismo durante estos años, porque sin ellos este trayecto habría sido más difícil. No puedo olvidarme tampoco de la gente de Sheffield, en especial de Ning, Bi y Phil Green, por su cálida acogida durante mis estancias, haciéndome sentir como si estuviera en casa.

También quiero expresar mi más profundo y sincero agradecimiento a mis directores Antonio Peinado y Ángel Gómez, por su guía, apoyo y, en especial, por su cercanía humana. Esta tesis no sería lo que es de no haber sido por ellos.

Gracias a mi familia y amigos por el cariño que me han dado durante estos años. A la gente de la montaña, especialmente a Raúl y Ana, por hacer este proceso más llevadero. No me puedo olvidar de mi tía Amalia y mi primo Antonio, porque también gracias a ellos estoy aquí. A mis suegros Manuela y Serafín les debo un agradecimiento eterno, por habernos apoyado en los momentos más duros. A mis hermanos Javi, Cristina y Mari les debo la felicidad durante tantos años. El tesón, cariño, y sacrificio de mis padres lo tengo siempre presente. Finalmente, gracias a mi compañera durante todo este camino, Nuria, porque sin ella nada de esto podría haber ocurrido.

Granada, 23 de enero de 2013

Resumen

De entre los problemas que aún se encuentran abiertos en el campo del reconocimiento automático del habla, uno de los que suscitan mayor interés entre la comunidad científica es el robustecimiento de estos sistemas frente al ruido acústico. Es bien conocido que, en presencia de ruido, el rendimiento de estos sistemas se deteriora hasta niveles en los que su uso resulta ineficaz. A fin de mitigar este problema, en este trabajo se desarrollan un conjunto de técnicas que permiten incrementar la robustez de estos sistemas en condiciones ruidosas.

Para alcanzar esta meta, en esta tesis se adopta un marco de trabajo en el que las características de voz usadas por el reconocedor son procesadas para mitigar la degradación producida por el ruido. Bajo este marco genérico, a lo largo de este trabajo proponemos varias técnicas que, usando estimación bayesiana MMSE e información a priori sobre la voz y/o el ruido, permiten obtener unas características más limpias.

En primer lugar, se proponen un conjunto de técnicas que, a partir de datos estéreo, derivan una serie de transformaciones que se aplican a la voz ruidosa para compensarla. Estos datos estéreo consisten en grabaciones de voz donde se cuenta con señales de voz limpia y sus correspondientes versiones ruidosas. Usando estas grabaciones, las técnicas propuestas estiman un conjunto de parámetros que se emplean posteriormente para realzar las características extraídas de la voz ruidosa.

A fin de eludir la necesidad de datos estéreo de las técnicas anteriores, en segundo lugar se proponen un conjunto de técnicas de reconstrucción basadas en un *modelo de enmascaramiento* de las características espectrales de voz. De acuerdo a este modelo, el ruido acústico enmascara (distorsiona) ciertas características del espectro de la voz dejando otras intactas. Para recuperarse de esta pérdida de información, se proponen dos técnicas alternativas. La primera de ellas, denominada TGI (*Truncated-Gaussian based Imputation*), estima el valor de las regiones enmascaradas del espectro supuesto que se conoce por adelantado la clasificación entre regiones limpias y ruidosas de la señal observada. La segunda técnica, conocida como MMSR (*Masking-Model based Spectral Reconstruction*), emplea modelos a priori de voz y ruido para llevar a cabo esta tarea. Como se verá, la reconstrucción obtenida por MMSR se reduce a una combinación

lineal entre el espectro original y un espectro estimado en el que se consideran los posibles efectos del enmascaramiento del ruido.

Basándonos en la formulación desarrollada para la técnica MMSR, también se propondrá un algoritmo iterativo para la estimación del modelo de ruido empleado por esta técnica. Este algoritmo permite ajustar los parámetros de un modelo de mezcla de gaussianas al ruido acústico presente en cada elocución que se reconoce, consiguiendo de esta forma modelar las características variantes de este tipo de ruidos. Asimismo, de esta formulación también se derivará un algoritmo para la estimación de las máscaras empleadas para identificar las regiones limpias y ruidosas del espectro de voz. Estas máscaras pueden emplearse por TGI y otras técnicas similares.

Finalmente, en esta tesis también se exploran otros dos cuestiones importantes para la compensación de características: el modelado temporal de la voz y el tratamiento de la incertidumbre. En relación a la primera cuestión, se proponen dos alternativas para representar la evolución temporal de la voz en los modelos usados por las técnicas de compensación propuestas: modelado de segmentos de voz y modelado basado en modelos ocultos de Markov. En cuanto a la segunda cuestión, se investiga el cómputo de medidas que describan la incertidumbre residual del proceso de compensación de características de voz, así como su posterior uso en el reconocedor.

Abstract

Among the open problems in automatic speech recognition (ASR), noise robustness is still a major issue in this research field. It is well-known that ASR performance significantly deteriorates in noisy environments to the point of making this technology useless. Therefore, this thesis will present a new effort to increase ASR robustness to acoustic noise.

To achieve this goal, we will focus on feature compensation in this work. This approach enhances the noisy speech features before they are feed into the speech recognizer, so that acoustics models trained on clean speech can be used for recognition. In particular, several feature compensation techniques based on MMSE estimation will be proposed here.

First, a set of feature compensation techniques based on stereo data are proposed. These techniques make use of stereo databases that include both clean and noisy recordings to estimate the statistical relationship between clean and noisy features. The learned statistical relationship is then used to derive a set of transformations that are applied later for enhancing noisy speech.

To overcome the requeriment of stereo data in the previous techniques, we move towards a missing-data approach for feature compensation. According to this approach, when speech is corrupted by additive noise, some regions of the noisy spectra are less affected by noise than others. In the log-spectral domain, some regions of the spectra are completely masked by the noise energy and can be considered unreliable or missing, while other parts where the speech energy dominates can be considered reliable. In order to estimate the missing regions, two alternative techniques are proposed in this work. The first one, namely TGI, estimates the missing features by assuming that reliable and unreliable regions are identified in advance by means of missing data masks. The second technique, namely MMSR, uses *a priori* speech and noise models for this task. As will be shown, the estimated spectra obtained by MMSR reduces to a linear combination of the original speech signal and an estimated spectra computed for the case of speech totally masked by noise.

Other main contribution of this work is the proposal of an EM-based ML noise

model estimation algorithm. This algorithm estimates the paramaters of a Gaussian Mixture Model (GMM) to represent the acoustic noise in noisy speech signals. Furthermore, an algorithm for estimating the missing data masks employed by the imputation techniques (e.g. TGI) will be also proposed.

Finally, this thesis also briefly explores other important aspects for MMSE-based speech estimation: temporal modeling and uncertainty processing. Regarding the first one, two different approaches are considered: a sliding-window approach to model spectro-temporal patches of speech and an HMM-based modeling of speech. As will be shown, both approaches greatly improve the recognition performance. Second, we also study joint schemes for feature compensation and uncertainty exploitation in the recognizer: the so-called soft-data decoding and weighted Viterbi algorithm approaches. The experimental results show that, by leveraging this kind of information, increased noise robustness can be achieved.

Acrónimos

Acrónimo	Significado
AFE	ETSI Advanced Front-End
ASA	Auditory Scene Analysis
CASA	Computational Auditory Scene Analysis
CBR	Cluster-Based Reconstruction
CDF	Cumulative Distribution Function
CMLLR	Constrained MLLR
CS	Compressed Sensing
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
EM	Expectation Maximisation
ETSI	European Telecommunications Standards Institute
FE	ETSI Front-end
FFT	Fast Fourier Transform
fMLLR	Feature space MLLR
GMM	Gaussian Mixture Model
GPS	Global Positioning System
HMM	Hidden Markov Model
HTK	Hidden Markok ToolKit
IDCT	Inverse Discrete Cosine Transform
J-VQMMSE	Joint VQMMSE
KL	Kullback-Leibler divergence
LVCSR	Large Vocabulary Continuous Speech Recognition
MAP	Maximum A Posteriori
MD	Missing Data
MEMLIN	Multi-Environment Model-based Linear Normalization
MFCC	Mel-Frequency Cepstral Coefficient
ML	Maximum Likelihood
MLLR	Maximum Likelihood Linear Regression
MMSE	Minimum Mean Squared Error

Acrónimo	Significado
MMSR	Masking-Model based Spectral Reconstruction
MSE	Mean Squared Error
NAT	Noise Adaptive Training
PCA	Principal Component Decomposition
PDF	Probability Density Function
PLP	Perceptual Linear Prediction
PMC	Parallel Model Combination
PNCC	Power-Normalized Cepstral Coefficients
Q-VQMMSE	Quantized VQMMSE
RAH	Reconocimiento Automático del Habla
RASTA	RelActive SpecTrA
SAT	Speaker Adaptive Training
SFD	Speech Fragmend Decoding
SNR	Signal-to-Noise Ratio
SPLICE	Stereo Piece-wise Linear Compensation for Environments
SPP	Speech Presence Probability
SSM	Stereo-based Stochastic Mapping
S-VQMMSE	Simple bias VQMMSE
TF-MFR	Time-Frequency Missing Feature Reconstruction
TGI	Truncated-Gaussian based Imputation
VAD	Voice Activity Detector
VQ	Vector Quantization
VQMMSE	VQ-based MMSE estimation
VTS	Vector Taylor Series
WAcc	Word Accuracy
WER	Word Error Rate
WVA	Weighted Viterbi Algorithm
W-VQMMSE	Whitening-transformation based VQMMSE

Notación

Convenciones tipográficas	
x	variable escalar
$\boldsymbol{x}$	vector
$\boldsymbol{X}$	matriz
$\hat{x}$	estimación del valor de x
Símbolos y operadores	
$\equiv$	equivalente a
$\approx$	aproximadamente igual a
$\propto$	proporcional a
$\circ$	productor vectorial elemento a elemento
$*$	convolución
$\text{argmax}_x\, f(x)$	valor de x que maximiza $f(x)$
$\text{máx}_x\, f(x)$	valor de $f(x)$ cuando x maximiza $f(x)$
$O(f(x))$	orden de eficiencia en tiempo de $f(x)$

Vectores y matrices	
$\boldsymbol{x}$	vector columna de dimensión arbitraria
x_i	elemento i del vector $\boldsymbol{x}$
$\mathbf{1}$	vector columna de unos
$\mathbf{0}$	vector columna de ceros
$\boldsymbol{A}$	matriz
a_{ij}	elemento (i, j) de $\boldsymbol{A}$
$\boldsymbol{A}^\top$	traspuesta de $\boldsymbol{A}$
$\boldsymbol{A}^{-1}$	inversa de $\boldsymbol{A}$
$\text{diag}(\boldsymbol{x})$	matriz diagonal formada por $\boldsymbol{x}$
$\boldsymbol{I}$	matriz identidad
$\lvert\boldsymbol{A}\rvert$	determinante de $\boldsymbol{A}$
$\text{tr}(\boldsymbol{A})$	traza de $\boldsymbol{A}$
$\text{máx}(\boldsymbol{x}, \boldsymbol{y})$	operación máximo elemento a elemento
$\log(\boldsymbol{x}), \exp(\boldsymbol{x})$	operaciones logaritmo y exponencial aplicadas por elementos
$\boldsymbol{C}$	matriz DCT
$\boldsymbol{C}^{-1}$	matriz IDCT

Distribuciones de probabilidad

$P(\cdot)$	función de masa de probabilidad
$p(\cdot)$	función de densidad de probabilidad
$p(x, y)$	función de densidad conjunta de x e y
$p(x\|y)$	función de densidad condicional de x dado y
$\mathcal{N}(x; \mu, \sigma)$	distribución normal de media μ y varianza σ^2
$\mathcal{N}(x)$	distribución normal estándar ($\mu = 0$ y $\sigma = 1$)
$\Phi(x; \mu, \sigma)$	distribución normal acumulada de media μ y varianza σ^2
$\Phi(x)$	distribución normal acumulada estándar
$p(\boldsymbol{x})$	función de densidad de probabilidad multivariante
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	distribución normal multivariante de media $\boldsymbol{\mu}$ y covarianza $\boldsymbol{\Sigma}$
$\delta(x)$	delta de Kronecker (igual a 1 si $x = 0$ y 0 en otro caso)
$\delta_a(x)$	delta de Dirac centrada en a (igual a 0 para $x \neq a$ e integra a 1)
$\mathcal{U}_a(x)$	escalón unitario centrado en a (0 para $x < a$ y 1 para $x \geq a$)
$\mathbb{1}_A$	función indicatriz (igual a 1 si A es cierto, 0 en otro caso)
$\mathbb{E}[x]$	valor esperado de x
$\mathrm{Var}[x]$	varianza de x
$\boldsymbol{\mu}_{x\|y}$	media de la distribución condicional $p(\boldsymbol{x}\|\boldsymbol{y})$
$\boldsymbol{\Sigma}_{xy}$	matriz de correlación cruzada entre x e y
$\tilde{\mu}$	valor esperado de la gaussiana en el intervalo $(-\infty, y]$
$\tilde{\sigma}^2$	varianza de la gaussiana en el intervalo $(-\infty, y]$

Señales

$\boldsymbol{x}$	vector de características de voz limpia
$\boldsymbol{y}$	vector de características de voz ruidosa
$\boldsymbol{n}$	vector de características de ruido aditivo
$\boldsymbol{h}$	vector de ruido convolutivo
$\boldsymbol{X}$	secuencia de vectores de características limpias
$\boldsymbol{Y}$	secuencia de vectores de características ruidosas

Índices

t	índice temporal
T	número de vectores de características en la elocución
i	índice para los elementos en el vector de características
D	dimensión del vector de características
k	índice de componentes en el modelo de mezcla
k^*	índice de la componente más probable
M	número de componentes (gaussianas o celdas VQ) en el modelo
e	índice de entornos acústicos
E	número total de entornos acústicos
s	índice de los estados del modelo acústico

$\cdot_x$	subíndice para la voz limpia
$\cdot_y$	subíndice para la voz ruidosa
$\cdot_n$	subíndice para el ruido aditivo

Modelos de fuente

$\mathcal{M}_x$	modelo de voz limpia
$\mathcal{M}_y$	modelo de voz ruidosa
$\mathcal{M}_n$	modelo de ruido
$\pi^{(k)}$	probabilidad a priori de la componente k del modelo
$\mu^{(k)}$	media de la componente k del modelo
$\boldsymbol{\Sigma}^{(k)}$	matriz de covarianza de la componente k del modelo

Modelo analítico de distorsión

$\boldsymbol{f}(\cdot)$	modelo analítico de distorsión de voz
$\boldsymbol{g}(\cdot)$	función de discrepancia
$\boldsymbol{J}_x, \boldsymbol{J}_n, \boldsymbol{J}_h$	matrices jacobianas del modelo analítico de distorsión respecto a la voz, al ruido aditivo y al convolutivo
$\boldsymbol{r}$	vector de corrección

Técnicas VQMMSE

$\mathcal{C}_x^{(k_x)}$	celda k_x del modelo de voz limpia
$\mathcal{C}_y^{(k_y)}$	celda k_y del modelo de voz ruidosa
$\mathcal{C}_x^{(k_x,k_y)}$	subregión de $\mathcal{C}_x^{(k_x)}$ cuyos vectores ruidosos pertenecen a $\mathcal{C}_y^{(k_y)}$
$\boldsymbol{\mu}_x^{(k_x,k_y)}, \boldsymbol{\Sigma}_x^{(k_x,k_y)}$	media y matriz de covarianza de la subregión $\mathcal{C}_x^{(k_x,k_y)}$

Datos perdidos

$\boldsymbol{y}_u$	subvector con las características perdidas (ruidosas) de $\boldsymbol{y}$
$\boldsymbol{y}_r$	subvector con las características fiables (limpias) de $\boldsymbol{y}$
$\boldsymbol{x}_u$	vector de características de voz correspondiente a $\boldsymbol{y}_u$
$\boldsymbol{x}_r$	vector de características de voz correspondiente a $\boldsymbol{y}_r$
$\boldsymbol{n}_u$	vector de ruido aditivo asociado a $\boldsymbol{y}_u$
$\boldsymbol{n}_r$	vector de ruido aditivo asociado a $\boldsymbol{y}_r$
U	dimensión de $\boldsymbol{y}_u$
R	dimensión de $\boldsymbol{y}_r$
$\boldsymbol{m}$	máscara de segregación
$\boldsymbol{s}_u$	conjunto con los índices de los elementos perdidos en $\boldsymbol{y}$
$\boldsymbol{s}_r$	conjunto con los índices de los elementos fiables en $\boldsymbol{y}$
$w_i^{(k_x,k_n)}$	probabilidad de no enmascaramiento del elemento i de acuerdo a las componentes (k_x, k_n)

Modelado temporal de la voz

$\boldsymbol{P}$	matriz de proyección PCA
$\boldsymbol{A}$	matriz de transición entre estados del HMM
$\boldsymbol{B}$	matriz con los parámetros del HMM
$\gamma_t^{(k)}$	probabilidad a posteriori de la componente k en el instante de tiempo t
$\alpha_t^{(k)}$	probabilidad de avance del algoritmo de avance-retroceso
$\beta_t^{(k)}$	probabilidad de retroceso del algoritmo de avance-retroceso
q_t	estado oculto del modelo acústico en el instante de tiempo t
$\boldsymbol{q}$	secuencia de estados $(q_1, \ldots, q_T)$
$\boldsymbol{W}$	secuencia de palabras $(w_1, w_2, \ldots, w_m)$

Índice general

Índice de figuras

Índice de tablas

CAPÍTULO 1

Introducción

Como consecuencia del desarrollo de la informática, Internet y la telefonía móvil, en los últimos años hemos experimentado el auge de los teléfonos inteligentes (*smartphones*, en inglés): dispositivos móviles con capacidad para ejecutar aplicaciones, acceder a información y, por supuesto, desempeñar las mismas funciones que antes realizaba un "teléfono tradicional". Debido a sus características físicas, estos dispositivos presentan interfaces que pueden limitar la interacción del usuario con los mismos. Estas limitaciones son la causa de que aún hoy día ciertos sectores de la población (p.ej. personas de avanzada edad) sean reticentes a su uso. En otros casos, por ejemplo en personas con discapacidad, estas limitaciones constituyen una barrera infranqueable. Una de las posibles soluciones a este problema es el uso de interfaces más naturales entre el usuario y el dispositivo, como son aquellas basadas en *reconocimiento automático del habla* (RAH)[1].

El reconocimiento automático del habla es el proceso mediante el cual la señal de voz se transforma en texto a través de un programa informático. Esta tecnología se sitúa dentro del marco más general del *procesamiento de la voz*, en el que se incluyen también otras tecnologías como la síntesis de voz, la codificación de voz y la biometría por voz (identificación y/o verificación de locutor). En general, estas tecnologías persiguen replicar mediante una máquina la habilidad humana de escuchar, identificar y pronunciar frases de una lengua dada. Otra tecnología muy relacionada con las anteriores y que suele ir de la mano de éstas es el procesamiento del lenguaje natural, cuyo objetivo es modelar la capacidad humana de comprender y procesar el contenido del lenguaje humano.

[1]En esta tesis emplearemos indistintamente los términos reconocimiento automático del habla y reconocimiento automático de voz.

Debemos destacar que, a pesar de considerarse como una tecnología madura, el rendimiento del RAH aún se encuentra lejos de los niveles de precisión alcanzados por los seres humanos. Así, en tareas sencillas de reconocimiento como la transcripción de secuencias de dígitos, los humanos cometemos una tasa de error de menos del 0,009 % [179], mientras que los mejores sistemas de RAH rondan el 0,55 % [280]. En tareas más difíciles las diferencias se reducen: por ejemplo, en la transcripción de una conversación telefónica la tasa de error de palabras en los humanos está en torno al 4 %, mientras que los sistemas de transcripción automática tienen una tasa de error tres veces mayor [50, 84].

La diferencia entre la ejecución humana y tecnológica se atribuye a una serie de causas que incluyen la alta variabilidad de la voz [194], el pobre modelado del habla espontánea [23, 24, 179], limitaciones en la extracción de características en el habla convencional [24, 202], el marco estadístico [24, 236] y, por último, la limitada *robustez* de los sistemas de reconocimiento actuales [24, 120]. Por robustez entendemos la habilidad de un sistema para mantener su ejecución o degradarse ligeramente cuando es expuesto a un rango de situaciones adversas. Esta falta de robustez de los sistemas de RAH se debe a su fragilidad frente a discrepancias entre las condiciones de entrenamiento y reconocimiento. Así, por ejemplo, pequeñas diferencias entre las señales de voz que se reconocen y las empleadas durante el entrenamiento de estos sistemas pueden mermar la precisión de los mismos. De entre las causas que originan estas discrepancias cabe mencionar las diferencias en la voz debidas al locutor (p.ej. género, edad, estado de ánimo, acento, etc.), el uso de palabras o abreviaturas no consideradas en el entrenamiento, distorsiones introducidas por los transductores de entrada (respuesta en frecuencia del micrófono y efecto del canal), acústica de la sala (reverberaciones y ecos) y el efecto del *ruido ambiental*. Esta última causa, el ruido ambiental, es de suma importancia en los sistemas de RAH integrados en dispositivos móviles.

A diferencia de otros contextos de uso en donde el entorno es relativamente silencioso, el carácter móvil de estos dispositivos hace que los sistemas de RAH deban enfrentarse a un mayor número de ruidos que distorsionan la voz y la hacen menos inteligible. De nuevo nos encontramos que, frente a la capacidad de los seres humanos para reconocer el habla en presencia de distorsiones severas, el rendimiento del RAH decae hasta hacerlo inservible en ambientes ruidosos, incluso aplicando distintas técnicas diseñadas para mitigar su efecto [120, 179, 236].

Esta tesis aborda el problema del reconocimiento automático de voz en condiciones de ruido. A fin de mejorar su robustez, a lo largo de este trabajo se examinarán distintas aproximaciones para reducir la discrepancia producida por el ruido en el RAH. En especial se hará hincapié en el desarrollo de técnicas orientadas a mitigar la distorsión

producida por el ruido en la señal de voz que se reconoce. Mediante la aplicación de estas técnicas, como ya veremos, se conseguirá robustecer el comportamiento del RAH frente al ruido, llegando en ciertos casos a equiparse el rendimiento al obtenido usando voz sin degradar.

1.1. Reconocimiento automático del habla

En la introducción anterior se han mencionado algunas de las motivaciones que existen para emplear los sistemas de RAH en algunas aplicaciones específicas. Asimismo, se ha comentado que uno de los problemas a los que se enfrentan estos sistemas es su limitada robustez al ruido acústico. Con el fin de que esta limitación quede más patente, así como sus causas y posibles soluciones, en esta sección se hará una breve revisión de los diferentes módulos que componen un sistema de reconocimiento actual. Una descripción más detallada de este tema puede encontrarse en [34, 147, 215, 221].

Un sistema de reconocimiento automático consta de dos módulos diferenciados: un *front-end*, que ejecuta la adquisición y extracción de las características de voz, y un *back-end*, que ejecuta el proceso de reconocimiento en sí mismo). En los siguientes apartados se describen estos dos módulos.

1.1.1. Extracción de características de voz

El objetivo del bloque de extracción de características de un reconocedor consiste en, partiendo de la señal de voz digitalizada, extraer una serie de parámetros[1] que representen la voz y sean útiles de cara al reconocimiento. En principio los parámetros extraídos deberían de cumplir una serie de propiedades entre las que podemos mencionar las siguientes:

- *Compacidad*: el proceso de extracción debería de eliminar la información redundante y superflua de la voz, a fin de conseguir una representación eficiente de esta señal de cara a la tarea de reconocimiento. Esto redunda también en una mejor estimación posterior de los modelos empleados por el *back-end* del reconocedor, ya que el número de parámetros a estimar será menor

- *Discriminantes*: las características extraídas deberían de tener la capacidad de modelar con precisión los aspectos relevantes del habla y, además, que estos as-

[1] A lo largo de este trabajo usaremos indistintamente los términos parámetros y características de voz para referirnos a la representación de la voz realizada.

pectos facilitasen la discriminación entre las distintas unidades acústicas de la lengua (fonemas, por ejemplo).

- *Robustez*: los parámetros extraídos deben ser robustos, inmunes en la medida de lo posible, a las variaciones no deseadas de la señal de voz, como por ejemplo al ruido o a las diferencias entre distintos locutores.

A lo largo de los años se han explorado varias representaciones alternativas de las características de voz que satisfacen, en mayor o menor medida, las propiedades anteriores. Algunas de las representaciones más conocidas son la codificación predictiva lineal (LPC, *Linear Predictive Coding*) [188], las características tándem [80], el análisis PLP (*Perceptual Linear Prediction*, predicción lineal perceptiva) [134], los coeficientes PNCC (*Power-Normalized Cepstral Coefficients*, coeficientes cepstrales normalizados en potencia) [158, 158] y los parámetros MFCC (*Mel-Frequency Cepstral Coefficients*, coeficientes cepstrales en escala Mel) [64], constituyendo estos la parametrización estándar en la mayoría de reconocedores actuales. Debido a esto, en este trabajo nos centraremos en el estudio del análisis cepstral de la voz.

Para extraer los parámetros MFCC se aplican los siguientes pasos. En primer lugar la señal de voz es capturada por un micrófono y digitalizada usando un proceso de muestreo y cuantificación. Puesto que la mayor parte de la energía de la voz está contenida en las frecuencias inferiores a 5 KHz, las aplicaciones de RAH suelen usar una tasa de muestreo entre 8 KHz y 16 KHz. Sobre la señal digitalizada se aplica un filtro de preénfasis para eliminar la componente de continua y, además, realzar las componentes de alta frecuencia del espectro.

La señal resultante es entonces segmentada y enventanada en tramas solapadas parcialmente. Si se usan tramas de longitud fija, la duración de cada trama está entre 20 ms y 30 ms, tiempo durante el cual se asume que la voz es cuasi estacionaria. El desplazamiento elegido entre tramas suele rondar los 10 ms, valor que permite capturar la mayor parte de la variabilidad temporal de la voz. En cuanto al tipo de ventana aplicada, es común utilizar una ventana tipo Hamming, ya que se adapta bien a las características espectrales de la señal de voz obteniendo un buen compromiso entre la anchura del lóbulo principal (resolución espectral) y la amplitud de los lóbulos secundarios (fenómeno de *leakage*).

El espectro de cada segmento de voz se calcula usando la transformada discreta de Fourier. Posteriormente, se aplica un banco de D filtros triangulares en escala Mel sobre la magnitud del espectro para obtener una representación espectral más suave. La escala Mel supone una aproximación a la resolución en frecuencia del oído humano: casi lineal en las bajas frecuencias, pero aproximadamente logarítmica a frecuencias altas. Después de aplicar el banco de filtros sus D salidas son comprimidas usando un

operador no lineal, típicamente el logaritmo, para modelar la sensibilidad perceptiva del oído humano.

Al conjunto de D características resultantes del paso anterior, denominadas características log-Mel, se les aplica finalmente la transformada discreta del coseno (DCT, *Discrete Cosine Transform*) dando lugar a los coeficientes MFCC. La DCT permite decorrelar las características de voz y, con ello, reducir el número de parámetros usados por el reconocedor (típicamente el número de MFCCs usado es $D/2$). Asimismo, al decorrelar las características resultantes, los MFCCs pueden modelarse mejor con matrices de covarianza diagonales. Esto simplifica los algoritmos de reconocimiento y reduce significativamente la carga computacional, así como las necesidades de datos durante la etapa de entrenamiento.

En los sistemas de RAH es común aumentar los coeficientes MFCC con sus derivadas temporales discretas (características dinámicas). Estas derivadas se añaden para modelar el comportamiento no estacionario de la voz a lo largo del tiempo. La mayoría de los sistemas estiman únicamente la primera y segunda derivada, proporcionando características de velocidad y aceleración, respectivamente.

1.1.2. Motor de reconocimiento

El motor de reconocimiento o *back-end* del reconocedor de voz se encarga de buscar la secuencia de palabras (frase) $\boldsymbol{W} = (w_1, w_2, \ldots, w_n)$ que mejor explica la señal de voz observada $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T)$, siendo $\boldsymbol{x}_t$ el vector de características de voz (p.ej. MFCCs y sus derivadas) extraído en el instante de tiempo t por el *front-end*. De acuerdo a la teoría de decisión bayesiana, el reconocimiento del habla se puede formular como un problema de optimización de máximo a posteriori, esto es, encontrar la secuencia de palabras $\boldsymbol{W}^*$ que maximiza la probabilidad $P(\boldsymbol{W}|\boldsymbol{X})$:

$$\begin{aligned} \boldsymbol{W}^* &= \underset{\boldsymbol{W}}{\operatorname{argmax}}\, P(\boldsymbol{W}|\boldsymbol{X}) \\ &= \underset{\boldsymbol{W}}{\operatorname{argmax}}\, \frac{p(\boldsymbol{X}|\boldsymbol{W})P(\boldsymbol{W})}{p(\boldsymbol{X})} \\ &\propto \underset{\boldsymbol{W}}{\operatorname{argmax}}\, p(\boldsymbol{X}|\boldsymbol{W})P(\boldsymbol{W}). \end{aligned} \tag{1.1}$$

En la optimización de la ecuación anterior, como podemos observar, intervienen dos términos: $p(\boldsymbol{X}|\boldsymbol{W})$ y $P(\boldsymbol{W})$. El primero de ellos, $p(\boldsymbol{X}|\boldsymbol{W})$, nos proporciona la probabilidad de observar la señal de voz $\boldsymbol{X}$ supuesto que se ha pronunciado la frase $\boldsymbol{W}$. Para calcular este término usaremos dos fuentes de información: el léxico del reconocedor y el modelo acústico. Por otro lado, el término $P(\boldsymbol{W})$ mide la probabilidad

de la secuencia de palabras $\boldsymbol{W}$ en el contexto en el que se emplea el sistema de RAH. Esta probabilidad nos la proporciona el modelo del lenguaje del reconocedor.

En los siguientes apartados se proporcionan más detalles de los tres subsistemas del reconocedor que hemos comentado: el léxico, el modelo del lenguaje y el modelo acústico.

1.1.2.1. Léxico

El léxico o diccionario contiene una lista de las palabras reconocibles junto a sus transcripciones fonéticas. La representación fonética determina la secuencia de fonemas que pueden darse durante el reconocimiento de cada palabra. Por ejemplo, una misma palabra puede tener varias pronunciaciones fonéticas asociadas, fruto de considerar distintos acentos de un mismo idioma.

El tamaño del vocabulario puede ser pequeño en las aplicaciones que sólo requieren pocas palabras (p.ej. control automático por comandos o marcado por voz), mientras que otras tareas de reconocimiento requieren gran cantidad de vocabulario (p.ej. dictado o subtitulado automático). En este último caso cobra especial relevancia disponer de una cobertura amplia del vocabulario, ya que la aparición durante el reconocimiento de palabras no consideradas en el léxico trae consigo errores de reconocimiento.

1.1.2.2. Modelo del lenguaje

El modelo del lenguaje determina la probabilidad a priori $P(\boldsymbol{W})$ de las hipótesis de palabras consideradas por el reconocedor. Esta probabilidad es independiente de la señal observada $\boldsymbol{X}$, siendo función únicamente de aspectos relacionados con el lenguaje como las probabilidades a priori de las palabras en cada lengua o el dominio específico al que se aplica el reconocedor (p.ej. dictado automático, marcado por voz o transcripción de informes jurídicos).

Dentro de los modelos usados por los reconocedores de voz podemos distinguir entre los determinísticos y los estadísticos. Los modelos determinísticos se definen mediante gramáticas formales (reglas) que restringen el lenguaje que el sistema es capaz de reconocer. Debido a su escasa flexibilidad pero gran compacidad para representar las construcciones gramaticalmente correctas, las gramáticas formales se suelen emplear en tareas de reconocimiento de pequeño vocabulario. Por otro lado, los modelos estadísticos más populares son las N-gramáticas. Una N-gramática modela la probabilidad de una palabra w_i supuesto que se conocen las N palabras que la preceden $w_{i-1}, \ldots, w_{i-N+1}$. En base a estas probabilidades podemos evaluar la probabilidad glo-

bal de cada hipótesis como

$$P(\boldsymbol{W}) = \prod_{i=1}^{m} P\left(w_i | w_{i-1}, \ldots, w_{i-N+1}\right), \tag{1.2}$$

donde $\boldsymbol{W}$ tiene m palabras.

Las probabilidades $P\left(w_i | w_{i-1}, \ldots, w_{i-N+1}\right)$ de la N-gramática se estiman usando conteos de palabras en corpus extensos. Para limitar el número de parámetros del modelo del lenguaje, habitualmente se utilizan modelos con $N = 2$ (bigramáticas) o $N = 3$ (trigramáticas).

1.1.2.3. Modelo acústico

La tarea del modelo acústico es calcular con precisión y eficiencia $p(\boldsymbol{X}|\boldsymbol{W})$, esto es, la probabilidad de que la señal de voz emitida sea $\boldsymbol{X}$ supuesto que el locutor ha pronunciado las palabras $\boldsymbol{W}$. Para modelar esta probabilidad, las palabras se suelen descomponer en unidades acústicas más pequeñas (fonemas, trifonemas con contexto o, en tareas de reconocimiento simples, incluso palabras) a través de la información que proporciona el léxico. Cada unidad acústica se representa entonces mediante un modelo estadístico independiente a fin de abarcar la variabilidad acústica inherente a la pronunciación. En los sistemas de RAH actuales la opción más popular para este cometido es el uso del modelo oculto de Markov (HMM, *Hidden Markov Model*) con funciones de densidad continuas. Por tanto, el modelo acústico del reconocedor consiste en una concatenación de estos HMMs básicos para representar palabras.

El HMM describe las diferentes unidades acústicas de la voz como una secuencia de S segmentos estacionarios de señal $(s_1, \ldots, s_S)$ denominados estados del modelo de Markov. Generalmente cada HMM suele contar entre 3 y 5 estados si las unidades acústicas consideradas son fonemas. En el caso de modelar unidades acústicas mayores (p.ej. palabras), el modelo requerirá un número mayor de estados. Para simplificar el proceso de modelado se asume que la señal de voz es un proceso estocástico de Markov de primer orden. Esto implica que la probabilidad de estar en un estado del modelo en el instante de tiempo t sólo depende del estado que se visitó en el instante de tiempo anterior $t-1$,

$$P(q_t = s_j | q_{t-1} = s_i, \ldots, q_1 = s_k) = P(q_t = s_j | q_{t-1} = s_j), \tag{1.3}$$

donde se ha considerado una secuencia temporal de estados $\boldsymbol{q} = (q_1, \ldots, q_T)$.

A estas probabilidades se les denomina probabilidades de transición entre estados y se emplean para representar el orden de los estados en el HMM y la duración de los segmentos de señal en cada estado. Otras probabilidades que aparecen en los HMMs son

las probabilidades de observación o emisión, que describen la relación entre los vectores de características observados $\boldsymbol{x}_t$ y los estados del modelo. De nuevo la propiedad de Markov simplifica la tarea de modelado, dependiendo la probabilidad de observación de $\boldsymbol{x}_t$ únicamente del estado actual q_t del HMM. Para modelar la distribución de los vectores de características emitidos por cada estado, se suelen emplear modelos de mezcla de gaussianas (GMM, *Gaussian Mixture Model*) multivariantes. De esta forma, la probabilidad de emisión de los vectores de características para el estado s viene dada por

$$p(\boldsymbol{x}|s) = \sum_{k=1}^{M} P(k|s)\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_s^{(k)}, \boldsymbol{\Sigma}_s^{(k)}), \tag{1.4}$$

donde M es el número de gaussianas del GMM, $P(k|s)$ es la probabilidad a priori (peso) de la gaussiana k-ésima y $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_s^{(k)}, \boldsymbol{\Sigma}_s^{(k)})$ es una distribución de probabilidad normal de media $\boldsymbol{\mu}_s^{(k)}$ y matriz de covarianza $\boldsymbol{\Sigma}_s^{(k)}$.

Los parámetros de cada HMM, esto es, las probabilidades de transición entre estados y los parámetros de los GMMs que modelan cada estado, se estiman a partir de datos de entrenamiento siguiendo un criterio de optimización dado. Generalmente el criterio adoptado trata de maximizar la verosimilitud (ML, *Maximum Likelihood*) de los datos de entrenamiento, aunque otros criterios alternativos (p.ej. maximizar la discriminación entre las distintas unidades acústicas) también se han propuesto en la literatura con buenos resultados [218]. Elegido el criterio de optimización apropiado, se recurre entonces al algoritmo de Baum-Welch [22, 33] para estimar de forma iterativa los parámetros de los HMMs que definen el modelo acústico.

Finalmente, la probabilidad del modelo acústico $p(\boldsymbol{X}|\boldsymbol{W})$ se calcula sumando las probabilidades de todas las posibles secuencias de estados $\boldsymbol{q} = (q_1, \ldots, q_T)$ que pueden generar la secuencias de palabras $\boldsymbol{W}$,

$$p(\boldsymbol{X}|\boldsymbol{W}) = \sum_{\boldsymbol{q}} \prod_{t=1}^{T} p(\boldsymbol{x}_t|q_t)P(q_t|q_{t-1}). \tag{1.5}$$

1.1.2.4. Algoritmo de búsqueda

La tarea del algoritmo de búsqueda consiste en encontrar la secuencia óptima de palabras $\boldsymbol{W}^*$ que maximice la probabilidad a posteriori dada en la ecuación (1.1). La optimización de esta probabilidad requiere, en principio, evaluar la probabilidad (1.5) para todas las posibles secuencias de palabras que admite el modelo de lenguaje considerado y quedarse únicamente con la de mayor probabilidad. Puesto que esta tarea resulta inviable en la mayoría de los casos, se aplican aproximaciones de programación dinámica para obtener una solución subóptima. En estas aproximaciones el problema

de encontrar $\boldsymbol{W}^*$ se reformula como la búsqueda del camino de coste mínimo a través de un grafo con pesos, siendo estos pesos las probabilidades dadas por los modelos acústico y del lenguaje. Un algoritmo eficiente de búsqueda apropiado para estos casos es el algoritmo de Viterbi [269].

1.2. Objetivos

El objetivo general de esta tesis es proporcionar una serie de mecanismos que mejoren el rendimiento de los sistemas de reconocimiento automático del habla cuando trabajan en ambientes ruidosos. En particular, la meta que nos marcamos consiste en robustecer la etapa de extracción de características de los sistemas de reconocimiento para que éstas sean más inmunes a los efectos del ruido. De entre la multitud de caminos existentes para alcanzar esta meta, nos centramos en el desarrollo de técnicas de estimación bayesiana para inferir las características correspondientes a la señal de voz limpia. La elección de la estimación bayesiana frente a otras metodologías de estimación se debe, en primer lugar, a los buenos resultados que esta metodología proporciona en otros problemas similares del procesado de señal y, en segundo lugar, a la posibilidad de emplear información a priori en forma de modelos de fuente (voz y/o ruido) durante la estimación de las características de voz.

Los objetivos particulares de este trabajo se desglosan a continuación:

- Hacer una revisión bibliográfica de las técnicas que han sido propuestas en la literatura para abordar el problema del reconocimiento robusto de voz.
- Estudiar el efecto de las distintas fuentes de ruido (aditivo y convolutivo) sobre las características de voz.
- Desarrollar nuevas técnicas de realce de características de voz usando para ello el marco estadístico provisto por la estimación bayesiana.
- Evaluar las técnicas propuestas en bases de datos estándar y compararlas con otras técnicas de referencia.

1.3. Estructura de la memoria

Esta tesis consta de siete capítulos y tres apéndices que se agrupan en tres grandes bloques temáticos. El primer bloque, que comprende este capítulo introductorio y el capítulo 2, sirve de introducción al problema que se pretender abordar, realizando

además una revisión bibliográfica de las técnicas ya propuestas en la literatura. Seguidamente se presenta un segundo bloque compuesto por los capítulos 3, 4 y 5, en el que se estudian las diferentes técnicas propuestas en este trabajo. Finalmente, el tercer bloque temático lo componen los capítulos 6 y 7 y presenta la evaluación experimental de las técnicas propuestas, así como las conclusiones derivadas del trabajo realizado. A continuación se describe brevemente el contenido de cada uno de los capítulos.

El **capítulo 2** describe la problemática que genera para los sistemas de RAH el reconocimiento con señales de voz degradadas. Este capítulo consta de tres secciones diferenciadas. En la sección 2.1 se presenta un estudio de las distintas fuentes de ruido que pueden afectar a un sistema de reconocimiento, concluyendo que éstas pueden ser categorizadas en fuentes de ruido aditivo y convolutivo. Como consecuencia de este estudio se deriva un modelo matemático que nos permite evaluar analíticamente el efecto del ruido (aditivo y convolutivo) en las características de la señal de voz (p.ej. espectro log-Mel o MFCCs).

La segunda parte de este capítulo (sección 2.2) hace una revisión bibliográfica de las distintas aproximaciones propuestas en la literatura para abordar el problema del ruido en los sistemas de RAH. Así, grosso modo, veremos que existen tres enfoques diferentes: (i) extracción robusta de características, (ii) adaptación de los modelos acústicos del reconocedor a las condiciones de reconocimiento y (iii) modificación de las características con las que se reconocen, siendo este último el enfoque que se sigue en este trabajo, aunque también se considerarán las modificaciones que es necesario introducir en el motor de reconocimiento para tener en cuenta la incertidumbre asociada a las características modificadas. Durante el estudio de las distintas aproximaciones, y dada la importancia que cobra en esta tesis, se hará especial hincapié en el denominado paradigma de datos perdidos (*missing data*, en inglés) aplicado al reconocimiento robusto de voz. De acuerdo con este paradigma, la distorsión producida por el ruido aditivo en la señal de voz se interpreta de forma alternativa como una pérdida de ciertas regiones del espectro de la voz. Finalmente, en la sección 2.3 se revisarán algunas de las técnicas para estimar las características del ruido, como por ejemplo la densidad de potencia espectral del ruido aditivo.

El **capítulo 3** supone el comienzo de las propuestas realizadas en esta tesis. Así, la primera mitad del capítulo sentará las bases de la estimación bayesiana que emplearemos en el resto de la memoria. En concreto, la estrategia elegida para inferir las características de voz limpia será una estimación de tipo MMSE (*Minimum Mean Square Error*, mínimo error cuadrático medio) fundamentada en modelos a priori para la voz. Basándonos en este estimador genérico, en la segunda mitad del capítulo desarrollaremos un conjunto de técnicas de compensación que mejoran la calidad de

las características de voz cuando ésta se ve sometida a un número indeterminado de distorsiones. Para simplificar el proceso de derivación de estas técnicas, en este capítulo se asumirá que las características de estas distorsiones son conocidas a priori. En este caso la tarea de compensación se simplifica, ya que se reduce a modelar con precisión el efecto de estas distorsiones en la voz, pero no en estimarlas.

La suposición bajo la cual se asume que se conocen las características de las distorsiones que afectan a la voz es, como puede preverse, poco realista. Es por ello que en el **capítulo 4** se propone un nuevo conjunto de técnicas de compensación orientadas a paliar esta deficiencia de las técnicas presentadas en el capítulo 3. Dos son los pilares en los que se fundamentarán este nuevo conjunto de técnicas: (i) el estimador MMSE presentado en el capítulo 3 y (ii) una aproximación al modelo analítico de distorsión de las características de voz al que denominaremos *modelo de enmascaramiento*. Tal modelo proporciona una expresión matemática para evaluar el efecto del ruido aditivo en las características de voz expresadas en el dominio log-Mel. De acuerdo a este modelo, la distorsión producida por el ruido aditivo se simplifica a un problema de enmascaramiento, es decir, ciertas regiones del espectro de la voz se encuentran enmascaradas por la energía del ruido y viceversa. En base a esta observación, en este capítulo se desarrollan dos técnicas de estimación orientadas a reconstruir las regiones enmascaradas de los espectros de voz: TGI y MMSR.

En la técnica TGI (*Truncated-Gaussian based Imputation*, imputación basada en gaussianas truncadas) el proceso de reconstrucción espectral se efectúa tomando como base información disponible a priori sobre la fiabilidad o no de cada elemento del espectro. Esta información se materializa en máscaras binarias que indican, para cada elemento del espectro de voz observado, si dicho elemento se encuentra dominado por la energía de la voz (fiable) o, por contra, por la del ruido (no fiable o perdido). Usando estas máscaras, la técnica TGI infiere el valor de los elementos perdidos explotando para ello las correlaciones existentes entre los distintos elementos del espectro. Dichas correlaciones se representan de forma eficiente mediante modelos a priori de las características de voz.

En lugar de usar máscaras binarias, la técnica MMSR (*Masking-Model based Spectral Reconstruction*, reconstrucción espectral basada en el modelo de enmascaramiento) emplea descripciones probabilísticas del ruido que contamina cada elocución. Como veremos, esta forma de proceder presenta varias ventajas. En primer lugar, parte de los cálculos efectuados por MMSR pueden considerarse, de forma alternativa, como una máscara continua que indica la fiabilidad de cada elemento del espectro. Esto supone que ya no es necesario estimar dichas máscaras en un paso previo al proceso de reconstrucción, como ocurre en TGI, sino que la reconstrucción del espectro y la

estimación de la fiabilidad de los elementos del espectro se realizan conjuntamente. Por otro lado, la formulación desarrollada para MMSR nos permite estimar iterativamente los propios modelos de ruido que esta técnica necesita. Para ello se propondrá una versión del algoritmo EM orientada a estimar, para cada elocución de *test*, un GMM que modela el ruido presente en dicha frase.

El **capítulo 5** se centrará en el estudio de otros aspectos relevantes para el proceso de compensación, como son el modelado temporal de la voz y el tratamiento de la incertidumbre de la estimación. En las técnicas de compensación descritas anteriormente, los modelos a priori de voz empleados únicamente representan la distribución en frecuencia de la voz, pero no así su evolución temporal. Debido a que esta información es una característica importante de este tipo de señales, en este capítulo se considera su explotación dentro de los modelos de voz a fin de mejorar, por consiguiente, la precisión de las distintas técnicas de compensación.

Otro de los aspectos clave que se tratarán en este capítulo es el tratamiento de la incertidumbre de las estimas de voz. Como cabe prever, las estimaciones obtenidas por las distintas técnicas propuestas no son idénticas a las limpias, sino que en función de determinados factores como la aleatoriedad del ruido, la relación señal-ruido de la señal observada, etc., serán más o menos fidedignas. Por tanto, sería deseable disponer de medidas que indiquen la fiabilidad de las distintas estimaciones. Asimismo, estas medidas deberían tenerse en cuenta durante el proceso de reconocimiento, de forma que las estimas menos fiables tengan un menor peso que las más fiables. Ambos aspectos también se tratarán en el capítulo 5.

El **capítulo 6** estará dedicado a evaluar el rendimiento de las distintas técnicas presentadas a lo largo de esta tesis. Para ello se llevarán a cabo una serie de experimentos de reconocimiento en bases de datos con voz degradada artificialmente. En concreto, las bases de datos empleadas serán Aurora2 [141], la cual define una tarea de reconocimiento de dígitos conectados, y Aurora4 [140], que define una tarea de gran vocabulario. Además de las técnicas propuestas, en este capítulo también se evalúan otras técnicas de referencia que se encuentran en la literatura.

Finalmente, en el **capítulo 7** se presentan las conclusiones, un resumen de las contribuciones realizadas y líneas futuras de investigación derivadas de este trabajo.

CAPÍTULO 2

Reconocimiento de voz robusto al ruido

Como ya se comentó en el capítulo 1, la precisión de los sistemas de reconocimiento del habla se deteriora rápidamente cuando las condiciones en las que son explotados difieren de las condiciones en las que fueron entrenados. En este sentido, factores tales como el ruido acústico, las diferencias ínter e intralocutor (género, edad, acento, estado anímico, etc.), y los medios usados para capturar la señal de voz y para transmitirla, influyen significativamente en el rendimiento y experiencia de uso de este tipo de sistemas. Por tanto, se hace necesario un estudio que permita analizar en qué medida estos factores afectan a los sistemas de reconocimiento del habla y qué técnicas se han propuesto en la literatura para contrarrestarlos. De todos los factores mencionados, nuestro estudio se centrará en el ruido acústico, por ser uno de los que más seriamente limitan el funcionamiento de los sistemas de reconocimiento en situaciones reales [24, 120, 258].

Este capítulo se estructura de la siguiente manera. En la sección 2.1 se estudiarán las fuentes de variabilidad que influyen en la precisión de los sistemas de reconocimiento del habla. Nuestro estudio profundizará en el efecto del ruido aditivo y de canal, presentando un modelo matemático que permite analizar cuantitativamente el grado de degradación de la voz cuando estas fuentes de ruido se encuentran presentes. Asimismo, se presentarán simulaciones de cómo se modifican las distribuciones estadísticas que modelan la voz limpian en presencia de ruido. Visto el efecto del ruido sobre las características de voz utilizados por los sistemas de reconocimiento, en la sección 2.2 se hará una revisión bibliográfica de las técnicas que han sido propuestas para compensar este problema, a saber: representaciones de la voz robustas al ruido, técnicas

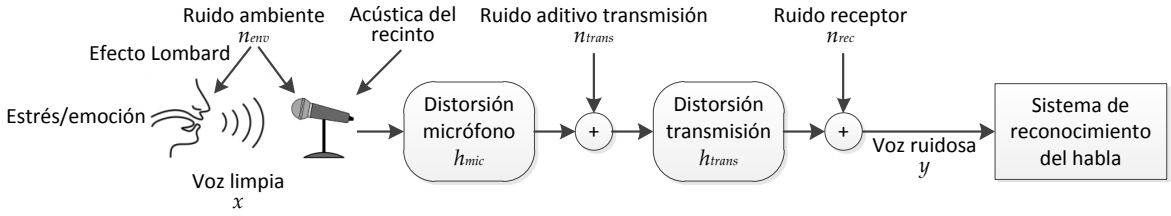

Figura 2.1: Fuentes de ruido y variabilidad que afectan a la voz (adaptada de [132]).

de adaptación de los modelos acústicos, técnicas de compensación y normalización de las características de voz, enfoques de reconocimiento basados en la incertidumbre y, por último, técnicas basadas en el paradigma de datos perdidos. Para terminar este capítulo, en la sección 2.3 se hará un repaso de las técnicas propuestas en la literatura para la estimación de las características del ruido.

2.1. Modelo de distorsión de las características voz

Cuando los sistemas de reconocimiento automático del habla funcionan en situaciones reales, estos se ven sometidos a un número de fuentes de ruido o variabilidad que distorsionan la voz emitida por el usuario. En la figura 2.1 se recogen la mayoría de estas fuentes, las cuales podemos clasificar en tres categorías diferentes: (i) variaciones debidas al locutor, (ii) ruido ambiental y (iii) efectos del canal de transmisión. El estado de físico, emocional y otros factores intrínsecos al locutor afectan a la voz emitida por éste alterando, por ejemplo, la amplitud y la distribución del tono fundamental de ésta [132, 253]. No sólo la voz se ve afectada por características intrínsecas al locutor, sino que otros factores externos, como son el ruido acústico, condicionan la forma en la que ésta se produce. El ejemplo más representativo lo tenemos en el llamado efecto Lombard [132, 152, 153]: en presencia de ruido los humanos tendemos a hiperarticular, enfatizando las vocales para que la voz sea más inteligible.

Por otro lado, y siendo el foco principal de estudio de esta tesis, tenemos los factores de distorsión debidos al entorno acústico donde el locutor desarrolla la interacción con el sistema de reconocimiento. Así, podemos mencionar el ruido aditivo como el representante más importante dentro de esta categoría (n_{env} en la figura 2.1), por ser el que afecta de forma más directa a la precisión de los sistemas de RAH. Por ruido aditivo entendemos la combinación de las señales emitidas por todas aquellas fuentes sonoras presentes mientras el locutor habla (p.ej. otras personas hablando, ruido del motor de un automóvil, etc.), y que se suman a la señal de voz en el dominio del tiempo. Este ruido es capturado por el transductor de entrada (micrófono) haciendo que la voz

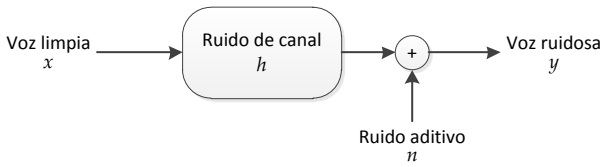

Figura 2.2: Modelo lineal simplificado de distorsión de la voz.

pierda calidad y sea menos inteligible. Las características del ruido aditivo pueden ser muy diversas, ya que incluye desde ruidos cuasi estacionarios (p.ej. ruido blanco o el ruido de un ventilador) fácilmente predecibles, hasta ruidos no estacionarios que cambian con el tiempo o ruidos espontáneos (p.ej. el ruido de un objeto al caer). Además del ruido aditivo, otra fuente de variabilidad es el ruido convolutivo (h_{mic} y h_{trans} en la figura). Este ruido se traduce en un filtrado de la señal de voz introduciendo distorsiones en ciertas frecuencias. Esencialmente, estas distorsiones vienen determinadas por las características de respuesta en frecuencia del micrófono que capta la señal, h_{mic}, y del canal de transmisión, h_{trans}. Otro tipo de distorsión debida a las características del entorno es la reverberación o eco: la señal de voz se mezcla con ecos de la misma reflejados en las paredes de la habitación.

Como últimas fuentes de distorsión, también podemos considerar los efectos debidos a la codificación de la voz para su transmisión eficiente sobre los canales de transmisión actuales (p.ej. VoIP), así como las distorsiones que se puedan generar a raíz de pérdidas de datos en estos [46, 47, 118, 148, 191, 215]. Por un lado, los codificadores de voz integran algoritmos de procesamiento de señal basados en consideraciones perceptivas que, en ciertas ocasiones, no son adecuados para las tareas de reconocimiento automático del habla [47]. Por otro lado, las congestiones en los nodos intermedios de encaminamiento de las redes de transmisión basadas paquetes de datos, suelen traducirse en pérdidas de datos que degradan la calidad de la señal sintetizada. Estos dos problemas, aunque de suma importancia en arquitecturas de reconocimiento distribuidas, no serán tratados en esta tesis.

El modelo de distorsión representado en la figura 2.1 puede simplificarse agrupando, por un lado, las distorsiones aditivas n_{env}, n_{trans} y n_{rec} en una sola fuente de ruido aditivo $n(t)$, y combinando, por otro lado, las distorsiones convolutivas representadas mediante las respuestas impulsivas h_{mic} y h_{trans} en una única respuesta $h(t)$. Este modelo lineal simplificado de distorsión, el cual fue propuesto inicialmente por Acero [8], se esquematiza en la figura 2.2. Como podemos ver, para cada instante de tiempo t

la señal de voz ruidosa $y(t)$ se obtiene a partir de una versión filtrada de la voz limpia emitida por el locutor $x(t)$, combinada de forma aditiva con el ruido ambiental $n(t)$. Una característica no considerada en este modelo simplificado, pero que sí recoge el modelo de la figura 2.1, es el efecto del ruido de canal $h(t)$ sobre $n(t)$. En efecto, la señal $n(t)$ contiene una versión filtrada del ruido ambiental n_{env}, donde el filtro aplicado viene determinado por las respuestas en frecuencia del transductor de entrada y el canal de transmisión. No obstante, no consideraremos esa situación aquí por simplicidad y, por tanto, supondremos que $n(t)$ y $h(t)$ son independientes.

2.1.1. Desarrollo matemático del modelo de distorsión

A continuación procedemos a desarrollar matemáticamente el modelo que relaciona los cuatro elementos que aparecen en la figura 2.2, a saber, la señal de voz ruidosa $y(t)$, la de voz limpia $x(t)$, el ruido aditivo $n(t)$ y la respuesta impulsiva de la distorsión lineal introducida por el canal $h(t)$. Para este desarrollo, nos apoyaremos en los trabajos [71, 164]. En el dominio del tiempo, el modelo de distorsión de la voz viene dado por,

$$y(t) = x(t) * h(t) + n(t), \tag{2.1}$$

donde $*$ es el operador de convolución.

Aplicando la transformada de Fourier discreta (DFT, *Discrete Fourier Transform*) sobre las señales enventanadas, la relación en el dominio de la frecuencia entre los términos de la ecuación (2.1) se traduce en

$$Y[j] = H[j]X[j] + N[j], \tag{2.2}$$

donde $j = 1, \ldots, N$ es el índice en el dominio de la DFT y $H[j]$ es la función de transferencia del canal. Como se puede apreciar, el canal se supone invariante al tiempo.

A partir de la ecuación (2.2), podemos obtener la densidad espectral de la señal de voz ruidosa como sigue,

$$\begin{aligned} |Y[j]|^2 &= |H[j]X[j] + N[j]|^2 \\ &= |H[j]|^2|X[j]|^2 + |N[j]|^2 + 2|H[j]X[j]||N[j]|\cos\theta_j, \end{aligned} \tag{2.3}$$

siendo θ_j el ángulo en el plano complejo entre los vectores $N[j]$ y $H[j]X[j]$ y coincide con la diferencia entre las fases de la señal de voz limpia y el ruido.

Aunque otros trabajos [69, 70, 71, 87, 164, 170, 251] consideran esta información de fase, en esta tesis supondremos que es nula. Esto se justifica porque, por un lado, dicha suposición es válida desde el punto de vista del valor esperado ($\mathbb{E}[\cos\theta_j] = 0$) y, por otro lado, en el tracto vocal no existe un mecanismo que sincronice la fase de la

voz con el ruido del entorno si exceptuamos el efecto Lombard. Por tanto, la ecuación anterior puede reescribirse como

$$|Y[j]|^2 \approx |H[j]|^2|X[j]|^2 + |N[j]|^2. \tag{2.4}$$

El siguiente paso en un algoritmo de extracción de característica (sección 1.1.1), suele involucrar el análisis de la señal a través de un banco de filtros distribuidos en una escala perceptiva que imita la resolución en frecuencia del oído humano. El filtrado más conocido es aquél que utiliza D filtros con respuesta triangular y linealmente espaciados en el dominio de la frecuencia logarítmica siguiendo la escala Mel [64]. Si denotamos por w_{ij} la respuesta en frecuencia del filtro i-ésimo ($w_{ij} \geq 0$, $\sum_j w_{ij} = 1$), la salida de ese filtro correspondiente a la voz ruidosa, $\tilde{Y}_i$, se puede calcular de la siguiente forma:

$$\begin{aligned}
\tilde{Y}_i &= \sum_j w_{ij}|Y[j]|^2 \\
&= \sum_j w_{ij}\left(|H[j]|^2|X[j]|^2 + |N[j]|^2\right) \\
&= \sum_j w_{ij}|H[j]|^2|X[j]|^2 + \sum_j w_{ij}|N[j]|^2,
\end{aligned} \tag{2.5}$$

con $i = 1, \ldots, D$.

La ecuación anterior puede expresarse en función de las salidas del banco de filtros para la voz limpia, el ruido aditivo y el convolutivo. Si definimos estos valores como

$$\begin{aligned}
\tilde{X}_i &= \sum_j w_{ij}|X[j]|^2, \\
\tilde{N}_i &= \sum_j w_{ij}|N[j]|^2, \\
\tilde{H}_i &= \frac{\sum_j w_{ij}|X[j]|^2|H[j]|^2}{\tilde{X}_i},
\end{aligned} \tag{2.6}$$

entonces podemos reescribir la ecuación (2.5) de la siguiente forma

$$\tilde{Y}_i = \tilde{H}_i\tilde{X}_i + \tilde{N}_i. \tag{2.7}$$

A los valores obtenidos a la salida del banco de filtros Mel, se les suele aplicar una compresión logarítmica que trata de imitar la resolución en amplitud del ser humano.

Aplicando logaritmos naturales a la ecuación (2.7) tenemos que

$$\begin{aligned}\log \tilde{Y}_i &= \log\left(\tilde{H}_i\tilde{X}_i + \tilde{N}_i\right)\\ &= \log\left(\tilde{H}_i\tilde{X}_i\left(1+\frac{\tilde{N}_i}{\tilde{H}_i\tilde{X}_i}\right)\right)\\ &= \log\left(\tilde{H}_i\tilde{X}_i\right) + \log\left(1+\exp\left(\log\frac{\tilde{N}_i}{\tilde{H}_i\tilde{X}_i}\right)\right)\\ &= \log\tilde{X}_i + \log\tilde{H}_i + \log\left(1+\exp\left(\log\tilde{N}_i - \log -\tilde{X}_i\log\tilde{H}_i\right)\right). \end{aligned} \tag{2.8}$$

Denotando por $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{n}$ y $\boldsymbol{h}$ a los vectores correspondientes a las salidas del banco de filtros Mel en escala logarítmica para la señales de voz ruidosa, voz limpia, ruido aditivo y ruido convolutivo, respectivamente, (p.ej. $\boldsymbol{y} = (\log\tilde{Y}_1, \log\tilde{Y}_2, \ldots, \log\tilde{Y}_L)^\top$), la ecuación (2.8) queda finalmente expresada de la siguiente forma

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{h} + \log\left(\mathbf{1} + e^{\boldsymbol{n}-\boldsymbol{x}-\boldsymbol{h}}\right) = \boldsymbol{x} + \boldsymbol{h} + \boldsymbol{g}(\boldsymbol{n}-\boldsymbol{x}-\boldsymbol{h}), \tag{2.9}$$

donde $\mathbf{1}$ es un vector de unos de tamaño D y las funciones $\exp(\cdot)$, $\log(\cdot)$ y $\boldsymbol{g}(\cdot)$ se aplican a cada componente de los vectores que tienen por argumento. La función $\boldsymbol{g}(\boldsymbol{z})$, denominada en la literatura como función de discrepancia (*mismatch function*, en inglés), se define de la siguiente forma:

$$\boldsymbol{g}(\boldsymbol{z}) = \log\left(\mathbf{1} + e^{\boldsymbol{z}}\right). \tag{2.10}$$

En ciertas ocasiones supondremos que el ruido convolutivo es nulo ($\boldsymbol{h} = \mathbf{0}$), por lo que la ecuación (2.9) se podrá simplificar de la siguiente forma:

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{n}) = \boldsymbol{x} + \boldsymbol{g}(\boldsymbol{n}-\boldsymbol{x}) = \log\left(e^{\boldsymbol{x}} + e^{\boldsymbol{n}}\right). \tag{2.11}$$

Finalmente, replicando el procesamiento de un extractor de características, podemos representar la relación entre las señales anteriores en el dominio del cepstrum. Sean $\boldsymbol{C}$ y $\boldsymbol{C}^{-1}$ las matrices de transformación de la transformada discreta del coseno (DCT) y su pseudoinversa (IDCT), entonces la relación dada por la ecuación (2.9) puede expresarse en el dominio del cepstrum como

$$\boldsymbol{y}^c = \boldsymbol{x}^c + \boldsymbol{h}^c + \boldsymbol{C}\log\left(\mathbf{1} + e^{\boldsymbol{C}^{-1}(\boldsymbol{n}^c - \boldsymbol{x}^c - \boldsymbol{h}^c)}\right), \tag{2.12}$$

donde los vectores que aparecen en esta expresión se obtienen tras proyectar los vectores con las energías en escala log-Mel al dominio cepstral, esto es,

$$\begin{aligned}\boldsymbol{y}^c &= \boldsymbol{C}\boldsymbol{y},\\ \boldsymbol{x}^c &= \boldsymbol{C}\boldsymbol{x},\\ \boldsymbol{n}^c &= \boldsymbol{C}\boldsymbol{n},\\ \boldsymbol{h}^c &= \boldsymbol{C}\boldsymbol{h}. \end{aligned} \tag{2.13}$$

De nuevo, en el caso de que el ruido convolutivo sea nulo, el modelo anterior se puede simplificar como

$$\boldsymbol{y}^c = \boldsymbol{f}_c(\boldsymbol{x}^c, \boldsymbol{n}^c) = \boldsymbol{x}^c + \boldsymbol{C}\log\left(\mathbf{1} + e^{\boldsymbol{C}^{-1}(\boldsymbol{n}^c - \boldsymbol{x}^c)}\right) \tag{2.14}$$

El modelo presentado contempla cómo se modifica la voz en los dominios que con mayor frecuencia usan los sistemas de reconocimiento automático del habla, esto es, los dominios log-Mel y cepstral. De ahí que el modelo haya sido exhaustivamente utilizado por las distintas técnicas de reconocimiento robusto al ruido propuestas en la literatura y, de forma análoga, será ampliamente empleado a lo largo de esta tesis.

2.1.2. Efecto del ruido sobre la distribución estadística de la voz

En el apartado anterior se ha presentado un modelo matemático que nos permite relacionar la señal de voz ruidosa observada, con las señales (ocultas) de voz limpia y ruido, tanto aditivo como convolutivo. A fin de tener un conocimiento más profundo de los efectos del ruido sobre la voz, en este apartado se presentan varias simulaciones sintéticas que recogen cómo se distorsionan las distribuciones de probabilidad de la voz debido al ruido. Para simplificar el análisis, sólo consideramos aquí el caso unidimensional, esto es, el efecto del ruido sobre una salida determinada del banco de filtros Mel. Asimismo, en dichas simulaciones se supondrá el modelo simplificado de distorsión con ruido de canal nulo dado en la ecuación (2.11).

La figura 2.11 muestra el efecto que tiene el ruido sobre las distribuciones de probabilidad de la voz limpia en el espacio log-Mel. En las simulaciones realizadas se ha supuesto que la distribución de probabilidad de la voz limpia $p(x)$ se puede aproximar mediante una distribución normal de media $\mu_x = 5$ y desviación típica $\sigma_x = 5$. Análogamente, la distribución del ruido $p(n)$ se ha supuesto gaussiana con una desviación típica fija de $\sigma_n = 1$ y cuya media oscila entre -3 y 9 con incrementos de 3 unidades en la escala logarítmica. La distribución de probabilidad de la voz ruidosa $p(y)$ se ha calculado utilizando el método de Montecarlo: en este método se generan N números aleatorios x_i, n_j $(1 \leq i, j \leq N)$ para cada distribución de probabilidad $p(x)$ y $p(n)$. A partir del modelo de distorsión de la ecuación (2.11) y los $2 \times N$ números aleatorios anteriores, podemos obtener el histograma de $y_i = \log(e^{x_i} + e^{n_i})$ por el cual aproximamos $p(y)$.

A partir de la figura 2.11, podemos concluir que [175, 199, 258]:

- A altas SNRs la distribución de probabilidad de la voz permanece inalterada por el ruido (ver figura 2.3-a), mientras que a bajas SNRs el ruido enmascara

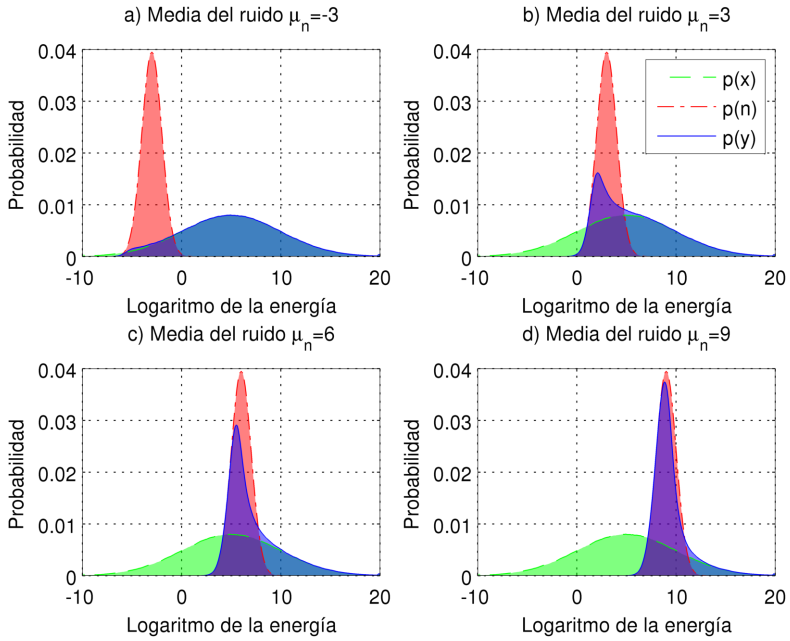

Figura 2.3: Efecto del ruido sobre las distribuciones de probabilidad de la voz limpia. La distribución de la voz limpia $p(x)$ se supone gaussiana $\mathcal{N}_x(\mu_x = 5, \sigma_x = 5)$, la distribución del ruido $p(n)$ también se supone normal con desviación estándar $\sigma_n = 1$ y medias distintas para cada panel; la distribución de la voz ruidosa $p(y)$, por último, se ha obtenido mediante el método de Montecarlo a partir de $p(x)$ y $p(n)$.

completamente la voz y $p(y)$ se aproxima a $p(n)$ (ver figura 2.3-d).

- El ruido introduce una distorsión no lineal que, como podemos ver en los paneles b) y c) de la figura 2.3, causa que $p(y)$ no sea normalmente distribuida. El efecto de esta distorsión sobre los parámetros de $p(y)$ viene determinado por el nivel de SNR y la varianza del ruido.

- Se produce una pérdida de información debido a la aleatoriedad introducida por el ruido, esto es, no existe biyectividad entre los elementos de $p(x)$ y $p(y)$. La razón principal de ello es que, debido a la varianza del ruido, varios (y distintos) valores de voz limpia pueden tener como imagen un mismo valor de voz ruidosa.

De forma análoga a lo visto anteriormente para el espacio de las energías logarítmicas, en la figura 2.4 se muestra el efecto del ruido sobre la distribución del coeficiente cepstral de orden 0 (C_0), el cual está íntimamente ligado al valor de la energía en escala logarítmica. Para el cálculo de los histogramas representados se han distorsionado las frases de entrenamiento de la base de datos Aurora2 con ruido tipo *subway* a SNRs de

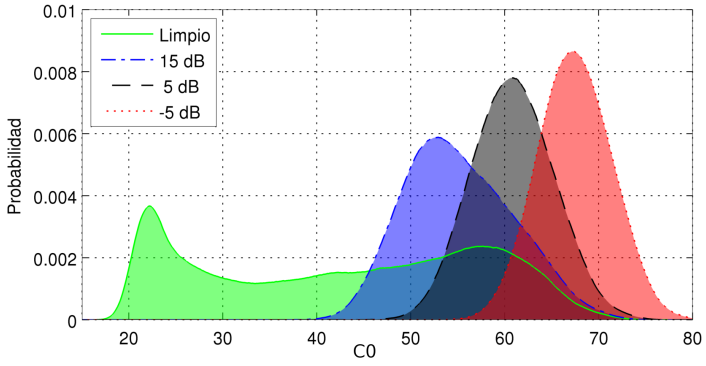

Figura 2.4: Distribución del coeficiente cepstral de orden 0 para distintas condiciones de ruido en la base de datos Aurora2.

15, 5 y -5 dB (ver sección 6.1.1.1) . Asimismo, se incluye la distribución del coeficiente C_0 calculada a partir de frases limpias (sin distorsionar). Como se puede observar, la característica distribución bimodal para el caso de la voz limpia debe sus dos picos a las estadísticas recopiladas para el silencio (pico centrado en, aproximadamente, 23) y para la voz (pico en 60). A medida que la SNR disminuye, la distribución va perdiendo su carácter bimodal y se va aproximando a una distribución normal con menor varianza centrada en el valor C_0 medio del ruido. En definitiva, el ruido provoca que los silencios presentes en las frases limpias pasen a ser segmentos de ruido.

2.2. Robustecimiento de los reconocedores de voz frente al ruido

En la sección anterior se ha estudiado cómo el ruido (aditivo y/o convolutivo) distorsiona las características de voz utilizadas por los sistemas de reconocimiento automático del habla, provocando una pérdida efectiva de información y modificando las distribuciones de probabilidad de la voz. Puesto que en condiciones acústicas desfavorables los modelos acústicos empleados por el reconocedor (entrenados, generalmente, con voz limpia) no modelan con propiedad la voz distorsionada, se produce una discrepancia en la etapa de reconocimiento que limita la precisión de estos sistemas. Con objeto de robustecer el funcionamiento de los reconocedores de voz frente a estas degradaciones, en la literatura se han propuesto una multitud de técnicas que podemos clasificar, a grandes rasgos, en cinco categorías [120, 215, 258]:

1. *Métodos de representación de la voz inherentemente robustos al ruido*, esto es, la

extracción de características que se vean menos afectados por el ruido.

2. *Adaptación de los modelos acústicos del reconocedor al entorno de reconocimiento.* Como se ha visto, el ruido modifica las distribuciones de probabilidad de la voz con la que se reconoce. Una posible solución al problema de reconocimiento en ruido es modificar los parámetros de los modelos acústicos para adaptarlos a la voz distorsionada.

3. *Técnicas de normalización y compensación de la voz ruidosa.* En vez de modificar los parámetros de los modelos acústico, estas técnicas procesan la voz ruidosa para que ésta se parezca lo máximo posible a la voz con la que fueron entrenados estos modelos. Distinguiremos entre técnicas de compensación, que intentan mitigar la distorsión producida por el ruido en las características de voz, y técnicas de normalización, que transforman las características a un dominio en el que les afecta menos el ruido.

4. *Técnicas basadas en la incertidumbre.* Podemos considerar ésta como una categoría híbrida de las dos anteriores. En primer lugar, la señal observada se compensa a fin de mitigar las degradaciones que afectan a la voz. Asimismo, a partir del proceso de compensación se extraen distintas medidas que indican la fiabilidad estimada de las características de voz procesadas. Estas medidas son usadas posteriormente por el reconocedor para otorgar un mayor peso en la etapa de descodificación a las características más fiables.

5. *Estrategias basadas en el paradigma de datos perdidos.* En ciertos dominios (p.ej. log-Mel) el efecto del ruido aditivo sobre la voz puede reducirse a un problema de datos perdidos (MD, *Missing data* en inglés), en donde ciertas de las características extraídas se consideran fiables y otras perdidas [57, 226]. Las características fiables son aquellas en las que la energía de la voz es dominante y, por tanto, apenas se ven alteradas por la acción del ruido. Por otra lodo, las características perdidas son aquellas donde la energía del ruido domina. Supuesto que se conoce a priori qué características del espectro observado son fiables y perdidas, las técnicas MD pueden, o bien intentar estimar las características perdidas del espectro, o bien reconocer con el espectro observado pero explotando la información sobre la fiabilidad de cada característica.

De entre las técnicas que se engloban en las cinco categorías anteriores, un gran grupo de ellas conlleva cierta transformación que reduce la discrepancia entre la estadística de la voz empleada en el entrenamiento y la de la voz con la que se reconoce. Desde este punto de vista, podemos realizar una segunda clasificación de las técnicas

en: (i) aquellas que modifican los parámetros de los modelos acústicos para adecuarlos a la voz con la que se reconoce y (ii) técnicas que modifican las características de la voz antes de reconocer con ellas. Existen técnicas específicamente diseñadas para adaptar los parámetros de los modelos acústicos (p.ej. MAP [109] o MLLR [168]), otras fueron diseñadas para limpiar las características de voz antes de ser usadas por el reconocedor (p.ej. sustracción espectral [37]), mientras que otras técnicas pueden usarse indistintamente como técnicas de adaptación de modelos o técnicas de compensación de características (p.ej. VTS [199] o Algonquin [165]). En este último caso, la elección entre implementar una técnica dada en su variante de adaptación de modelos o en su variante de compensación de características dependerá de varios factores, entre los cuales cabe considerar:

- En general, las técnicas de adaptación de modelos ofrecen mayor flexibilidad al permitir realizar transformaciones específicas en el espacio de estados del modelo acústico. Como ejemplo representativo podemos considerar la adaptación del modelo acústico a un locutor determinado, donde nos interesa transformar los parámetros específicos referidos a unidades acústicas (p.ej. fonemas, trifonemas, etc.).

- Aunque ofrecen una flexibilidad menor, las técnicas que modifican los vectores de características tienen como gran ventaja su mayor eficiencia computacional. A este respecto debemos considerar que en los actuales sistemas de reconocimiento de gran vocabulario, los cuales contienen cientos de miles de parámetros, el coste de adaptar el sistema sería prohibitivo. En cambio, la normalización de los vectores de características de la voz puede llegar a realizarse en tiempo real.

- Además, la mayoría de técnicas de adaptación propuestas en la literatura suponen implícitamente una arquitectura del tipo HMM-GMM (HMM con mezclas de gaussianas por estado) en su desarrollo. Por tanto, la aplicación de estas técnicas a otro tipo de arquitecturas (p.ej. HMMs con redes neuronales) puede resultar difícil.

En base a los factores anteriores, muchos de ellos contrapuestos, podemos concluir que la elección entre un tipo de técnica u otra dependerá del tipo de aplicación en la que se use el sistema de reconocimiento y los requisitos de ésta. Si bien es cierto esto, las técnicas de reconocimiento con incertidumbre nombradas anteriormente conforman una estrategia híbrida que auna beneficios de uno y otro enfoque (adaptación y compensación) y pueden considerarse como el punto de encuentro entre ambos. A la eficiencia computacional heredada de su implementación en el lado del cliente (*front-*

end), en estas técnicas hay que añadir la flexibilidad que permite el poder modificar los parámetros del modelo acústico mientras se reconoce.

Revisadas las características generales de las estrategias de reconocimiento robusto de voz, en las siguientes secciones procedemos a describir en detalle cada una de ellas, así como las técnicas propuestas que se enmarcan dentro de cada categoría.

2.2.1. Extracción robusta de características de voz

En esta categoría se encuadran aquellas técnicas que procesan la voz y extraen de ella una serie de características que, hasta cierto punto, son inmunes al ruido. En general, son técnicas que se aplican en las primeras etapas del proceso extracción de características y que, al contrario que los otros grupos de técnicas robustas para reconocimiento del habla, consiguen robustecer al sistema haciendo pocas suposiciones sobre el ruido [120]. Así nos encontramos que en la mayoría de casos las técnicas van dirigidas a calcular de forma robusta el espectro de la voz, trabajando para ello en dominios menos sensibles al ruido (p.ej. el dominio de la autocorrelación).

A pesar de las ventajas anteriores, podemos decir que la efectividad de estas técnicas será limitada. En efecto, es poco probable que, bajo las premisas consideradas por estas técnicas, sea posible la extracción de características robustas a una gran variedad de condiciones de ruido y niveles SNRs. Un ejemplo representativo de lo que acabamos de decir lo tenemos en el reconocimiento del habla en presencia de otro hablante [183]: en casos como éste donde las características del ruido son parecidas a las de la propia voz, el rendimiento de estas técnicas será limitado. De ahí que en la gran mayoría de los casos, las técnicas de extracción robusta de características sean un componente más dentro de un sistema más complejo de reconocimiento robusto de voz.

Dentro de las técnicas robustas de extracción de características podemos diferenciar distintas categorías. Por un lado, encontramos aquellas técnicas tales como SMC (*Short-term Modified Coherence*, coherencia modificada a corto plazo) [189] u OSALPC (*One-Sided Autocorrelation Linear Prediction Coding*, predicción lineal de la parte causal de la autocorrelación) [138], en las que el espectro de la señal se obtiene a partir de un modelado autorregresivo (AR) de la parte causal de la secuencia de autocorrelación. La razón de ello es la mayor robustez frente al ruido aditivo de la función de autocorrelación frente a la señal original en el dominio del tiempo [137]. Esto se debe a que el ruido aditivo suele afectar únicamente a los primeros coeficientes de la autocorrelación, o dicho de otro modo, en general el ruido está menos autocorrelacionado que la señal de voz, lo cual es válido para ruidos cuasi-blancos como por ejemplo el ruido del motor de un automóvil.

La propiedad anterior ha sido empleada por un gran número de técnicas para obte-

ner un espectro de voz más limpio. Por ejemplo, la técnica HASE (*Higher-lag Autocorrelation Spectrum Estimation*, estimación espectral a partir de los coeficientes altos de la autocorrelación) [242] consigue obtener un espectro más limpio para los segmentos de voz sonoros, eliminando para ello los primeros coeficientes de la autocorrelación. Hay que hacer notar que al eliminar estos coeficientes no se produce pérdida de información alguna, al menos para los segmentos sonoros de voz, ya que la periodicidad de la voz se ve reflejada también en la autocorrelación, siendo, por tanto, posible estimar el espectro de la señal a partir de los coeficientes altos. En el caso de los segmentos de voz no periódicos, la supresión de los coeficientes de orden bajo sí que supone una pérdida de información, obteniendo un espectro distinto a aquél que se obtendría utilizando la función de autocorrelación completa. No obstante, resultados de reconocimiento demuestran que al entrenar los modelos acústicos con este espectro modificado, el rendimiento en condiciones limpias no empeora, al tiempo que se obtienen mejores resultados de reconocimiento en condiciones ruidosas [195, 242].

Otros trabajos han investigado técnicas robustas de estimación espectral que combinan las ventajas de trabajar en el dominio de la autocorrelación, con el conocimiento de la frecuencia fundamental (F0 ó *pitch*) del hablante. En [195, 197] se proponen dos técnicas distintas para tal fin. Ambas técnicas basan su funcionamiento en la tabla de productos del segmento señal, esto es, una matriz simétrica que contiene los productos entre cada par de muestras del segmento de señal de voz a considerar. La primera de ellas, denominada *estimación espectral mediante promediado*, mejora el nivel de SNR de los segmentos sonoros de voz, supuesto que la frecuencia fundamental del hablante es conocida. Para ello calcula una autocorrelación promediada, realizando un promedio de los productos (o muestras de la señal ruidosa) separados por una distancia igual a la frecuencia fundamental. Esta autocorrelación promediada ha demostrado ser robusta frente a una gran variedad de tipos de ruidos, siempre que estos cumplan la condición de no ser armónicos con el mismo periodo que la voz. La segunda técnica, denominada *autocorrelación cribada*, puede considerarse una extensión de la técnica de *autocorrelación promediada*, ya que además se descartan los productos de las δ diagonales principales de la tabla de productos, siendo δ un parámetro empírico que depende del tipo de ruido. El valor de δ controla el número de coeficientes de autocorrelación de orden bajo que se descartan, por encontrarse estos más afectados por el ruido.

En [195, 198] se propone el uso de ventanas asimétricas centradas en la frecuencia fundamental del hablante, para una estimación robusta del espectro de la voz. En particular, la técnica propuesta implementa un procesado de la región causal de la autocorrelación mediante un enventanado de ésta a través de una ventana asimétrica de tipo DDR (*Double Dynamic Range*, rango dinámico doble). Se demuestra que la

ventana que proporciona la mejor tasa de reconocimiento es aquella centrada en la frecuencia fundamental promedio del ser humano, la cual consigue, por un lado, descartar los primeros coeficientes de la autocorrelación, y por otro explotar la periodicidad de la voz en los segmentos sonoros.

Hasta ahora las técnicas vistas de estimación espectral robusta han tenido como objetivo el robustecimiento frente al ruido aditivo. No obstante, en el modelo de distorsión presentado en la sección 2.1 vimos que otra fuente de distorsión a considerar es aquella provocada por el canal. La técnica RASTA (*RelAtive SpecTrAl*, espectro relativo) [135, 136] lleva a cabo un filtrado pasa-banda que elimina las distorsiones debidas a la respuesta en frecuencia del canal, suponiendo que ésta varía lentamente en el tiempo. Esta técnica ha sido mejorada dando lugar a la técnica conocida como J-RASTA [135, 163, 201], la cual permite abordar tanto el problema de los ruidos convolutivos, como el de los ruidos aditivos. En general, las técnicas RASTA y J-RASTA se han aplicado en conjunción con técnicas de extracción de características basadas en modelos auditivos. La técnica más conocida es la denominada como PLP (*Perceptual Linear Prediction*, predicción lineal perceptiva) [134]. El análisis PLP es similar al análisis LP (Linear Prediction, predicción lineal) de la voz, pero introduce tres elementos de la psicoacústica de la audición para calcular el espectro de la voz: la resolución espectral en banda crítica, las curvas de igual potencia de audición y un filtrado no uniforme basado en una escala de Bark.

Recientemente, Kim y Stern [157, 158] han propuesto un nuevo algoritmo de extracción de características de voz denominado PNCC (*Power-Normalized Cepstral Coefficients*, coeficientes cepstrales normalizados en potencia) que puede considerarse como el estado del arte en el campo de extracción de características robustas al ruido. El análisis PNCC toma muchas de las ideas introducidas por los algoritmos de extracción MFCC y PLP, como son el uso de un banco de filtros en escala auditiva, la compresión no lineal de las salidas de estos filtros y el análisis homomórfico para la obtención de coeficientes cepstrales. Tomando como base estas ideas, el algoritmo PNCC introduce ciertos bloques de procesamiento orientados a robustecer el análisis de voz frente al ruido. Los bloques más importantes son:

- Uso de un banco de filtros gammatone [212] con 40 filtros linealmente espaciados según la escala ERB (*Equivalent Rectangular Bandwidth*, ancho de banda rectangular equivalente) [192] entre 200 Hz y 8000 Hz. Los autores justifican el uso de los filtros gammatone por su robustez ligeramente mayor a la obtenida por los filtros triangulares Mel frente al ruido blanco.

- Filtrado asimétrico no lineal del nivel de ruido en cada canal. Este filtrado permite mejorar la calidad de la señal de voz para aquellos ruidos que varíen lentamente

con el tiempo. Los parámetros del ruido utilizados por el filtro se calculan a partir del nivel mínimo que toma cada canal sobre un rango de tiempo (generalmente 50-120 ms).

- Enmascaramiento temporal y suavizado espectral. Por un lado, los bloques de procesamiento anteriores suelen operar de forma independiente tanto a nivel de trama (tiempo) como a nivel de canal (frecuencia), por lo que es posible que se produzcan discontinuidades entre elementos vecinos en el espectrograma de la señal resultante. Estas discontinuidades suelen resultar en errores de inserción que merman el rendimiento del reconocedor. Para evitar estas discontinuidades, PNCC lleva a cabo un filtrado paso-bajo en el bloque de suavizado espectral. Por otro lado, distintos estudios psicoacústicos han demostrado que el oído humano es más sensible a los *onsets* en el espectro que a las regiones donde la energía cae. El bloque de enmascaramiento temporal incorpora esta observación a la extracción de características, suprimiendo la energía instantánea que se encuentre por debajo de un valor pico que se va actualizando continuamente.

Finalmente, comentamos aquí también aquellas técnicas que se han propuesto para salvar algunas de las limitaciones del modelado LP del espectro de la voz. En particular, es conocido que la predicción lineal modela de forma pobre la envolvente espectral en las frecuencias armónicas de los segmentos sonoros de voz [205]. Para contrarrestar este problema, se propone una estimación de la envolvente espectral basada en el modelo MVDR (*Minimum Variance Distortionless Response*, respuesta de mínima varianza sin distorsión) [205, 276]. Este modelo, además de proporcionar una envolvente espectral más precisa para segmentos sonoros y sordos de voz, demuestra ser más robusto a las distorsiones producidas por el ruido aditivo que, en el dominio de la densidad espectral de potencia, tiende a colmar los valles del espectro, dejando apenas inalterados los picos. El modelado todo-polos que MVDR implementa permite obtener una representación fiel de los picos y descartar la información sobre la estructura fina en los valles.

2.2.2. Adaptación de los modelos acústicos del reconocedor

Otra estrategia que ha dado muy buenos resultados a la hora de robustecer los sistemas de reconocimiento frente condiciones acústicas adversas es la denominada como adaptación de modelos. Las técnicas de adaptación de modelos modifican los parámetros de los modelos acústicos utilizados por el reconocedor (HMMs generalmente), para reducir la discrepancia con las condiciones de evaluación en las que son empleados. Así, tal y como se representa en la figura 2.5, los modelos entrenados generalmente con voz

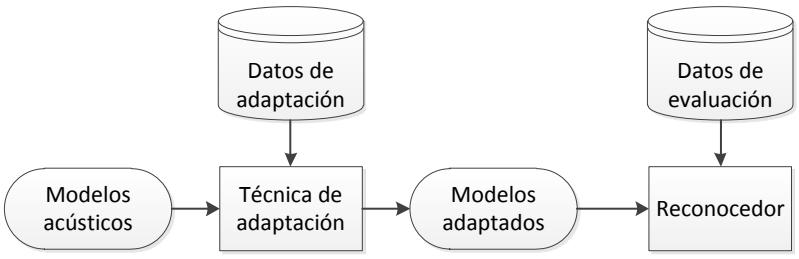

Figura 2.5: Esquema de la adaptación de modelos acústicos.

limpia se modifican para que modelen la voz adquirida en el entorno acústicamente adverso.

El proceso de adaptación puede hacerse *offline*, es decir, usando datos de adaptación disponibles antes de empezar el proceso de reconocimiento, u *online* a medida que avanza el reconocimiento de voz y usando los mismos datos de evaluación. Asimismo, la disponibilidad o no de las transcripciones asociadas a los datos de adaptación da lugar a algoritmos de adaptación supervisados y no supervisados, respectivamente. En este caso la elección de un algoritmo u otro dependerá de la disponibilidad de estas transcripciones: la adaptación supervisada suele proporcionar mejores resultados y, por tanto,es la que normalmente se emplea para hacer adaptación *offline*. Por otro lado, en la adaptación *online* no se cuenta con la transcripción de la voz que se reconoce, por lo que este proceso se realiza en base a una transcripción hipotética obtenida por el propio reconocedor. Así, la calidad de los modelos adaptados resultantes dependerá no sólo de la cantidad de datos de adaptación utilizados, sino también de la calidad de las transcripciones usadas para adaptar. En el caso extremo, los datos utilizados para adaptar los modelos pueden coincidir con los mismos datos de test que se reconocen: esto es lo que se conoce como *auto-adaptación* [102].

Además de su buen rendimiento, las técnicas de adaptación ofrecen gran flexibilidad para abordar distintos tipos de fuentes de variabilidad que afectan al reconocimiento como son el ruido (aditivo o de canal) o las diferencias entre locutores. En esta tesis, no obstante, sólo consideraremos las técnicas de adaptación de modelos en el marco de reconocimiento de voz robusto frente al ruido. Dentro de este marco específico podemos distinguir tres categorías diferentes: (i) técnicas de combinación de modelos, (ii) técnicas de compensación de modelos y (iii) técnicas de adaptación basadas en modelos de distorsión. En los siguientes apartados se describirán en detalle cada una de estas tres categorías y se aportarán ejemplos representativos de técnicas de adaptación que se encuadran dentro de cada de ellas.

Para terminar la sección de técnicas de adaptación de modelos, se hará una breve

introducción a las técnicas de *entrenamiento adaptativo* [16, 102, 281]; un esquema de entrenamiento en el que, por un lado, se dispone de un modelo canónico que representa los factores lingüísticos o variaciones deseadas de la voz, mientras que, por otro lado, se disponen de un conjunto de transformaciones lineales que modelan las variaciones no deseadas de la voz (p.ej. diferencias entre locutores o diferencias debidas al entorno acústico).

2.2.2.1. Adaptación estadística

Una de las estrategias más simples y al mismo tiempo más efectivas para abordar el problema de adaptación de los modelos acústicos al entorno de evaluación, consiste en entrenar dichos modelos directamente con voz distorsionada. Al entrenar los modelos con voz ruidosa se consigue modelar tanto las variaciones propias de la voz como las de la distorsión debida al ruido. En este sentido, el conocido como entrenamiento multicondición o entrenamiento *cocktail* [180] puede considerarse como una de las estrategias de entrenamiento robusto que mejores resultados de reconocimiento proporciona. En el entrenamiento multicondición los HMMs del reconocedor se entrenan con voz adquirida en los entornos donde el sistema va a trabajar. De esta forma, los modelos del reconocedor suelen entrenarse con voz contaminada usando diferentes tipos de ruido a diferentes niveles de SNR [141].

Generalmente esta estrategia de entrenamiento proporciona muy buenos resultados de reconocimiento para los entornos considerados en el entrenamiento, pero se degrada en presencia de ruidos y/o SNRs desconocidos. El problema en estos casos se debe a que el reconocedor no ha sido entrenado para representar las características de voz en esos entornos desconocidos. Para cubrir el mayor número de entornos, el entrenamiento multicondición requiere de gran volumen de datos adquiridos en múltiples entornos. En la práctica esto suele resultar muy difícil, por lo que esta estrategia de entrenamiento se suele utilizar preferentemente en aquellos sistemas para los que el número de entornos acústicos con los que el reconocedor va a interaccionar es conocido (p.ej. reconocedores empotrados en sistemas de navegación para automóviles o reconocedores empleados en oficinas).

A continuación se exponen algunas técnicas que proporcionan una alternativa al entrenamiento multicondición, mediante la adaptación de los modelos a partir de un número limitado de datos.

Máximo a posteriori. Cuando la cantidad de datos de adaptación es pequeña, el criterio de estimación ML usado por el algoritmo Baum-Welch provoca un sobreentrenamiento de los parámetros de los HMMs utilizados por el reconocedor, que deriva

hacia un rendimiento pobre del mismo. Para suavizar este problema, la técnica MAP (*Maximum a Posteriori*, máximo a posteriori) [109] considera los parámetros a adaptar como variables aleatorias, permitiendo con ello incorporar información a priori que guíe el proceso de estimación. La adaptación de los modelos acústicos en este caso se traduce en encontrar aquellos parámetros que maximizan la probabilidad a posteriori de los mismos. Formalmente, sea $\boldsymbol{X}$ la matriz con los datos de adaptación cuyas transcripciones asociadas vienen dadas por $\mathcal{H}$ y Θ el conjunto de parámetros de los HMMs que se quieren adaptar (medias, matrices de covarianza y pesos de las componentes en los GMMs). La estimación MAP del modelo de voz es aquella que maximiza la siguiente probabilidad a posteriori:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, p\left(\Theta|\boldsymbol{X}, \mathcal{H}\right) = \underset{\Theta}{\operatorname{argmax}}\, p\left(\boldsymbol{X}|\Theta, \mathcal{H}\right) p\left(\Theta|\Phi\right), \tag{2.15}$$

donde $p(\Theta|\Phi)$ es la distribución a priori de los parámetros Θ definida en base a los hiperparámetros Φ. Esta distribución previene que los parámetros del modelo sufran sobre-entrenamiento sobre el conjunto reducido de datos de adaptación.

Para estimar los parámetros del HMM que satisfacen la ecuación (2.15), la técnica MAP recurre a una estimación iterativa basada en el algoritmo EM [66]. La función auxiliar utilizada por el algoritmo EM en esta caso se compone de la función auxiliar utilizada por la estimación ML, más un término que depende de la probabilidad a priori de los parámetros del modelo [109]. En resumen, la función auxiliar viene dada por

$$\begin{aligned}&\mathcal{Q}(\hat{\Theta}, \Theta)\\ &= -\frac{1}{2}\sum_{t,k}\gamma_t^{(k)}\left[\log|\boldsymbol{\Sigma}^{(k)}| + (\boldsymbol{x}_t - \boldsymbol{\mu}^{(k)})^\top \boldsymbol{\Sigma}^{(k)^{-1}}(\boldsymbol{x}_t - \boldsymbol{\mu}^{(k)})\right] + \log p(\hat{\Theta}|\Phi),\end{aligned} \tag{2.16}$$

donde la sumatoria se realiza sobre todos los instantes de tiempo t y componentes k (gaussianas) del HMM. El término $\gamma_t^{(k)}$ en la ecuación anterior define la probabilidad de ocupación a posteriori de la gaussiana k-ésima en el instante de tiempo t. Este término se puede calcular de forma eficiente utilizando el algoritmo *forward-backward* [220] sobre los parámetros actuales Θ del modelo.

Una cuestión importante en la estimación MAP es la elección de la distribución a priori $p(\Theta|\Phi)$. La elección de una distribución a priori apropiada hará que las fórmulas de estimación obtenidas tengan una forma analítica cerrada. Desafortunadamente, no existe tal distribución cuando la función de verosimilitud empleada es un GMM (suponemos que los estados del HMM se modelan mediante GMMs). Para solventar este problema, en [109] se asume independencia estadística entre los pesos de las componentes y sus parámetros (media y covarianza). Bajo esta suposición, se demuestra que las distribuciones a priori conjugadas para los pesos de las componentes siguen

una distribución de Dirichlet, mientras $p(\hat{\Theta}|\Phi)$ es una distribución Normal-Wishart en el caso de los parámetros de cada componente. Finalmente, la estimación MAP de la media de cada componente k se calcula de forma iterativa atendiendo a la siguiente ecuación:

$$\hat{\boldsymbol{\mu}}^{(k)} = \frac{\tau\tilde{\boldsymbol{\mu}}^{(k)} + \sum_t \gamma_t^{(k)}\boldsymbol{x}_t}{\tau + \sum_t \gamma_t^{(k)}}, \tag{2.17}$$

donde $\tilde{\boldsymbol{\mu}}^{(k)}$ es la media de la distribución $p(\hat{\Theta}|\Phi)$ y τ es un parámetro de la distribución Normal-Wishart que controla la influencia de la distribución a priori y los datos de adaptación en el cálculo de $\hat{\boldsymbol{\mu}}^{(k)}$.

En la mayoría de los sistemas, y por razones de eficiencia, únicamente se actualizan los valores de las medias. No obstante, también es posible actualizar las matrices de covarianza y los pesos (probabilidades a priori) asociados con cada distribución. Las fórmulas de estimación utilizadas para adaptar estos dos términos pueden consultarse en [109].

Cuando el volumen de datos de adaptación tiende a infinito, la técnica de adaptación MAP converge a una estimación ML de los parámetros del HMM. No obstante, para conjuntos de adaptación de pequeños, la estimación MAP presenta deficiencias severas, ya que sólo modificara los parámetros de las unidades acústicas observadas en el conjunto de adaptación. Esto supone un problema muy grave a la hora de aplicar esta técnica a sistemas de reconocimiento de gran vocabulario, donde es normal que los modelos acústicos dispongan del orden de miles de gaussianas. Para paliar este problema, en la literatura se han propuesto varias extensiones de la técnica MAP. Por ejemplo, la técnica RMP (*Regression based Model Prediction*, predicción de modelos basada en regresión) [12] modela las correlaciones entre los parámetros del modelo acústico usando modelos de regresión lineal. Estas correlaciones son utilizadas luego para adaptar los parámetros de las unidades no observadas en base a aquellas observadas. Por otro lado, la técnica SMAP (*Structured MAP*, MAP estructurado) [244] utiliza un árbol para organizar las gaussianas jerárquicamente. El algoritmo de adaptación usa entonces una estrategia arriba-abajo a partir del nodo raíz, el cual contiene todas las componentes del modelo acústico. En cada nivel del árbol la distribución a priori utilizada por el algoritmo de adaptación es la obtenida para el nivel anterior. En comparación con la técnica MAP, la técnica SMAP es más eficiente computacionalmente hablando. Asimismo, ambas técnicas convergen si el volumen de datos de adaptación es suficiente.

Transformaciones lineales de los parámetros del modelo. Otro enfoque alternativo que se ha empleado ampliamente para adaptar los modelos acústicos del reconocedor es el uso de transformaciones lineales que, aplicadas sobre los parámetros

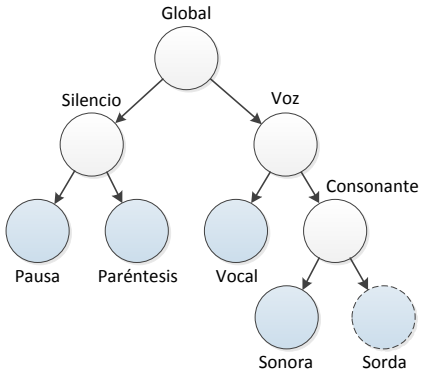

Figura 2.6: Ejemplo de árbol de regresión utilizado para jerarquizar las transformaciones empleadas durante el proceso de adaptación.

del modelo acústico, permiten reducir la discrepancia con los datos de adaptación. Por tanto, este enfoque puede utilizarse tanto para adaptar un modelo independiente del locutor a un locutor dado en cuestión, como para adaptar un HMM entrenado con voz limpia a un ambiente ruidoso determinado. En su versión más básica, todas las PDFs del modelo acústico comparten la misma transformación. No obstante, es posible obtener un conjunto de transformaciones que se aplican a conjuntos de PDFs específicas, consiguiendo con ello una mayor precisión en el proceso de adaptación.

Antes de estudiar los detalles relativos al uso y estimación de estas transformaciones, en los siguientes párrafos se presentan brevemente los *árboles de regresión* [98, 131, 167]: una estructura de datos que es empleada ampliamente para organizar las transformaciones de forma jerárquica . En los árboles de regresión las componentes similares del modelo acústico se agrupan formando un único nodo del árbol. En lugar de aplicar una transformación global a todos los parámetros del modelo, a cada grupo de componentes se le aplica una transformación específica estimada para dicho grupo. Partiendo del nodo raíz que contiene todas las PDFs del modelo acústico y, por tanto, una sola transformación global, cada nodo es divido en tantos hijos como fuere necesario. Para cada uno de los nodos hijos se estima una transformación específica siempre y cuando el conjunto de PDFs agrupadas bajo dicho nodo tenga suficientes datos de adaptación observados. En caso contrario, es decir, si no se disponen de suficientes datos de adaptación, las componentes de dicho grupo se adaptan con la transformación estimada para el nodo padre.

En la figura 2.6 se muestra un árbol de regresión de ejemplo en el que las distribuciones de probabilidad se han dividido atendiendo al tipo de unidad acústica que modelan. El árbol mostrado contiene cinco nodos terminales (pausa, paréntesis, vo-

cal, sonora y sorda), que a su vez tendrán asociados una transformación lineal si hay suficientes datos de adaptación observados para el nodo. Cuando no hay suficientes datos de adaptación (p.ej. el nodo *sorda* de la figura representado con línea punteada), la transformación empleada se calcula con los datos observados para el nodo padre (la transformación para las consonantes sordas sería la calculada para el grupo de las consonantes, sean sordas o sonoras).

Los árboles de regresión se pueden construir de dos formas: mediante la información proporcionada por un experto o de forma automática. En la figura 2.6 tenemos un ejemplo de árbol de regresión basado en conocimiento experto, esto es, la clasificación de las unidades acústicas atendiendo a distintos criterios fonéticos. Éste es el tipo de árboles que se suelen emplear para agrupar los estados comunes durante el entrenamiento de sistemas de reconocimiento de gran vocabulario. Por otra parte, también es posible agrupar automáticamente las distribuciones de probabilidad usando técnicas de *clustering* que exploten distintas medidas de similitud, como puede ser la distancia de Kullback-Leibler [245, 275] o medidas basadas en la correlación entre las unidades acústicas [131].

Una vez visto cómo se pueden estructurar las transformaciones, pasamos a ver cómo se estiman. De entre las técnicas de adaptación que emplean transformaciones lineales, la más conocida es la denominada como MLLR (*Maximum Likelihood Linear Regression*, regresión lineal de máxima verosimilitud) [99, 168]. La técnica MLLR estima una transformación afín que maximiza la probabilidad de los datos de adaptación. En su versión más simple, esta técnica sólo adapta las medias de las distribuciones de probabilidad,

$$\hat{\boldsymbol{\mu}}^{(k)} = \boldsymbol{A}^{(r)}\boldsymbol{\mu}^{(k)} + \boldsymbol{b}^{(r)}, \tag{2.18}$$

donde $\boldsymbol{\mu}^{(k)}$ es el vector media de la k-ésima gaussiana que se adapta, $\hat{\boldsymbol{\mu}}^{(k)}$ es el vector adaptado y $(\boldsymbol{A}^{(r)}, \boldsymbol{b}^{(r)})$ son los parámetros que modelan la transformación. Estos dos parámetros se estiman para la clase r del árbol de regresión a la que pertenece la componente k.

Definiendo las variables extendidas $\boldsymbol{W}^{(r)} = [\boldsymbol{A}^{(r)}\ \boldsymbol{b}^{(r)}]$ y $\boldsymbol{\xi}^{(k)} = [\boldsymbol{\mu}^{(k)T}\ 1]^\top$, entonces podemos reescribir la ecuación anterior de forma compacta como:

$$\hat{\boldsymbol{\mu}}^{(k)} = \boldsymbol{W}^{(r)}\boldsymbol{\xi}^{(k)}. \tag{2.19}$$

Por consiguiente, el objetivo de la técnica MLLR es estimar la matriz $\boldsymbol{W}^{(r)}$ que maximiza la probabilidad de los datos de adaptación, esto es,

$$\hat{\boldsymbol{W}}^{(r)} = \underset{\boldsymbol{W}}{\operatorname{argmax}}\, p\left(\boldsymbol{X}|\boldsymbol{W}, \Theta, \mathcal{H}\right), \tag{2.20}$$

donde $\boldsymbol{X}$ es el conjunto de datos de adaptación, Θ son los parámetros del HMM y $\mathcal{H}$ son las transcripciones de los datos de adaptación.

Dado que la ecuación anterior no presenta una solución analítica, se recurre a una solución iterativa basada en el algoritmo EM para el cálculo de los parámetros de la transformación. La función auxiliar a maximizar en este caso es

$$\mathcal{Q}\left(\boldsymbol{W}, \hat{\boldsymbol{W}}; \Theta\right) = -\frac{1}{2} \sum_{t} \sum_{k \in r} \gamma_t^{(k)} \left(\boldsymbol{x}_t - \hat{\boldsymbol{W}} \boldsymbol{\xi}^{(k)}\right)^{\top} \boldsymbol{\Sigma}^{(k)^{-1}} \left(\boldsymbol{x}_t - \hat{\boldsymbol{W}} \boldsymbol{\xi}^{(k)}\right), \tag{2.21}$$

siendo la primera sumatoria para todos los instantes de tiempo t y la segunda para todas las componentes k que pertenecen a la clase de regresión r.

Derivando la ecuación anterior respecto a $\hat{\boldsymbol{W}}$ e igualando a cero se obtiene un sistema de ecuaciones a partir del cual podemos calcular los parámetros de la transformación afín. Para el caso de matrices de covarianza diagonales, la fila d-ésima de la matriz $\hat{\boldsymbol{W}}$, $\hat{\boldsymbol{w}}_d$, se calcula de la siguiente manera,

$$\hat{\boldsymbol{w}}_d = \left(\boldsymbol{G}_d\right)^{-1} \boldsymbol{k}_d, \tag{2.22}$$

donde las variables auxiliares $\boldsymbol{G}_d$ y $\boldsymbol{k}_d$ se definen de la siguiente forma

$$\boldsymbol{G}_d = \sum_{t} \sum_{k \in r} \frac{\gamma_t^{(k)}}{\left(\sigma_d^{(k)}\right)^2} \boldsymbol{\xi}^{(k)} \boldsymbol{\xi}^{(k)^{\top}}, \tag{2.23}$$

$$\boldsymbol{k}_d = \sum_{t} \sum_{k \in r} \frac{\gamma_t^{(k)}}{\left(\sigma_d^{(k)}\right)^2} \boldsymbol{\xi}^{(k)} x_{t,d}, \tag{2.24}$$

siendo $\sigma_d^{(k)^2}$ el elemento d-ésimo de la diagonal principal de la matriz de covarianza $\boldsymbol{\Sigma}^{(k)}$ y $x_{t,d}$ representa el elemento d del vector $\boldsymbol{x}_t$.

Aunque la mayoría de los sistemas de reconocimiento sólo modifican las medias de las distribuciones de probabilidad que constituyen el modelo acústico, también es posible modificar las matrices de covarianza. En [99, 104] se propone una adaptación de la matriz de covarianza de la siguiente forma

$$\hat{\boldsymbol{\Sigma}}^{(k)} = \boldsymbol{L}^{(k)^{\top}} \boldsymbol{H} \boldsymbol{L}^{(k)}, \tag{2.25}$$

siendo $\boldsymbol{H}$ la transformación que se aplica a las matrices de covarianza y $\boldsymbol{L}^{(k)} = \boldsymbol{C}^{(k)^{-1}}$, con $\boldsymbol{C}^{(k)}$ la matriz triangular superior obtenida al factorizar la matriz de precisión $\boldsymbol{\Sigma}^{(k)^{-1}}$ por el método de Cholesky.

Una alternativa más eficiente para adaptar las matrices de covarianza es expresar la función de transformación de la siguiente forma [99, 208],

$$\hat{\boldsymbol{\Sigma}}^{(k)} = \boldsymbol{H} \boldsymbol{\Sigma}^{(k)} \boldsymbol{H}^{\top}. \tag{2.26}$$

Como ventaja de este enfoque, tenemos que la probabilidad de observación de la gaussiana adaptada puede calcularse de manera más eficiente de acuerdo a la siguiente expresión

$$\log \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}^{(k)}, \hat{\boldsymbol{\Sigma}}^{(k)}\right) = \log \mathcal{N}\left(\boldsymbol{H}^{-1}\boldsymbol{x}; \boldsymbol{H}^{-1}\boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\right) - \log |\boldsymbol{H}|. \tag{2.27}$$

Para ambos enfoques, los detalles sobre el cálculo de los parámetros de la transformación pueden encontrarse en [99].

En las fórmulas de adaptación anteriores de la técnica MLLR, las matrices de transformación empleadas para adaptar la media son distintas de aquellas utilizadas para adaptar las matrices de covarianza. En [75, 99], no obstante, se propone una versión de la técnica MLLR denominada CMLLR (*Constrained MLLR*, MLLR restringido), en la que la transformación es empleada tanto para adaptar las medias como las covarianzas. Entre otras ventajas, disponer de una transformación compartida reduce el coste computacional requerido para estimar los parámetros de ésta. Asimismo, el número de datos de adaptación necesarios para estimar de forma robusta dichos parámetros también se reduce. De forma genérica, la transformación que la técnica CMLLR aplica tiene la siguiente forma:

$$\hat{\boldsymbol{\mu}}^{(k)} = \boldsymbol{A}^{(r)}\boldsymbol{\mu}^{(k)} - \boldsymbol{b}^{(r)}, \tag{2.28}$$

$$\hat{\boldsymbol{\Sigma}}^{(k)} = \boldsymbol{A}^{(r)}\boldsymbol{\Sigma}^{(k)}\boldsymbol{A}^{(r)^{\top}}. \tag{2.29}$$

A la técnica CMLLR también se le conoce con el nombre de fMLLR (*feature space MLLR*, MLLR aplicada al espacio de características) [174, 219, 234], ya que la transformación afín que se aplica para adaptar los parámetros del modelo acústico también puede llevarse a cabo en el espacio de características. De esta forma se consigue una mayor eficiencia computacional, al no tener que modificar los parámetros del modelo acústico mientras se reconoce. En el espacio de características de la voz, la técnica CMLLR (o fMLLR) se traduce en la siguiente transformación afín

$$\tilde{\boldsymbol{x}}_t^{(r)} = \boldsymbol{A}^{(r)^{-1}}\boldsymbol{x}_t + \boldsymbol{A}^{(r)^{-1}}\boldsymbol{b}^{(r)}. \tag{2.30}$$

Como puede observarse, existe un vector de características transformado $\tilde{\boldsymbol{x}}_t^{(r)}$ para cada una de las clases de regresión. Esto equivale a propagar sobre el reconocedor r flujos de datos, uno por cada clase de regresión. Transformado el vector de características, la probabilidad de observación de éste dada la gaussiana k se puede calcular eficientemente como

$$\log \mathcal{N}\left(\boldsymbol{x}; \hat{\boldsymbol{\mu}}^{(k)}, \hat{\boldsymbol{\Sigma}}^{(k)}\right) = \log \mathcal{N}\left(\tilde{\boldsymbol{x}}^{(r)}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}\right) + \log |\boldsymbol{A}^{(r)^{-1}}|, \tag{2.31}$$

es decir, la probabilidad de observación es aquella calculada sobre el vector transformado usando los parámetros originales de la PDF, más un termino de normalización que depende de la matriz de transformación empleada.

De nuevo, los parámetros de la transformación CMLLR se estiman usando el algoritmo EM sobre los datos de adaptación y la transcripción de estos. Hay que hacer notar que en los sistemas no supervisados no se dispone de tal transcripción, sino que se emplea una transcripción hipotética obtenida por el propio reconocedor. Esta transcripción, por consiguiente, puede contener errores que afectarán al cálculo de los parámetros de la transformación CMLLR. No obstante, la técnica CMLLR ha demostrado ser bastante robusta a estos errores [266].

Una extensión de la técnica MLLR especialmente diseñada para situaciones en las que se dispone de muy pocos datos de adaptación es la conocida como MAPLR (*Maximum A Posteriori Linear Regression*, regresión lineal mediante máximo a posteriori) [51, 246]. Cuando no hay suficientes datos de adaptación disponibles, las transformaciones que MLLR calcula pueden distorsionar el modelo acústico. Para evitar este problema, la técnica MAPLR introduce una distribución a priori sobre los parámetros de la transformación afín. De esta forma, la transformación $\hat{\boldsymbol{W}}^{(r)}$ empleada por la técnica MAPLR es aquella que maximiza la siguiente probabilidad:

$$\hat{\boldsymbol{W}}^{(r)} = \underset{\boldsymbol{W}}{\operatorname{argmax}}\, p\left(\boldsymbol{X}|\boldsymbol{W}, \Theta, \mathcal{H}\right) p\left(\boldsymbol{W}|\Phi^{(r)}\right), \tag{2.32}$$

donde $p(\boldsymbol{W}|\Phi^{(r)})$ es la distribución a priori sobre los parámetros de la transformación. Esta distribución se define en base a los hiperparámetros $\Phi^{(r)}$, de la clase de regresión r.

Al igual que ocurría en la técnica MAP, la elección de la distribución a priori cobra una especial importancia en la técnica MAPLR. En [51] se propone utilizar una PDF gaussiana como distribución a priori $p(\boldsymbol{W}|\Phi^{(r)})$. Los hiperparámetros $\Phi^{(r)}$ de esta distribución (media y matriz de covarianza) se calculan en este caso usando un enfoque empírico Bayesiano, esto es, a partir de un conjunto de matrices de transformación conocidas, se estiman los hiperparámetros de la distribución utilizando un criterio ML.

Comparativa. Para concluir este apartado, dedicaremos unos últimos comentarios a comentar las ventajas e inconvenientes de las principales técnicas que se han estudiado, a saber, MAP y MLLR. Como se vio antes, podemos considerar que la adaptación MAP supone una combinación de unos modelos acústicos bien entrenados pero que no modelan fielmente los datos de adaptación, y de unos parámetros estimados a partir de un volumen limitado de datos de adaptación. Cuando la cantidad de datos de adaptación es limitada, la técnica MLLR genera unos modelos adaptados más precisos

que los de la técnica MAP [147]. Esto es coherente con el modo en el que se adaptan los modelos en cada técnica: mientras que la técnica MLLR estima una transformación global que se aplica a todas las PDFs del modelo, MAP adapta una por una las PDFs en base a los datos observados. No obstante, en aquellas situaciones en las que se disponga de un número suficientemente alto de datos de adaptación, la técnica MAP proporcionará mejores resultados que la técnica MLLR.

2.2.2.2. Adaptación basada en modelos de distorsión

Como se ha visto, las técnicas de adaptación de modelos presentadas en el apartado anterior (MAP, MLLR y sus variantes) no hacen consideración alguna sobre la forma en la que el ruido modifica las características de voz, sino que estiman una transformación lineal (MLLR) o los parámetros de la distribución de voz ruidosa (MAP) usando un conjunto de datos de adaptación. Por un lado, esta forma de proceder les brinda una gran flexibilidad, pudiendo ser aplicadas tanto a adaptación de locutor como a adaptación frente al ruido acústico. Asimismo, estas técnicas no requieren de un conocimiento preciso del ruido, sino tan sólo de un conjunto de datos de adaptación ruidosos con sus transcripciones asociadas. No obstante, esta flexibilidad implica una mayor necesidad de datos para estimar de forma robusta los parámetros del modelo adaptado. Por otra parte, podemos considerar que en la práctica es difícil disponer de transcripciones fiables de las frases que se reconocen: de hecho ése es el objetivo último del reconocedor.

Frente a las técnicas anteriores, en la literatura podemos encontrar un segundo grupo de técnicas especialmente diseñadas para adaptar los modelos acústicos a ambientes ruidosos. Estas técnicas emplean modelos analíticos de distorsión, como el presentado en la sección 2.1.1, para derivar las expresiones que permiten adaptar los parámetros del modelo acústico al entorno de reconocimiento. Recordemos que estos los modelos de distorsión permiten relacionar el vector de voz ruidosa $\boldsymbol{y}$ con el vector de voz limpia $\boldsymbol{x}$ y los vectores de ruido aditivo $\boldsymbol{n}$ y de canal $\boldsymbol{h}$. Usando esta relación analítica y los modelos estimados a priori para las variables aleatorias independientes ($\boldsymbol{x}$, $\boldsymbol{n}$ y $\boldsymbol{h}$), es posible calcular los parámetros de $p(\boldsymbol{y})$, siendo éste el modelo acústico con el que se reconoce. No obstante, dependiendo del dominio en el que se trabaje, es posible que la relación entre las variables aleatorias anteriores dada por el modelo de distorsión no sea lineal (p.ej. en los dominios log-Mel y cepstral). Esto, en la práctica, puede suponer que las expresiones de adaptación obtenidas no dispongan de solución analítica. En los siguientes puntos estudiaremos distintas aproximaciones para resolver este problema, lo que dará lugar a distintos métodos de adaptación.

Antes de pasar a la exposición de dichos métodos debemos mencionar que, aunque

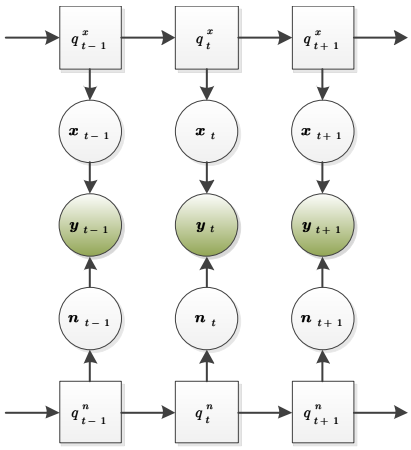

Figura 2.7: Combinación de modelos acústicos: las observaciones de voz ruidosa $\boldsymbol{y}_t$ se obtienen combinando los vectores de características de voz $\boldsymbol{x}_t$ y ruido $\boldsymbol{n}_t$, los cuales, a su vez, son generados por el modelo de voz y el modelo de ruido, respectivamente.

las propuestas de esta tesis se centran en la compensación de las características de voz, las técnicas de adaptación descritas en este apartado también tienen su equivalente como técnica de compensación de características. En su variante de compensación estas técnicas serán ampliamente utilizadas y referenciadas a lo largo de este trabajo. Aunque implique una carga de trabajo mayor, para alcanzar una visión más global de estas técnicas se ha preferido describirlas inicialmente como técnicas de adaptación. Posteriormente, en la sección 2.2.3.3, se indicarán brevemente las diferencias principales entre ambas versiones (adaptación y compensación).

Descomposición de modelos. Consideremos la red bayesiana dinámica representada en la figura 2.7. En esta red se ha supuesto que los modelos estadísticos que representan la voz y el ruido son conocidos a priori (p.ej. HMMs). Dados estos modelos, la evolución temporal de cada fuente (voz o ruido) puede determinarse a través de la secuencia temporal $t = 1, \ldots, T$ de los estados de voz q_t^x y ruido q_t^n de los modelos subyacentes. Estos estados, a su vez, generarán con una probabilidad dada los vectores de características no observados de voz $\boldsymbol{x}_t$ y ruido $\boldsymbol{n}_t$, que finalmente se combinan para formar la observación de voz ruidosa $\boldsymbol{y}_t$.

La idea recogida en la red bayesiana de la figura 2.7 puede aplicarse directamente a la etapa de descodificación de la voz dando lugar a lo que se conoce como descomposición de modelos [267]. Esta técnica supone que la voz y el ruido pueden modelarse de forma independiente mediante sendos HMMs $\mathcal{M}_x$ y $\mathcal{M}_n$, con M_x y M_n estados, respectivamente. A partir de estos dos modelos, es fácil ver que el modelo de voz rui-

dosa $\mathcal{M}_y$ se puede expresar de forma factorizada como el producto cartesiano de $\mathcal{M}_x$ y $\mathcal{M}_n$: para cada estado del modelo de voz $\mathcal{M}_x$ pueden darse M_n posibles estados del modelo de ruido, resultando, por tanto, en un total de $M_y = M_x \times M_n$ estados para $\mathcal{M}_y$. En lugar construir el modelo $\mathcal{M}_y$ de forma explícita, la técnica de descomposición de modelos modifica el algoritmo de Viterbi para que busque el camino de máxima probabilidad entre todos aquellos pares (s_i, s_j) de estados de voz y ruido,

$$p_t(s_i, s_j) = \max_{s_u, s_v} p_{t-1}(s_u, s_v) a^x_{u,i} a^n_{v,j} p(\boldsymbol{y}|s_i, s_j), \tag{2.33}$$

donde $p_t(s_i, s_j)$ es la probabilidad en el instante de tiempo t de que el estado del modelo de voz sea el i-ésimo ($q^x_t = s_i$) y que el de ruido sea el j-ésimo ($q^n_t = s_j$), $a^x_{u,i}$ es la probabilidad de transición entre los estados s_u y s_i del modelo de voz, $a^n_{v,j}$ es, de forma análoga, la probabilidad de transicionar del estado s_v al estado s_j del modelo $\mathcal{M}_n$ y, por último, $p(\boldsymbol{y}|s_i, s_j)$ es la probabilidad de observar $\boldsymbol{y}$ supuesto que los estados de voz y ruido en el instante de tiempo t son $q^x_t = s_i$ y $q^n_t = s_j$, respectivamente.

El único término desconocido en (2.33) es la probabilidad de observación $p_y(\boldsymbol{y}|s_i, s_j)$. De forma genérica esta probabilidad se puede obtener marginalizando la probabilidad conjunta $p(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{y}|s_i, s_j)$ sobre las variables ocultas $\boldsymbol{x}$ y $\boldsymbol{n}$:

$$p(\boldsymbol{y}|s_i, s_j) = \iint p(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{y}|s_i, s_j) d\boldsymbol{x} d\boldsymbol{n} = \iint p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n}) p(\boldsymbol{x}|s_i) p(\boldsymbol{n}|s_j) d\boldsymbol{x} d\boldsymbol{n}, \tag{2.34}$$

donde se ha supuesto que $\boldsymbol{y}$ es independiente de los estados de voz y ruido supuesto que $\boldsymbol{x}$ y $\boldsymbol{n}$ son conocidos. En esta ecuación $p(\boldsymbol{x}|s_i)$ y $p(\boldsymbol{n}|s_j)$ son directamente computables, mientras que en el cálculo de $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n})$ influirá la aproximación que se haga del modelo de distorsión definido en la ecuación (2.9). Esta cuestión será discutida con mayor detalle en el siguiente punto.

La técnica de descomposición de modelos tiene como ventaja su generalidad, siendo aplicable siempre y cuando se dispongan de modelos estadísticos para cada fuente (voz y ruido en nuestro caso), y, asimismo, se pueda modelar la interacción entre las distintas fuentes. De ahí que numerosos problemas en el campo del procesado y/o reconocimiento de señales hayan adoptado esta técnica para su resolución. Un caso ilustrativo de ello lo tenemos en el problema de reconocimiento multilocutor donde se dispone de una grabación, generalmente monocanal, con la voz de varios locutores y se pretende descodificar el mensaje emitido por cada uno de ellos. Varios son los trabajos [229, 232, 233] que, partiendo de modelos acústicos adaptados a la voz de cada locutor, han aplicado la técnica de descomposición de modelos a este problema con resultados prometedores.

Como principales problemas de la técnica de descomposición de modelos podemos mencionar dos: su alto coste computacional y la dificultad para estimar el modelo de

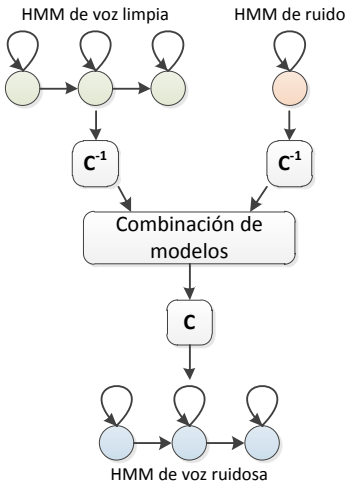

Figura 2.8: Esquema de la técnica PMC.

ruido. No obstante, como se estudiará en la sección 4.3, estos problemas se reducen mediante una aplicación cuidadosa de las ideas en las que se fundamente esta técnica. En concreto, en dicha sección se propondrá una técnica de compensación de características basada en la idea de combinar (o, de forma alternativa, descomponer) modelos acústicos independientes. Además de ser más eficiente que la técnica de descomposición de modelos, la técnica propuesta en la sección 4.3 contará como gran ventaja el permitir estimar de forma no supervisada el modelo $\mathcal{M}_n$ para cada elocución de evaluación.

Combinación de modelos paralelos. Otra técnica de adaptación similar a la de descomposición de modelos es la denominada como PMC (*Parallel Model Combination*, combinación de modelos paralelos) [97, 103, 105]. En la figura 2.8 podemos ver el esquema básico de la técnica PMC: supuesto que se dispone de modelos que representan las estadísticas de la voz limpia y del ruido en el dominio cepstral (p.ej. HMMs), la técnica PMC los combina para dar lugar a un modelo de voz ruidosa. Como veíamos en las ecuaciones (2.7) y (2.9), la interacción entre la voz y el ruido se expresa de forma más natural en el dominio log-Mel o en el dominio lineal del banco de filtros, de ahí que la combinación de los modelos se lleve a cabo en uno de estos dominios y, finalmente, el HMM resultante se transforme al dominio cepstral mediante el operador lineal $\boldsymbol{C}$ que implementa la DCT.

Como vemos, la mayor diferencia entre PMC y la técnica de descomposición de modelos propuesta por Varga y Moore en [267], radica en que PMC obtiene explícitamente el modelo de voz distorsionada $\mathcal{M}_y$, mientras que en la técnica de descomposición de modelos esta tarea se lleva a cabo implícitamente mediante un algoritmo de reco-

nocimiento modificado. En la figura 2.8 también aparece reflejada otra consideración práctica que se suele hacer respecto al modelo de ruido: dado que en situaciones realistas éste se debe estimar a partir de los datos de test, se elige un modelo simple para representar al ruido, generalmente una única gaussiana.

Entrando en detalle, en primer lugar PMC transforma los parámetros de las PDFs de los modelos de voz (y ruido) del dominio cepstral al dominio log-Mel de la siguiente forma,

$$\boldsymbol{\mu} = \boldsymbol{C}^{-1}\boldsymbol{\mu}^c, \tag{2.35}$$

$$\boldsymbol{\Sigma} = \boldsymbol{C}^{-1}\boldsymbol{\Sigma}^c\left(\boldsymbol{C}^{-1}\right)^\top, \tag{2.36}$$

donde $\boldsymbol{\mu}$ y $\boldsymbol{\mu}^c$ son dos vectores de media expresados en el dominio log-Mel y el dominio cepstral, respectivamente. Asimismo, $\boldsymbol{\Sigma}$ y $\boldsymbol{\Sigma}^c$ son las matrices de covarianza de una distribución normal multivariante expresadas en los dominios log-Mel y cepstral ($\boldsymbol{C}^{-1}$ denota la IDCT).

Una vez transformados los parámetros de las gaussianas al dominio log-Mel, el siguiente paso de la técnica PMC es combinar los modelos de voz y ruido. Si observamos la expresión analítica del modelo de distorsión de la voz definido en la ecuación (2.11), vemos que dicha expresión es no lineal respecto a $\boldsymbol{x}$ y $\boldsymbol{n}$, lo cual implica que la combinación de la PDF de voz $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ con la PDF de ruido $\mathcal{N}(\boldsymbol{n}; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ va a resultar en una PDF no gaussiana. Las distintas versiones de la técnica PMC afrontan este problema de forma diferente. En primer lugar tenemos la versiones básicas no iterativas en las que se realizan distintas aproximaciones para obtener una forma analítica de la PDF combinada. Dos han sido las aproximaciones que se han empleado con mayor éxito: aproximación *log-add* y aproximación *log-normal*. En la aproximación *log-add* se asume que la varianza del ruido es suficientemente pequeña respecto a la varianza de la voz, por lo que sólo se modifican las medias de las distribuciones de voz limpia. La fórmula de adaptación utilizada se obtiene aproximando la ecuación (2.11) mediante expansión por series de Taylor vectoriales de orden cero, tomando como punto de expansión las medias de la voz y el ruido,

$$\boldsymbol{\mu}_y^c \approx \boldsymbol{\mu}_x^c + \boldsymbol{C}\log\left(\mathbf{1} + e^{\boldsymbol{\mu}_n - \boldsymbol{\mu}_x}\right). \tag{2.37}$$

Por otra parte, si suponemos que las distribuciones de voz y ruido se distribuyen según una distribución normal en el dominio log-Mel, dichas distribuciones serán log-normales en el dominio lineal del banco de filtros. Se podría pensar en sumar dichas distribuciones en dicho dominio, pero desafortunadamente la suma de dos variables aleatorias distribuidas según la distribución log-normal no necesariamente resulta en

otra distribución log-normal. No obstante, la aproximación *log-normal* asume dicha distribución con los siguientes parámetros:

$$\boldsymbol{\mu}_y^l \approx \boldsymbol{\mu}_x^l + \boldsymbol{\mu}_n^l, \tag{2.38}$$

$$\boldsymbol{\Sigma}_y^l \approx \boldsymbol{\Sigma}_x^l + \boldsymbol{\Sigma}_n^l, \tag{2.39}$$

donde el superíndice l indica que los parámetros se expresan en el dominio lineal del banco de filtros.

La relación entre los parámetros $(\boldsymbol{\mu}^l, \boldsymbol{\Sigma}^l)$ de una distribución log-normal y su distribución gaussiana asociada $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ viene dada por [97],

$$\mu_i^l = \exp\left(\mu_i + \frac{\Sigma_{ii}}{2}\right), \tag{2.40}$$

$$\Sigma_{ij}^l = \mu_i \mu_j \left[\exp\left(\Sigma_{ij}\right) - 1\right]. \tag{2.41}$$

Finalmente, los parámetros de la distribución de voz ruidosa $\mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$ en el dominio log-Mel se obtienen de acuerdo a las siguientes expresiones

$$\mu_{y,i} \approx \log\left(\mu_{y,i}^l\right) - \frac{1}{2}\log\left(\frac{\Sigma_{ii}^l}{\left(\mu_{y,i}^l\right)^2} + 1\right), \tag{2.42}$$

$$\Sigma_{y,ij}^l \approx \log\left(\frac{\Sigma_{y,ij}^l}{\mu_{y,i}^l \mu_{y,j}^l} + 1\right). \tag{2.43}$$

Otro método alternativo para el cálculo de la distribución de voz distorsionada es aquél propuesto por la técnica DPMC (*Data-driven* PMC, PMC basado en datos). En lugar de recurrir a aproximaciones como las que hemos estudiado para la técnica PMC básica, DPMC aproxima la distribución de voz ruidosa teórica por la gaussiana que minimiza la divergencia de Kullback-Leibler (KL) con ésta. Formalmente, sea $p(\boldsymbol{y}|k_x, \mathcal{M}_x, k_n, \mathcal{M}_n)$ la distribución de voz distorsionada obtenida a partir de muestras de voz generadas por la componente (gaussiana) k_x-ésima del modelo de voz limpia $\mathcal{M}_x$ y combinadas con muestras de ruido generadas por la componente k_n-ésima del modelo de ruido $\mathcal{M}_n$. De forma genérica esta distribución se puede expresar como

$$\begin{aligned} p(\boldsymbol{y}|k_x, \mathcal{M}_x, k_n, \mathcal{M}_n) &= \iint p(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{y}|k_x, \mathcal{M}_x, k_n, \mathcal{M}_n) d\boldsymbol{x} d\boldsymbol{n} \\ &= \iint p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n}) p(\boldsymbol{x}|k_x, \mathcal{M}_x) p(\boldsymbol{n}|k_n, \mathcal{M}_n) d\boldsymbol{x} d\boldsymbol{n} \\ &= \iint \delta_{\boldsymbol{f}(\boldsymbol{x},\boldsymbol{n})}(\boldsymbol{y}) p(\boldsymbol{x}|k_x, \mathcal{M}_x) p(\boldsymbol{n}|k_n, \mathcal{M}_n) d\boldsymbol{x} d\boldsymbol{n}, \end{aligned} \tag{2.44}$$

siendo $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{n})$ el modelo de distorsión simplificado de la ecuación (2.11) y $\delta_{\boldsymbol{z}}(\boldsymbol{y})$ es la función delta de Dirac centrada en el punto $\boldsymbol{z}$.

Asimismo, sea $q^*(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$ la función de densidad de probabilidad por la se desea aproximar $p(\boldsymbol{y}|k_x, \mathcal{M}_x, k_n, \mathcal{M}_n)$. De acuerdo a la divergencia KL, $q^*(\boldsymbol{y})$ será aquella PDF que satisfaga la siguiente expresión:

$$\begin{aligned} q^*(\boldsymbol{y}) &= \operatorname*{argmin}_q \mathcal{KL}(p||q) \\ &= \operatorname*{argmin}_q \int p(\boldsymbol{y}|k_x, \mathcal{M}_x, k_n, \mathcal{M}_n) \log \frac{p(\boldsymbol{y}|k_x, \mathcal{M}_x, k_n, \mathcal{M}_n)}{q(\boldsymbol{y})} d\boldsymbol{y} \\ &= \operatorname*{argmax}_q \int p(\boldsymbol{y}|k_x, \mathcal{M}_x, k_n, \mathcal{M}_n) \log q(\boldsymbol{y}) d\boldsymbol{y}. \end{aligned} \tag{2.45}$$

Para calcular $q^*(\boldsymbol{y})$, la técnica DPMC aproxima la distribución teórica de la voz ruidosa mediante una versión muestreada $\hat{p}(\boldsymbol{y}|k_x, \mathcal{M}_x, k_n, \mathcal{M}_n)$ obtenida usando el método de Montecarlo sobre las distribuciones marginales $p(\boldsymbol{x}|k_x, \mathcal{M}_x)$ y $p(\boldsymbol{n}|k_n, \mathcal{M}_n)$. Sean $(\boldsymbol{y}^{(1)}, \cdots, \boldsymbol{y}^{(n)})$ las n muestras por las que se aproxima la PDF. Al aplicar la divergencia KL sobre esta distribución muestreada resulta en la siguiente función a optimizar:

$$\begin{aligned} q^*(\boldsymbol{y}) &= \operatorname*{argmin}_q \mathcal{KL}(\hat{p}||q) \\ &= \operatorname*{argmax}_q \sum_{i=1}^{n} \log q(\boldsymbol{y}^{(i)}), \end{aligned} \tag{2.46}$$

lo que equivale a estimar los parámetros de $q^*(\boldsymbol{y})$ según el criterio ML (*Maximum Likelihood*, máxima verosimilitud) a partir de las n muestras por las que se ha aproximada la distribución teórica.

Para una distribución normal, es bien conocido que la estimación ML de la media y la varianza viene dada por

$$\boldsymbol{\mu}_y^* = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}^{(i)}, \tag{2.47}$$

$$\boldsymbol{\Sigma}_y^* = \left(\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}^{(i)} \boldsymbol{y}^{(i)T} \right) - \boldsymbol{\mu}_y^*. \tag{2.48}$$

El desarrollo anterior de las técnicas PCM y DPCM se ha centrado en las componentes estáticas de los vectores de características. No obstante, en la literatura se han propuesto extensiones de estas técnicas para adaptar también las componentes dinámicas (características Δ y Δ^2). En [97] se puede encontrar más información al respecto.

Desarrollo en series de Taylor vectoriales. El objetivo de la técnica VTS es estimar la función de densidad de probabilidad de la voz ruidosa, supuesto que la PDF

de la voz limpia, del ruido aditivo y del ruido convolutivo son conocidas. Una vez estimada $p(\boldsymbol{y})$, ésta se puede emplear para descodificar la voz ruidosa (VTS aplicado a la adaptación de modelos) o para compensar el ruido de las características de voz (VTS aplicado a la compensación de la voz). Sea como fuere, para el cálculo de $p(\boldsymbol{y})$ la técnica VTS explota el modelo de distorsión de la voz presentado en la sección 2.1.1. En particular, si el dominio en el que se expresan los características de voz es el cepstral, la ecuación empleada es la (2.12), la cual volvemos a reescribir aquí por conveniencia,

$$\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{h}) = \boldsymbol{x} + \boldsymbol{h} + \boldsymbol{C} \log\left(\mathbf{1} + e^{\boldsymbol{C}^{-1}(\boldsymbol{n}-\boldsymbol{x}-\boldsymbol{h})}\right), \tag{2.49}$$

en donde se han omitido los superíndices presentes en (2.12) que indican que el modelo se expresa en el dominio cepstral.

Como se puede apreciar, el modelo es no lineal en sus parámetros, lo que dificulta su aplicación a la hora de obtener expresiones de adaptación manejables. La solución que VTS aporta consiste en aproximar (2.49) mediante una función lineal obtenida mediante desarrollo en serie de Taylor. Recordemos que la expansión por serie de Taylor de una función $g(x)$ infinitamente derivable en torno a un punto a se define como

$$g(x) = \sum_{n=0}^{\infty} \frac{g^{(n)}(a)}{n!} (x-a)^n, \tag{2.50}$$

donde $n!$ representa el factorial de n y $g^{(n)}(a)$ es la derivada n-ésima de la función g evaluada en $x = a$ (la derivada de orden 0 es la propia función g).

Para el caso de funciones con d variables, p.ej. $g(\boldsymbol{x}) = g(x_1, x_2, \cdots, x_d)$, el desarrollo en serie de Taylor alrededor del punto $\boldsymbol{a} = (a_1, \cdots, a_d)$ se define como

$$g(\boldsymbol{x}) = \sum_{n_1=0}^{\infty} \cdots \sum_{n_d=0}^{\infty} \frac{\partial^{n_1}}{\partial x_1^{n_1}} \cdots \frac{\partial^{n_d}}{\partial x_d^{n_d}} \frac{g(\boldsymbol{a})}{n_1! \cdots n_d!} (x_1 - a_1)^{n_1} \cdots (x_d - a_d)^{n_d}. \tag{2.51}$$

Usando las expresiones anteriores, se puede comprobar que la aproximación de Taylor de primer orden de la función $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{h})$ definida en (2.49) usando como punto de expansión $(\boldsymbol{x}_0^\top, \boldsymbol{n}_0^\top, \boldsymbol{h}_0^\top)^\top$ viene dada por [9],

$$\boldsymbol{y}_{\text{vts-1}} = \boldsymbol{f}(\boldsymbol{x}_0, \boldsymbol{n}_0, \boldsymbol{h}_0) + \boldsymbol{J}_x^{(k)} (\boldsymbol{x} - \boldsymbol{x}_0) + \boldsymbol{J}_n^{(k)} (\boldsymbol{n} - \boldsymbol{n}_0) + \boldsymbol{J}_h^{(k)} (\boldsymbol{h} - \boldsymbol{h}_0), \tag{2.52}$$

donde $\boldsymbol{J}_x^{(k)}$, $\boldsymbol{J}_n^{(k)}$ y $\boldsymbol{J}_h^{(k)}$ son las matrices jacobianas respecto a la voz, al ruido aditivo y al convolutivo, respectivamente, del modelo de distorsión. La dependencia de estas matrices respecto al índice k de las gaussianas del modelo acústico será discutido más

adelante. Estas matrices pueden obtenerse del siguiente modo (ver p.ej. [9]),

$$\boldsymbol{J}_x^{(k)} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0,\boldsymbol{n}_0,\boldsymbol{h}_0} = \begin{bmatrix} \left.\frac{\partial y_1}{\partial x_1}\right|_{\boldsymbol{x}_0,\boldsymbol{n}_0,\boldsymbol{h}_0} & \cdots & \left.\frac{\partial y_1}{\partial x_d}\right|_{\boldsymbol{x}_0,\boldsymbol{n}_0,\boldsymbol{h}_0} \\ \vdots & \ddots & \vdots \\ \left.\frac{\partial y_d}{\partial x_1}\right|_{\boldsymbol{x}_0,\boldsymbol{n}_0,\boldsymbol{h}_0} & \cdots & \left.\frac{\partial y_d}{\partial x_d}\right|_{\boldsymbol{x}_0,\boldsymbol{n}_0,\boldsymbol{h}_0} \end{bmatrix} = \boldsymbol{C}\boldsymbol{F}\boldsymbol{C}^{-1}, \tag{2.53}$$

$$\boldsymbol{J}_n^{(k)} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{n}}\right|_{\boldsymbol{x}_0,\boldsymbol{n}_0,\boldsymbol{h}_0} = \boldsymbol{I} - \boldsymbol{C}\boldsymbol{F}\boldsymbol{C}^{-1}, \tag{2.54}$$

$$\boldsymbol{J}_h^{(k)} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{h}}\right|_{\boldsymbol{x}_0,\boldsymbol{n}_0,\boldsymbol{h}_0} = \boldsymbol{J}_x^{(k)}, \tag{2.55}$$

donde $\boldsymbol{I}$ es la matriz identidad de dimensión d y $\boldsymbol{F}$ es una matriz diagonal cuyos elementos se calculan de la siguiente forma

$$\boldsymbol{F} = \operatorname{diag}\left(\frac{1}{1+\exp\left(\boldsymbol{C}^{-1}\left(\boldsymbol{n}_0 - \boldsymbol{x}_0 - \boldsymbol{h}_0\right)\right)}\right). \tag{2.56}$$

Para un desarrollo de orden cero, la aproximación del modelo de distorsión obtenida es

$$\boldsymbol{y}_{\text{vts-0}} = \boldsymbol{f}\left(\boldsymbol{x}_0, \boldsymbol{n}_0, \boldsymbol{h}_0\right). \tag{2.57}$$

Como podemos ver, tanto para el desarrollo de orden cero de la ecuación anterior, como el de primer orden reflejado en (2.52), el modelo de distorsión no lineal de la ecuación (2.49) se aproxima mediante un polinomio. Para la aproximación de orden cero dada en la ecuación anterior, dicho polinomio se reduce a un vector constante igual a $\boldsymbol{f}(\boldsymbol{x}_0, \boldsymbol{n}_0, \boldsymbol{h}_0)$. La aproximación de primer orden de la ecuación (2.52) puede expresarse en una forma más conveniente como,

$$\boldsymbol{y}_{\text{vts-1}} = \boldsymbol{J}_x^{(k)}\boldsymbol{x} + \boldsymbol{J}_n^{(k)}\boldsymbol{n} + \boldsymbol{J}_h^{(k)}\boldsymbol{h} + \underbrace{\boldsymbol{f}\left(\boldsymbol{x}_0, \boldsymbol{n}_0, \boldsymbol{h}_0\right) - \boldsymbol{J}_x^{(k)}\boldsymbol{x}_0 - \boldsymbol{J}_n^{(k)}\boldsymbol{n}_0 - \boldsymbol{J}_h^{(k)}\boldsymbol{h}_0}_{\boldsymbol{b}} \tag{2.58}$$

Una cuestión importante en la técnica VTS es la elección del punto $(\boldsymbol{x}_0^\top, \boldsymbol{n}_0^\top, \boldsymbol{h}_0^\top)^\top$ en torno al cual se aproxima el modelo de distorsión y, por tanto, alrededor del cual la aproximación VTS es válida. En nuestro caso, el objetivo es modificar los parámetros del modelo acústico para que estos representen de forma más precisa el entorno acústico de evaluación. Sin pérdida de generalidad, podemos decir que el objetivo de la adaptación es obtener una distribución adaptada $\mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_y^{(k)}, \boldsymbol{\Sigma}_y^{(k)})$ para cada componente $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x^{(k)}, \boldsymbol{\Sigma}_x^{(k)})$ del modelo acústico, de forma que se modele mejor (en el sentido del riesgo de Bayes) la unidad acústica correspondiente en el espacio de las características de la voz distorsionadas. Por tanto, a la hora de adaptar dicha componente se escoge como valor de expansión $\boldsymbol{x}_0$ la media de dicha gaussiana, esto es, $\boldsymbol{x}_0 = \boldsymbol{\mu}_x^{(k)}$. De ahí que

las matrices jacobianas sean función del componente k a partir del cual se aproxima el modelo de distorsión.

El ruido aditivo $\boldsymbol{h}$ y el de canal $\boldsymbol{h}$ suelen ser desconocidos a priori, no obstante, en la práctica pueden estimarse de una u otra forma[1]. En tal caso, es posible caracterizar los ruidos mediante sendas distribuciones $\boldsymbol{h} \sim \mathcal{N}(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$ y $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$, donde las medias corresponden al valor de ruido estimado y las matrices de covarianza indican el error esperado del estimador empleado. Asimismo, es también habitual disponer de una estimación de cada ruido para cada instante de tiempo. No obstante, y por simplicidad en la notación, omitimos la dependencia respecto al tiempo en las PDFs que modelan los ruidos. En base a estas distribuciones, y de forma análoga al caso de la voz, podemos escoger las medias de estas distribuciones como puntos de expansión para el desarrollo en serie de Taylor, es decir, $\boldsymbol{n}_0 = \boldsymbol{\mu}_n$ y $\boldsymbol{h}_0 = \boldsymbol{\mu}_h$.

Para el cálculo de los parámetros del modelo acústico adaptado consideraremos la propiedad de combinación lineal de las variables aleatorias (v.a.) normales [216]: dado que $\boldsymbol{x}$, $\boldsymbol{n}$ y $\boldsymbol{h}$ son v.a. normales, cualquier combinación lineal de las mismas también lo será. En particular, según la aproximación VTS de la ecuación (2.58), $\boldsymbol{y}$ es una combinación lineal de las variables anteriores. Por tanto, $p(\boldsymbol{y}|k)$ será una distribución gaussiana con los siguientes parámetros:

$$p(\boldsymbol{y}|k) = \mathcal{N}\left(\boldsymbol{y}; \boldsymbol{\mu}_y^{(k)}, \boldsymbol{\Sigma}_y^{(k)}\right), \tag{2.59}$$

donde la media viene dada por

$$\boldsymbol{\mu}_y^{(k)} = \boldsymbol{f}\left(\boldsymbol{\mu}_x^{(k)}, \boldsymbol{\mu}_n, \boldsymbol{\mu}_h\right) = \boldsymbol{\mu}_x^{(k)} + \boldsymbol{\mu}_h + \boldsymbol{C}\log\left(\mathbf{1} + e^{\boldsymbol{C}^{-1}\left(\boldsymbol{\mu}_n - \boldsymbol{\mu}_x^{(k)} - \boldsymbol{\mu}_h\right)}\right) \tag{2.60}$$

y la matriz de covarianza por

$$\boldsymbol{\Sigma}_y^{(k)} = \boldsymbol{J}_x^{(k)}\boldsymbol{\Sigma}_x^{(k)}\boldsymbol{J}_x^{(k)\top} + \boldsymbol{J}_n^{(k)}\boldsymbol{\Sigma}_n\boldsymbol{J}_n^{(k)\top} + \boldsymbol{J}_h^{(k)}\boldsymbol{\Sigma}_h\boldsymbol{J}_h^{(k)\top}. \tag{2.61}$$

En el caso de la aproximación VTS de orden cero, la media de la distribución $p(\boldsymbol{y}|k)$ es la dada en la ecuación (2.60), pero la matriz de covarianza de esta distribución coincide con la covarianza de la PDF limpia, es decir, $\boldsymbol{\Sigma}_y^{(k)} = \boldsymbol{\Sigma}_x^{(k)}$.

Antes de continuar con el desarrollo de la técnica VTS, merece la pena estudiar el comportamiento de los parámetros de $p(\boldsymbol{y}|k)$ en función de la energía relativa de las medias ($\boldsymbol{\mu}_x^{(k)}$, $\boldsymbol{\mu}_n$ y $\boldsymbol{\mu}_h$) que intervienen en el cálculo de dichos parámetros. Para simplificar el análisis, supondremos que la distorsión del canal es nula ($\boldsymbol{h} = \mathbf{0}$). Bajo

[1] Por ejemplo en el caso de $\boldsymbol{n}$ puede emplearse un detector de actividad de voz (VAD, *Voice Activity Detector*) que identifique los segmentos de silencio de la elocución.

esta simplificación, la media de la distribución ruidosa viene dada por

$$\boldsymbol{\mu}_y^{(k)} = \boldsymbol{\mu}_x^{(k)} + \boldsymbol{C}\log\left(\mathbf{1} + e^{\boldsymbol{C}^{-1}\left(\boldsymbol{\mu}_n - \boldsymbol{\mu}_x^{(k)}\right)}\right). \tag{2.62}$$

Como podemos observar, $\boldsymbol{\mu}_y^{(k)}$ se obtiene modificando el vector de voz limpia original, $\boldsymbol{\mu}_x^{(k)}$, con una cantidad en la que intervienen el propio vector de voz limpia y el vector de ruido, ambos expresados en el dominio log-Mel, ya que van multiplicados por el operador $\boldsymbol{C}^{-1}$. A esta cantidad se la denominó en la sección 2.1.1 función de discrepancia. Si la energía de $\boldsymbol{\mu}_x^{(k)}$ en el dominio log-Mel es mucho mayor que la de $\boldsymbol{\mu}_n$ (también en ese dominio), entonces la función de discrepancia tiende a cero y, por tanto, $\boldsymbol{\mu}_y^{(k)} \approx \boldsymbol{\mu}_x^{(k)}$. En el caso contrario, esto es, si la energía de $\boldsymbol{\mu}_x^{(k)}$ en el dominio log-Mel es pequeña, entonces $\boldsymbol{\mu}_y^{(k)}$ tiende a $\boldsymbol{\mu}_n$ como cabría esperar.

En el cálculo de la matriz de covarianza $\boldsymbol{\Sigma}_y^{(k)}$ de la ecuación (2.61) intervienen las matrices jacobianas $\boldsymbol{J}_x^{(k)}$, $\boldsymbol{J}_n^{(k)}$ y $\boldsymbol{J}_h^{(k)}$. En el cómputo de estas matrices participa, a su vez, la matriz $\boldsymbol{F}$ definida en la ecuación (2.56). Como se puede apreciar, la expresión para el cálculo de los elementos de esta matriz diagonal tiene forma de función sigmoide. Para ver esto más claro, consideremos el cálculo del elemento i-ésimo de esta matriz, esto es,

$$F_{ii} = \frac{1}{1 + \exp\left(\boldsymbol{c}_i^{-1}\left(\boldsymbol{\mu}_n - \boldsymbol{\mu}_x^{(k)}\right)\right)} = \frac{1}{1 + \exp\left(\mu_{n,i}^{log} - \mu_{x,i}^{(k)log}\right)}, \tag{2.63}$$

siendo $\boldsymbol{c}_i^{-1}$ la fila i-ésima de la matriz $\boldsymbol{C}^{-1}$.

Al encontrarnos ante una función sigmoide, sabemos que el valor de F_{ii} variará entre 0 y 1 en función de las energías relativas entre el ruido y la voz en el domino log-Mel. Si la energía del ruido es mucho mayor que la de la voz en el dominio log-Mel, entonces $F_{ii} \to 0$ y la matriz jacobiana $\boldsymbol{J}_x^{(k)}$ se anula, por lo que la matriz de covarianza $\boldsymbol{\Sigma}_y^{(k)} \approx \boldsymbol{\Sigma}_n$. Por otra parte, si la energía de la voz es mucho mayor que la del ruido (de nuevo en el dominio log-Mel), $F_{ii} \to 1$ y $\boldsymbol{J}_x^{(k)} \to \boldsymbol{I}$. En este segundo caso, el ruido no tiene ninguna influencia y los parámetros del modelo acústico permanecen inalterados durante el proceso de adaptación. Para niveles de SNR alrededor de 0 dB, es decir, cuando la energía de la voz y el ruido sean similares, el valor de F_{ii} será aproximadamente 0,5; por lo que tanto el ruido como la voz intervendrán en la adaptación.

En resumen, para niveles de SNR altos las componentes del modelo acústico apenas se ven alteradas, mientras que para niveles de SNR muy bajos sus parámetros (medias y covarianzas) se reemplazan por los del modelo de ruido $p(\boldsymbol{n})$. En este caso, durante el proceso de descodificación de la voz, la probabilidad de observación del vector de entrada $\boldsymbol{y}_t$ será la misma para todas las componentes de todos los estados del modelo

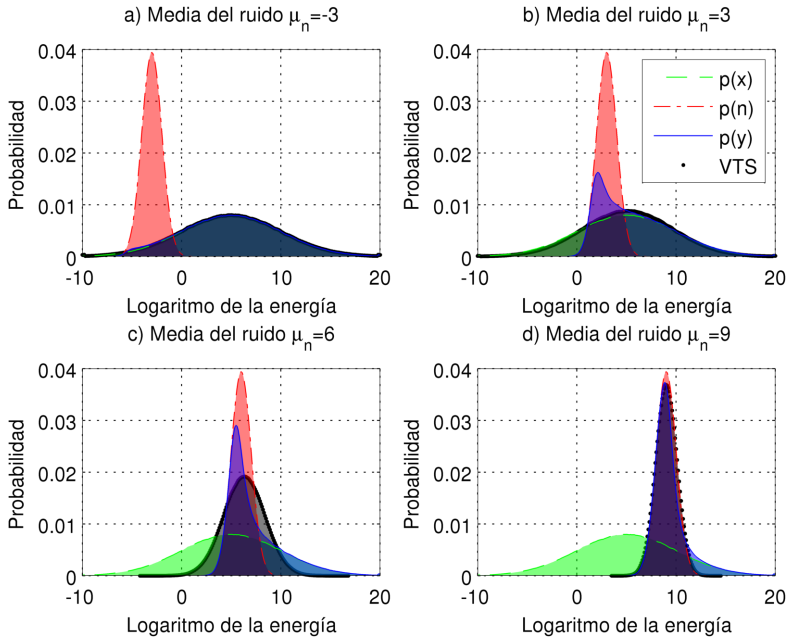

Figura 2.9: Ejemplos de la aproximación realizada por la técnica VTS de la distribución $p(y)$ para distintos niveles de ruido.

acústico, es decir, $p_y(\boldsymbol{y}_t|k) = p_n(\boldsymbol{y}_t)\ \forall k$. Luego la descodificación vendrá guiada únicamente por las probabilidades a priori de las componentes y por las probabilidades de transición entre estados.

La figura 2.9 muestra distintos ejemplos de las aproximaciones obtenidas por la técnica VTS para el caso unidimensional. En los ejemplos se ha asumido que el la distorsión del canal es nula y, por sencillez, que las variables aleatorias se expresan en el dominio del banco de filtros logarítmico en lugar del cepstrum. Asimismo, consideramos que las distribuciones de probabilidad $p(x)$ y $p(n)$ son gaussianas: $p(x) = \mathcal{N}_x(\mu_x = 5, \sigma_x = 5)$, mientras que $p(n)$ cuenta con una desviación típica fija de $\sigma_n = 1$ y cuya media oscila entre -3 y 9 con incrementos de 3 (gráficas a), b), c) y d) de la figura). A partir de estas distribuciones, se ha calculado la distribución teórica $p(y)$ mediante el método de Montecarlo usando el modelo de distorsión de la voz de la ecuación (2.11). Como se puede observar, la aproximación VTS de la distribución $p(y)$ es muy fiel para SNRs altas y bajas (gráficas a) y d) de la figura). Para SNRs intermedias (gráficas b) y c)), la aproximación VTS se aleja de la distribución teórica debido a la no gaussianidad de la segunda.

Para terminar con la descripción con la adaptación de modelos basada en VTS, veremos ahora cómo se pueden modificar los parámetros que modelan las caracterís-

ticas dinámicas de la voz. De forma estándar estas características se calculan usando diferencias simples o regresión lineal de los características estáticas en una ventana de tiempo dada [279]. Por ejemplo, asumiendo una ventana de tamaño 3, el cálculo de las componentes dinámicas en el instante de tiempo t, $\boldsymbol{y}_t^{\Delta}$, se puede expresar de forma genérica mediante la siguiente relación lineal

$$\boldsymbol{y}_t^{\Delta} = \boldsymbol{D}^{\Delta} \begin{bmatrix} \boldsymbol{y}_{t-1} \\ \boldsymbol{y}_t \\ \boldsymbol{y}_{t+1} \end{bmatrix} = \boldsymbol{D}^{\Delta} \begin{bmatrix} \boldsymbol{f}\left(\boldsymbol{x}_{t-1}, \boldsymbol{n}_{t-1}, \boldsymbol{h}_{t-1}\right) \\ \boldsymbol{f}\left(\boldsymbol{x}_t, \boldsymbol{n}_t, \boldsymbol{h}_t\right) \\ \boldsymbol{f}\left(\boldsymbol{x}_{t+1}, \boldsymbol{n}_{t+1}, \boldsymbol{h}_{t+1}\right) \end{bmatrix}, \tag{2.64}$$

donde $\boldsymbol{D}^{\Delta}$ es la matriz que implementa el cálculo de las características dinámicas y $\boldsymbol{f}(\cdot)$ es el modelo de distorsión de la voz de la ecuación (2.49).

El uso de la ecuación (2.64) para el cálculo de las componentes dinámicas, no obstante, dificulta la obtención de una expresión analítica para la adaptación de estos parámetros. Así, en la práctica la función de adaptación para estos parámetros se obtiene recurriendo a una aproximación de tiempo continuo [97, 128], en la cual se asume que las características dinámicas representan realmente la derivada respecto al tiempo (o derivada segunda para las características Δ^2) del modelo de distorsión de la voz,

$$\boldsymbol{y}_t^{\Delta} \simeq \frac{\partial \boldsymbol{f}\left(\boldsymbol{x}_t, \boldsymbol{n}_t, \boldsymbol{h}_t\right)}{\partial t}. \tag{2.65}$$

Combinando la aproximación de tiempo continuo y la aproximación del modelo de distorsión realizada por VTS en la ecuación (2.52), se puede demostrar que las características dinámicas de la voz ruidosa vienen dadas por la siguiente expresión (ver Apéndice B de [175] para más detalles),

$$\begin{aligned} \boldsymbol{y}_t^{\Delta} &\approx \frac{\partial \boldsymbol{y}}{\partial t} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial t} + \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{n}}\frac{\partial \boldsymbol{n}}{\partial t} + \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{h}}\frac{\partial \boldsymbol{h}}{\partial t} \\ &\approx \boldsymbol{J}_x^{(k)}\boldsymbol{x}_t^{\Delta} + \boldsymbol{J}_n^{(k)}\boldsymbol{n}_t^{\Delta} + \boldsymbol{J}_h^{(k)}\boldsymbol{h}_t^{\Delta}, \end{aligned} \tag{2.66}$$

siendo $\boldsymbol{x}_t^{\Delta}$, $\boldsymbol{n}_t^{\Delta}$ y $\boldsymbol{h}_t^{\Delta}$ las características dinámicas de la voz limpia, el ruido aditivo y el ruido de canal, respectivamente.

Generalmente el ruido de canal se supone constante ($\boldsymbol{h}_t^{\Delta} = 0\ \forall t$), luego la expresión anterior se simplifica como

$$\boldsymbol{y}_t^{\Delta} \approx \boldsymbol{J}_x^{(k)}\boldsymbol{x}_t^{\Delta} + \boldsymbol{J}_n^{(k)}\boldsymbol{n}_t^{\Delta}. \tag{2.67}$$

Siguiendo un procedimiento análogo, la expresión obtenida para las aceleraciones (derivadas segundas) es

$$\boldsymbol{y}_t^{\Delta^2} \approx \boldsymbol{J}_x^{(k)}\boldsymbol{x}_t^{\Delta^2} + \boldsymbol{J}_n^{(k)}\boldsymbol{n}_t^{\Delta^2}. \tag{2.68}$$

De nuevo el cálculo de las características dinámicas de la voz ruidosa se aproxima mediante una combinación lineal de las características dinámicas de la voz sin distorsionar y el ruido. Supuesto que estas últimas se modelan mediante distribuciones gaussianas, el resultado final será también una distribución gaussiana con los siguientes parámetros:

$$\boldsymbol{\mu}_y^{(k)} = \begin{bmatrix} \boldsymbol{\mu}_y^{(k)} \\ \boldsymbol{\mu}_y^{\Delta(k)} \\ \boldsymbol{\mu}_y^{\Delta^2(k)} \end{bmatrix} \approx \begin{bmatrix} \boldsymbol{\mu}_x^{(k)} + \boldsymbol{\mu}_h + \boldsymbol{C}\log\left(\mathbf{1} + e^{\boldsymbol{C}^{-1}\left(\boldsymbol{\mu}_n - \boldsymbol{\mu}_x^{(k)} - \boldsymbol{\mu}_h\right)}\right) \\ \boldsymbol{J}_x(k)\boldsymbol{\mu}_x^{\Delta(k)} + \boldsymbol{J}_n^{(k)}\boldsymbol{\mu}_n^{\Delta} \\ \boldsymbol{J}_x(k)\boldsymbol{\mu}_x^{\Delta^2(k)} + \boldsymbol{J}_n^{(k)}\boldsymbol{\mu}_n^{\Delta^2} \end{bmatrix}, \tag{2.69}$$

$$\boldsymbol{\Sigma}_y^{(k)} = \begin{bmatrix} \boldsymbol{\Sigma}_y^{(k)} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_y^{\Delta(k)} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Sigma}_y^{\Delta^2(k)} \end{bmatrix}, \tag{2.70}$$

donde

$$\boldsymbol{\Sigma}_y^{(k)} \approx \boldsymbol{J}_x^{(k)}\boldsymbol{\Sigma}_x^{(k)}\boldsymbol{J}_x^{(k)^\top} + \boldsymbol{J}_n^{(k)}\boldsymbol{\Sigma}_n\boldsymbol{J}_n^{(k)^\top}, \tag{2.71}$$

$$\boldsymbol{\Sigma}_y^{\Delta(k)} \approx \boldsymbol{J}_x^{(k)}\boldsymbol{\Sigma}_x^{\Delta(k)}\boldsymbol{J}_x^{(k)^\top} + \boldsymbol{J}_n^{(k)}\boldsymbol{\Sigma}_n^{\Delta}\boldsymbol{J}_n^{(k)^\top}, \tag{2.72}$$

$$\boldsymbol{\Sigma}_y^{\Delta^2(k)} \approx \boldsymbol{J}_x^{(k)}\boldsymbol{\Sigma}_x^{\Delta^2(k)}\boldsymbol{J}_x^{(k)^\top} + \boldsymbol{J}_n^{(k)}\boldsymbol{\Sigma}_n^{\Delta^2}\boldsymbol{J}_n^{(k)^\top} \tag{2.73}$$

donde se ha considerado que el ruido convolutivo (distorsión del canal) es constante durante toda la elocución.

Para concluir con la técnica VTS, merece la pena mencionar que las matrices de covarianza obtenidas en la ecuación (2.70) son diagonales por bloques, sin embargo, los reconocedores de voz suelen trabajar con matrices diagonales por razones de eficiencia. Por tanto, en la práctica se suelen diagonalizar las matrices obtenidas. Igualmente, para la estimación del ruido (aditivo y/o de canal) antes hemos mencionado que se pueden utilizar métodos tradicionales como el uso de un VAD para identificar los silencios y, posteriormente, recopilar estadísticas del ruido en estos. Otros trabajos, por contra, han empleado el algoritmo EM para estimar los parámetros de esta distribución conjunta [70, 94, 159, 166, 171]. En la sección 2.3 se profundizará en la estimación de los parámetros del ruido.

Algonquin. La última técnica de adaptación basada en modelos de distorsión que estudiaremos es Algonquin [94, 95, 165]. Algonquin puede considerarse como una extensión de la técnica VTS, diferenciándose de ella en los siguientes aspectos: (i) uso de un modelo de distorsión de la voz más detallado y (ii) los puntos de expansión en torno a los cuales se aproxima el modelo de distorsión son actualizados de forma iterativa en función de la observación ruidosa. Estas dos diferencias, que serán analizadas con

mayor profundidad a continuación, suponen una mayor precisión en el proceso de adaptación a costa de incrementar el coste computacional. Aunque inicialmente Algonquin se propuso como una técnica de realce de las características de la voz ruidosa, aquí la estudiaremos en su vertiente de técnica de adaptación de modelos.

Como se ha dicho, la primera diferencia básica respecto a la técnica VTS es el uso de un modelo de distorsión de la voz más detallado. Para simplificar la comparativa entre ambas técnicas, vamos a suponer que el ruido convolutivo es nulo ($\boldsymbol{h} = \mathbf{0}$), aunque la extensión para considerar este ruido es trivial. Bajo esta premisa, en el apartado anterior hemos visto que el modelo de distorsión en el que se basa VTS es el de la ecuación (2.14). Este modelo, no obstante, no es completamente exacto. Así, la ecuación (2.14) no contempla las posibles interacciones entre las fases del ruido aditivo y la voz limpia que sí aparecían recogidas en el tercer término de la ecuación (2.3). El algoritmo Algonquin, emplea un modelo de distorsión más sofisticado en el que sí aparece recogido explícitamente dicho término en fase. En el dominio del cepstrum, el modelo de distorsión empleado por Algonquin es

$$\boldsymbol{y} \approx \ddot{\boldsymbol{f}}(\boldsymbol{x}, \boldsymbol{n}) = \boldsymbol{x} + \boldsymbol{C} \log\left(\mathbf{1} + e^{\boldsymbol{C}^{-1}(\boldsymbol{n}-\boldsymbol{x})}\right) + \boldsymbol{r}. \tag{2.74}$$

El vector $\boldsymbol{r}$ de la ecuación anterior define la relación entre las fases de los vectores de voz y ruido. Empíricamente se puede demostrar (ver [70, 71]) que los elementos de $\boldsymbol{r}$ se pueden modelar mediante una distribución gaussiana de media 0 ($\mathbb{E}[\boldsymbol{r}] = \mathbf{0}$) y cuyo soporte está definido en el intervalo $[-1, 1]$. Por esta razón, en lugar de modelar la distribución condicional $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n})$ mediante una delta de Dirac tal y como hace VTS, es decir,

$$p_{\text{vts}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n}) = \delta_{\boldsymbol{f}(\boldsymbol{x},\boldsymbol{n})}(\boldsymbol{y}), \tag{2.75}$$

siendo $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{n})$ el modelo de la (2.14), Algonquin considera $\boldsymbol{r}$ como un vector de error residual y, por consiguiente, modela $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n})$ de manera más efectiva mediante una gaussiana con los siguientes parámetros,

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n}) = \int \delta_{\ddot{\boldsymbol{f}}(\boldsymbol{x},\boldsymbol{n})}(\boldsymbol{y}) p(\boldsymbol{r}) d\boldsymbol{r} \simeq \mathcal{N}(\boldsymbol{y}; \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{n}), \boldsymbol{\Psi}) \tag{2.76}$$

De aquí en adelante nos referiremos a la ecuación (2.76) como el modelo Algonquin de distorsión de la voz. Como se puede apreciar, la dependencia de este modelo respecto al factor de fase $\boldsymbol{r}$ desaparece, siendo la matriz de covarianza $\boldsymbol{\Psi}$ la que modela el error cuadrático medio esperado del modelo. En las primeras versiones del algoritmo Algonquin dicha matriz se consideraba fija. No obstante, en posteriores mejoras también se han considerados matrices de covarianza dependientes del nivel de SNR [164].

Una segunda diferencia del algoritmo Algonquin respecto a la técnica VTS es el modo en el que el modelo de distorsión de la ecuación (2.14) se aproxima. Al igual

que VTS, el algoritmo Algonquin también utiliza un desarrollo en serie de Taylor para aproximar linealmente el modelo de distorsión. No obstante, Algonquin aproxima de forma iterativa el modelo, escogiendo en cada iteración la moda de la distribución a posteriori $p(\boldsymbol{x}, \boldsymbol{n}|\boldsymbol{y})$ como punto de expansión. Notemos como $(\boldsymbol{x}_0^{(i)}, \boldsymbol{n}_0^{(i)})$ el punto de expansión empleado en la iteración i. En dicho caso la aproximación del modelo de distorsión viene dada por

$$\boldsymbol{f}^{(i)}(\boldsymbol{x}, \boldsymbol{n}) = \boldsymbol{f}(\boldsymbol{x}_0^{(i)}, \boldsymbol{n}_0^{(i)}) + \boldsymbol{J}_x^{(i)} \left(\boldsymbol{x} - \boldsymbol{x}_0^{(i)}\right) + \boldsymbol{J}_n^{(i)} \left(\boldsymbol{n} - \boldsymbol{n}_0^{(i)}\right), \tag{2.77}$$

donde $\boldsymbol{J}_x^{(i)}$ y $\boldsymbol{J}_n^{(i)}$ son las matrices jacobianas derivadas a partir de la ecuación (2.14) y evaluadas en el punto de expansión de la iteración i-ésima, esto es,

$$\boldsymbol{J}_x^{(i)} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\Big|_{\boldsymbol{x}_0^{(i)}, \boldsymbol{n}_0^{(i)}}, \quad \boldsymbol{J}_n^{(i)} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{n}}\Big|_{\boldsymbol{x}_0^{(i)}, \boldsymbol{n}_0^{(i)}}. \tag{2.78}$$

Sean $\mathcal{N}(\boldsymbol{\mu}_x^{(k)}, \boldsymbol{\Sigma}_x^{(k)})$ y $\mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ la componente k-ésima del modelo acústico que se quiere adaptar y la estimación del ruido aditivo obtenida para el instante de tiempo t, respectivamente. Con objeto de simplificar la notación, de aquí en adelante omitiremos la dependencia respecto al tiempo en la estimación del ruido respecto, así como el índice de la componente k en los parámetros del modelo acústico, dándose por supuesto que las fórmulas de adaptación presentadas deben de aplicarse para todos los instantes de tiempo t y componentes k en el modelo acústico. Teniendo esto en cuenta y bajo la aproximación lineal de la ecuación (2.77), es fácil de ver que la distribución conjunta de $\boldsymbol{z} = (\boldsymbol{x}^\top, \boldsymbol{n}^\top, \boldsymbol{y}^\top)^\top$ es gaussiana con los siguientes parámetros:

$$p^{(i)}(\boldsymbol{z}) = \mathcal{N}\left(\begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{n} \\ \boldsymbol{y} \end{bmatrix}; \begin{bmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_n \\ \boldsymbol{\mu}_y^{(i)} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_x & \mathbf{0} & \boldsymbol{\Sigma}_{xy}^{(i)} \\ \mathbf{0} & \boldsymbol{\Sigma}_n & \boldsymbol{\Sigma}_{ny}^{(i)} \\ \boldsymbol{\Sigma}_{yx}^{(i)} & \boldsymbol{\Sigma}_{yn}^{(i)} & \boldsymbol{\Sigma}_y^{(i)} \end{bmatrix}\right). \tag{2.79}$$

Esta distribución conjunta es uno de los puntos claves de Algonquin. A partir de ella, y usando las herramientas matemáticas desarrolladas para la distribución normal, podemos derivar distintas distribuciones marginales y/o condicionadas. Por ejemplo, las distribuciones marginales de $\boldsymbol{x}$ y $\boldsymbol{n}$ son las mismas que las distribuciones a priori de estas variables y no dependen de la iteración actual. Por otro lado, los parámetros de $p^{(i)}(\boldsymbol{y})$ vendrán en función de la iteración i y del error inicial asumido por el modelo Algonquin ($\boldsymbol{\Psi}$). Es posible demostrar que los parámetros de esta distribución son (ver

p.ej. [260]):

$$\boldsymbol{\mu}_y^{(i)} = \mathbb{E}\left[\boldsymbol{f}^{(i)}(\boldsymbol{x},\boldsymbol{n})\right] = \boldsymbol{f}(\boldsymbol{x}_0^{(i)},\boldsymbol{n}_0^{(i)}) + \boldsymbol{J}_x^{(i)}\left(\boldsymbol{\mu}_x - \boldsymbol{x}_0^{(i)}\right) + \boldsymbol{J}_n^{(i)}\left(\boldsymbol{\mu}_n - \boldsymbol{n}_0^{(i)}\right) \tag{2.80}$$

$$\begin{aligned}\boldsymbol{\Sigma}_y^{(i)} &= \mathbb{E}\left[\left(\boldsymbol{f}^{(i)}(\boldsymbol{x},\boldsymbol{n}) - \boldsymbol{\mu}_y\right)\left(\boldsymbol{f}^{(i)}(\boldsymbol{x},\boldsymbol{n}) - \boldsymbol{\mu}_y\right)^\top\right] + \boldsymbol{\Psi}, \\ &= \mathbb{E}\left[\boldsymbol{J}_x^{(i)}\left(\boldsymbol{x}-\boldsymbol{\mu}_x\right)\left(\boldsymbol{J}_x^{(i)}\left(\boldsymbol{x}-\boldsymbol{\mu}_x\right)\right)^\top + \boldsymbol{J}_n^{(i)}\left(\boldsymbol{n}-\boldsymbol{\mu}_n\right)\left(\boldsymbol{J}_n^{(i)}\left(\boldsymbol{n}-\boldsymbol{\mu}_n\right)\right)^\top\right] + \boldsymbol{\Psi} \\ &= \boldsymbol{J}_x^{(i)}\mathbb{E}\left[\left(\boldsymbol{x}-\boldsymbol{\mu}_x\right)\left(\boldsymbol{x}-\boldsymbol{\mu}_x\right)^\top\right]\boldsymbol{J}_x^{(i)^\top} + \boldsymbol{J}_n^{(i)}\mathbb{E}\left[\left(\boldsymbol{n}-\boldsymbol{\mu}_n\right)\left(\boldsymbol{n}-\boldsymbol{\mu}_n\right)^\top\right]\boldsymbol{J}_n^{(i)^\top} + \boldsymbol{\Psi} \\ &= \boldsymbol{J}_x^{(i)}\boldsymbol{\Sigma}_x\boldsymbol{J}_x^{(i)^\top} + \boldsymbol{J}_n^{(i)}\boldsymbol{\Sigma}_n\boldsymbol{J}_n^{(i)^\top} + \boldsymbol{\Psi}. \end{aligned} \tag{2.81}$$

Debemos notar la similitud de las expresiones de adaptación anteriores con aquellas obtenidas para VTS en las ecuaciones (2.60) y (2.61). La diferencia más significativa entre estas expresiones es la dependencia $\boldsymbol{\mu}_y^{(i)}$ respecto a las matrices jacobianas en Algonquin. Estas matrices no aparecen en la fórmula de adaptación de VTS, ya que se anulan cuando el punto de expansión escogido son las propias medias de la voz limpia y el ruido.

Las matrices de covarianzas cruzadas que aparecen en (2.79) y que, como veremos más adelante, intervienen en el cálculo de los puntos de expansión en torno a los cuales se aproxima el modelo de distorsión, se pueden calcular de la siguiente manera,

$$\begin{aligned}\boldsymbol{\Sigma}_{yx}^{(i)} &= \mathbb{E}\left[(\boldsymbol{f}^{(i)}(\boldsymbol{x},\boldsymbol{n}) - \boldsymbol{\mu}_y)(\boldsymbol{x}-\boldsymbol{\mu}_x)^\top\right] \\ &= \mathbb{E}\left[(\boldsymbol{J}_x^{(i)}(\boldsymbol{x}-\boldsymbol{\mu}_x))(\boldsymbol{x}-\boldsymbol{\mu}_x)^\top\right] \\ &= \boldsymbol{J}_x^{(i)}\mathbb{E}\left[\left(\boldsymbol{x}-\boldsymbol{\mu}_x\right)\left(\boldsymbol{x}-\boldsymbol{\mu}_x\right)^\top\right] \\ &= \boldsymbol{J}_x^{(i)}\boldsymbol{\Sigma}_x, \end{aligned} \tag{2.82}$$

$$\boldsymbol{\Sigma}_{yn}^{(i)} = \boldsymbol{J}_n^{(i)}\boldsymbol{\Sigma}_n. \tag{2.83}$$

Usando las covarianzas cruzadas anteriores y los parámetros de la PDF de voz ruidosa, es posible derivar la distribución a posteriori $p^{(i)}(\boldsymbol{x},\boldsymbol{n}|\boldsymbol{y})$ a partir de la distribución conjunta dada en la ecuación (2.79). Como se ha comentado anteriormente, Algonquin emplea la media de la distribución a posteriori como punto de expansión para la iteración $i+1$. Los parámetros de $p^{(i)}(\boldsymbol{x},\boldsymbol{n}|\boldsymbol{y})$ se pueden obtener utilizando el conocido complemento de Schur [211],

$$p^{(i)}(\boldsymbol{x},\boldsymbol{n}|\boldsymbol{y}) = \mathcal{N}\left((\boldsymbol{x}^\top,\boldsymbol{n}^\top)^\top;\boldsymbol{\mu}_{x,n|y}^{(i)},\boldsymbol{\Sigma}_{x,n|y}^{(i)}\right), \tag{2.84}$$

donde

$$\boldsymbol{\mu}_{xn|y}^{(i)} = \begin{bmatrix}\boldsymbol{\mu}_x \\ \boldsymbol{\mu}_n\end{bmatrix} + \begin{bmatrix}\boldsymbol{\Sigma}_{xy}^{(i)} \\ \boldsymbol{\Sigma}_{ny}^{(i)}\end{bmatrix}\boldsymbol{\Sigma}_y^{(i)^{-1}}\left(\boldsymbol{y} - \boldsymbol{\mu}_y^{(i)}\right), \tag{2.85}$$

$$\boldsymbol{\Sigma}_{xn|y}^{(i)} = \begin{bmatrix}\boldsymbol{\Sigma}_x & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_n\end{bmatrix} - \begin{bmatrix}\boldsymbol{\Sigma}_{xy}^{(i)} \\ \boldsymbol{\Sigma}_{ny}^{(i)}\end{bmatrix}\boldsymbol{\Sigma}_y^{(i)^{-1}}\left[\boldsymbol{\Sigma}_{yx}^{(i)}\ \boldsymbol{\Sigma}_{yn}^{(i)}\right]. \tag{2.86}$$

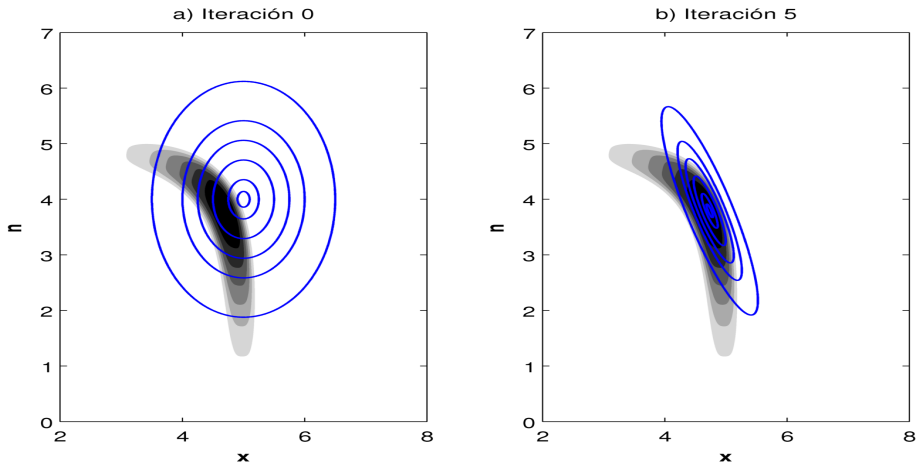

Figura 2.10: Aproximación de la distribución a posteriori $p(x, n|y)$ obtenida por el algoritmo Algonquin con $x \sim \mathcal{N}(\mu_x = 5, \sigma_x^2 = 1)$, $n \sim \mathcal{N}(\mu_n = 4, \sigma_n^2 = 2)$, $\psi = 0,04$ (varianza de $p(y|x, n)$) y $y = 5$. a) Distribución a posteriori exacta y aproximación obtenida al inicializar el algoritmo con la distribución a priori $p(x, n)$. b) Distribución exacta y aproximación gaussiana tras 5 iteraciones del algoritmo.

Luego en la iteración $i+1$ los valores de los puntos de expansión $\boldsymbol{x}_0^{(i+1)}$ y $\boldsymbol{n}_0^{(i+1)}$ son

$$\boldsymbol{x}_0^{(i+1)} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}^{(i)} \boldsymbol{\Sigma}_y^{(i)^{-1}} \left(\boldsymbol{y} - \boldsymbol{\mu}_y^{(i)} \right), \tag{2.87}$$

$$\boldsymbol{n}_0^{(i+1)} = \boldsymbol{\mu}_n + \boldsymbol{\Sigma}_{ny}^{(i)} \boldsymbol{\Sigma}_y^{(i)^{-1}} \left(\boldsymbol{y} - \boldsymbol{\mu}_y^{(i)} \right). \tag{2.88}$$

Inicialmente, al igual que en la técnica VTS, los puntos de expansión se inicializan al valor de las medias de las distribuciones a priori:

$$\boldsymbol{x}_0^{(0)} = \boldsymbol{\mu}_x, \tag{2.89}$$

$$\boldsymbol{n}_0^{(0)} = \boldsymbol{\mu}_n. \tag{2.90}$$

Finalmente, en la figura 2.10 podemos observar un ejemplo de la aproximación gaussiana que el algoritmo Algonquin realiza de la distribución a posteriori $p(x, n|y)$. Como se puede apreciar en la gráfica de la izquierda, Algonquin inicializa el método iterativo con la distribución a priori de la voz y el ruido, esto es, $p^{(0)}(x, n|y) = p(x, n)$. Tras varias iteraciones (cinco en el ejemplo mostrado) la aproximación de la distribución a posteriori converge, coincidiendo además la media de la distribución aproximada con la moda de la distribución teórica. Hay que resaltar también que mientras en la distribución a priori $p(x, n)$ la voz y el ruido son independientes, en la distribución a posteriori $p(x, n|y)$ la correlación entre dichas variables es no nula.

2.2.2.3. Entrenamiento adaptativo

En las técnicas de adaptación de modelos descritas hasta ahora se ha supuesto que los modelos acústicos de los que se parte (los modelos a adaptar) están entrenados con voz limpia. No obstante, en la práctica esto no suele ser así, encontrándonos que los modelos suelen entrenarse también con voz distorsionada (p.ej. modelos multicondición) para mejorar la robustez al ruido de los sistemas de RAH. Esto, por tanto, puede suponer un problema a la hora de extrapolar las ideas en las que se basan las técnicas de adaptación estudiadas anteriormente.

A lo anterior hay que añadir la dificultad práctica para adquirir grandes volúmenes de datos con voz de alta calidad grabada en un estudio. Debido a lo anterior, en los últimos años ha habido un gran interés por parte de la comunidad científica en usar fuentes de voz disponibles, por ejemplo, en Internet para entrenar los modelos acústicos. El problema de utilizar estas fuentes para el entrenamiento de los modelos del reconocedor es su mayor heterogeneidad en comparación con la voz grabada en un estudio, lo que se traduce en una degradación del rendimiento del sistema del reconocimiento debido al mayor número de locutores, acentos, condiciones acústicas y de canal, etc. El enfoque clásico empleado para compensar esta heterogeneidad consiste en implementar una extracción robusta de las características de la voz que elimine, en la medida de lo posible, las variaciones no deseadas. No obstante, no parece razonable que durante el proceso de extracción de características sea posible suprimir todas estas variaciones. De ahí que este procesado robusto se haya combinado con un entrenamiento del tipo multicondición o multiestilo con el fin de modelar tanto las variaciones deseadas de la voz como aquellas no deseadas.

En este apartado haremos una breve introducción a un enfoque alternativo, y más potente, para el entrenamiento de sistemas de reconocimiento sobre datos heterogéneos que permite resolver las dos problemas anteriores: el entrenamiento adaptativo. El entrenamiento adaptativo (del inglés *adaptive training*) [16, 102, 228, 281] fue concebido inicialmente como una estrategia para abordar el problema de las variaciones interlocutor de forma más efectiva y, en algunos sentidos, puede considerarse como una extensión al entrenamiento multicondición. En lugar de entrenar un modelo acústico independiente del locutor con la voz de todos los locutores, el conjunto de datos de entrenamiento $(\boldsymbol{X}, \mathcal{H})$ se divide en en un conjunto bloques homogéneos $\boldsymbol{X} = (\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(B)})$ y $\mathcal{H} = (\mathcal{H}^{(1)}, \cdots, \mathcal{H}^{(B)})$ ($\mathcal{H}^{(i)}$ son las transcripciones asociadas al bloque de datos $\boldsymbol{X}^{(i)}$), donde cada bloque representa una condición acústica determinada (p.ej. un locutor o un entorno acústico específico). Usando estos bloques, el entrenamiento adaptativo estima los siguientes modelos:

1. **Modelo canónico** $\mathcal{M}$: el modelo canónico captura la variabilidad deseada de la voz (en el sentido de ser útil para el reconocimiento de la voz) independientemente de la condición acústica en la que se haya grabado. Este modelo, expresado generalmente en forma de HMM, representa las datos de entrenamiento dadas las transformaciones oportunas.

2. **Conjunto de transformaciones** $\mathcal{T} = (\mathcal{T}^{(1)}, \cdots, \mathcal{T}^{(B)})$: modelan las variaciones acústicas no deseadas (p.ej. distorsión producida por un entorno acústico o particularidades de un locutor determinado). Estas transformaciones pueden, por tanto, utilizarse para adaptar el modelo canónico $\mathcal{M}$ a una condición acústica en particular o para normalizar un bloque de datos determinado.

De esta forma, puede considerarse que el entrenamiento adaptativo es una versión factorizada de los modelos tradicionales, donde por un lado se modelan las características relevantes de la voz (modelo canónico) y por otro las variaciones no deseadas (conjunto de transformaciones). Al modelar las variaciones no lingüísticas de la voz de forma mediante $\mathcal{T}$, el modelo acústico $\mathcal{M}$ resultante será más compacto y, por tanto, su adaptación a nuevas condiciones acústicas será más precisa que la adaptación de un modelo acústico multicondición, por ejemplo. Por otro lado, también cabe decir que el modelo canónico siempre deberá utilizarse en combinación con una transformación dada, es decir, el modelo deberá ser adaptado antes de emplearse para descodificar la voz (incluso voz limpia). Para ello será necesario estimar, a partir de los datos de *test*, la transformación que modele los factores irrelevantes de la voz, como pueden ser las diferencias entre la voz de distintos locutores, dando lugar a lo que se conoce como SAT (*Speaker Adaptive Training*, entrenamiento adaptativo por locutor) [16, 100], o la degradación de las características de voz por el ruido, dando lugar a las técnicas NAT (*Noise Adaptive Training*, entrenamiento adaptativo por ruido) [68, 154, 155, 176], o ambas fuentes de variabilidad de forma factorizada [101, 272]. Como se puede apreciar, en todos los casos se pretende eliminar del proceso de entrenamiento todos los factores irrelevantes para el reconocimiento automático de la voz, de ahí que el entrenamiento adaptativo también se conozca con el nombre de IVN (*Irrelevant Variability Normalization*, normalización de la variabilidad irrelevante) [146, 243, 282].

Dentro del entrenamiento adaptativo podemos encontrar dos grandes grupos de técnicas [281]: aquellas eliminan la variabilidad no deseada en las características de la voz y las que se aplican a los modelos acústicos. En el primer grupo encontramos las técnicas de normalización y compensación de características. El objetivo es, por tanto, extraer un conjunto de características más robusto a las variabilidades no deseadas. El segundo grupo de técnicas lo conforman aquellas que estiman de forma conjunta,

generalmente usando el criterio de máxima verosimilitud, el modelo canónico y el conjunto de transformaciones. Por ejemplo, la técnica SAT suele emplearse en conjunción con la técnica MLLR para estimar las transformaciones que modelan las variabilidades irrelevantes debidas al locutor [16, 100]. Para el caso del ruido, NAT suele emplearse en combinación con la técnica VTS. En este último caso las transformaciones modelan la distorsión introducida por el ruido, estimándose los parámetros oportunos de éstas durante el proceso de entrenamiento (p.ej. en VTS se estiman los parámetros de los modelos de ruido aditivo y convolutivo [155]).

Dados el modelo canónico $\mathcal{M}$ y el conjunto de transformaciones $\mathcal{T}$, la probabilidad de los datos de entrenamiento $\boldsymbol{X} = (\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(B)})$ se calcula suponiendo independencia entre los distintos bloques de datos homogéneos, esto es,

$$p\left(\boldsymbol{X}|\mathcal{H}, \mathcal{M}, \mathcal{T}\right) = \prod_{b=1}^{B} p(\boldsymbol{X}^{(b)}|\mathcal{H}^{(b)}, \mathcal{M}, \mathcal{T}^{(b)}), \tag{2.91}$$

donde el índice b indica las distintas condiciones acústicas (p.ej. locutores o tipos de ruido) presentes en los datos de entrenamiento.

En el entrenamiento adaptativo el modelo canónico empleado es aquél que maximiza la probabilidad dada en la ecuación (2.91),

$$\hat{\mathcal{M}} = \underset{\mathcal{M}}{\operatorname{argmax}}\, p(\boldsymbol{X}|\mathcal{H}, \mathcal{M}, \hat{\mathcal{T}}), \tag{2.92}$$

siendo $\hat{\mathcal{T}}$ la estimación actual del conjunto de transformaciones.

Dado un bloque de datos homogéneo s, la transformación para dicho bloque se estima a partir de la estimación actual del modelo canónico,

$$\hat{\mathcal{T}}^{(b)} = \underset{\mathcal{T}}{\operatorname{argmax}}\, p(\boldsymbol{X}^{(b)}|\mathcal{H}^{(b)}, \hat{\mathcal{M}}, \mathcal{T}). \tag{2.93}$$

La estimación de (2.92) y (2.93) se lleva a cabo de forma iterativa mediante el algoritmo EM: inicialmente se parte de un modelo acústico independiente del locutor o multicondición y de un conjunto de transformaciones lineales identidad. En base a estos valores iniciales, el algoritmo EM va iterando hasta que se alcance una cierta condición de finalización. En cada iteración se estiman, en primer lugar, los parámetros del conjunto de transformaciones usando para ello los valores obtenidos en la iteración anterior y, en segundo lugar, el modelo canónico dado el conjunto de transformaciones obtenido en esa iteración [281].

Debe quedar claro que las transformaciones estimadas durante el entrenamiento adaptativo se descartan, por no ser de utilidad en la fase de evaluación. Así, es necesario estimar, generalmente para cada frase de *test*, la transformación oportuna que modele

las variabilidades irrelevantes de la voz. Por ejemplo, bajo el esquema NAT-VTS para cada frase de *test* se estiman los parámetros de las distribuciones de ruido usando el algoritmo EM [154, 155].

2.2.3. Modificación de las características de voz

Al contrario que las técnicas de adaptación estudiadas en la sección anterior, las técnicas que estudiaremos en este apartado trabajan en el dominio de las características de voz. A pesar de esta diferencia, la meta de ambos enfoques es la misma: reducir la discrepancia entre las condiciones de entrenamiento y las de evaluación. En el caso de las técnicas estudiadas en este apartado, la consecución de esta meta se alcanza modificando convenientemente las características con las que se reconocen. En las siguientes secciones veremos que existen principalmente dos estrategias para llevar a cabo esta tarea: reducción de ruido y normalización de características.

Las técnicas de reducción de ruido intentan mitigar la distorsión producida por el ruido en la señal de voz antes de emplearla para reconocer. Así, en lugar de emplear voz ruidosa, el reconocedor trabaja ahora sobre la señal pseudo-limpia resultante del proceso de compensación. Dentro de las técnicas de compensación consideraremos, a su vez, las siguientes categorías: realce de voz y de compensación de características.

Los algoritmos de realce de voz que estudiaremos en la sección 2.2.3.2 son métodos genéricos que mejoran la calidad de la voz en presencia de ruido. Como norma general, nos encontramos que estos algoritmos no han sido diseñados específicamente con el objetivo de mejorar la robustez al ruido del RAH, sino que inicialmente fueron concebidos para otros propósitos (p.ej. realce de voz en telefonía). Debido a esto, el objetivo que se suele perseguir es la mejora de la calidad perceptiva de la señal realzada. Generalmente el proceso de realce se lleva a cabo en dominios como el temporal o el de la DFT, si bien en los últimos años ha habido un creciente interés por el uso de otros dominios tradicionalmente usados en RAH (p.ej. el domino cepstral [207]).

Al contrario que los algoritmos de realce, las técnicas de compensación de características sí que persiguen como meta específica el incremento de la robustez de los sistemas de RAH, es decir, minimizar el error de reconocimiento de estos sistemas. Para lograr este cometido, las distintas técnicas procesan la señal de voz ruidosa expresada en distintos dominios (p.ej. log-Mel o MFCC) para compensar la distorsión producida por el ruido, de forma que a la salida se obtenga una señal de voz más parecida a la original limpia. Un aspecto importante de estas técnicas es la estimación de las transformaciones que aplican a las características ruidosas para compensarlas. Al igual que en el caso de las técnicas de adaptación, encontramos técnicas que derivan estas transformaciones a partir de modelos analíticos de distorsión (sección 2.2.3.3) o

usando grabaciones estéreo con voz limpia y distorsionada (sección 2.2.3.4).

Como último grupo de técnicas que modifican las características de voz encontramos a las de normalización. En este caso no se persigue obtener unas características más limpias, sino transformarlas a un dominio en el que se encuentren menos afectadas por el ruido. Aparte de esto, otra diferencia más con respecto a las técnicas de cancelación de ruido es que la transformación se aplica tanto a las elocuciones empleadas en el entrenamiento como a las que se reconocen.

En los siguientes subapartados se realiza una revisión de los tres tipos de técnicas mencionados, centrándonos principalmente en la compensación de características, tema central de esta tesis.

2.2.3.1. Normalización de características

Las técnicas de normalización tienen como objeto transformar las características extraídas de la voz a un dominio en el que la variabilidad introducida por el ruido se vea minimizada, consiguiendo de esta forma reducir la discrepancia entre las distribuciones de los datos empleados para entrenamiento y los usados para reconocer. Desde este punto de vista, algunos autores consideran a las técnicas de normalización como parte integrante de los extractores robustos de la voz estudiados en la sección 2.2.1. A diferencia de las técnicas de compensación, las técnicas de normalización suelen realizar pocas o ninguna suposición sobre las características del ruido que distorsiona la voz. Esto les permite abordar de forma satisfactoria distintos tipos de distorsiones que puedan degradar la señal de voz, aunque, por otro lado, al no modelar de forma explícita las características de estas distorsiones (p.ej. la densidad de potencia espectral del ruido), limita la precisión con la éstas pueden mitigarse. Otra ventaja de estas técnicas es su simplicidad, siendo asaz eficientes tanto en tiempo como en memoria. Como consecuencia de todas estas características, encontramos que estas técnicas suelen estar presentes en la gran mayoría de los sistemas de reconocimiento actuales.

Dentro de las técnicas de normalización propuestas en la literatura, en este apartado estudiaremos aquellas que normalizan los momentos de la distribución de los datos observados. En particular, nuestro estudio se centrará en las técnicas de normalización de la media cepstral (CMN, *Cepstral Mean Normalization*) [20], normalización de la media y la varianza cepstral (MVN, *Mean and Variance Normalization*) y la técnica de ecualización de histogramas (HEQ, *Histogram EQualization*) [108, 258, 259], las cuales normalizan, respectivamente, el primer, los dos primeros y todos los momentos de la distribución de los datos observados. En líneas generales, cuanto mayor sea el número de momentos normalizados, mayor será el volumen de datos necesario para estimar de forma robusta los parámetros de las transformaciones implicadas en la normalización

de los datos [108, 238].

CMN. Consideremos el modelo de distorsión de la voz expresado en las ecuaciones (2.9) y (2.12) para los dominios log-Mel y cepstral, respectivamente. Existen ciertas situaciones en las que podemos considerar que la principal fuente de distorsión de la voz se debe al ruido de canal $\boldsymbol{h}$, siendo nula, por tanto, la componente del ruido aditivo $\boldsymbol{n}$. Estas situaciones se pueden producir, por ejemplo, cuando la interacción con el sistema de reconocimiento se desarrolla sobre ambientes silenciosos, pero existen diferencias entre las funciones de transferencia de los micrófonos usados para grabar las frases de entrenamiento y las de *test*. Otras fuentes de discrepancia que también involucran un filtrado lineal del espectro son la acústica de la sala, la distancia del locutor al micrófono o la respuesta en frecuencia del canal empleado para transmitir la voz. En todas estas situaciones, los modelos de distorsión mencionados anteriormente nos indican que el efecto principal del filtrado lineal $\boldsymbol{h}$ es un desplazamiento aditivo constante de las características en ambos dominios. Una de las técnicas más simples, pero a la vez más efectivas, para abordar este tipo de distorsiones es la técnica de normalización de la media cepstral [20] la cual describimos a continuación.

Consideremos la secuencia de vectores de características expresados en el dominio cepstral $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$. La secuencia normalizada $\{\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_T\}$ devuelta por CMN se obtiene tras sustraer la media muestral $\bar{\boldsymbol{x}}$ a cada vector $\boldsymbol{x}_t$,

$$\bar{\boldsymbol{x}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t, \tag{2.94}$$

$$\hat{\boldsymbol{x}}_t = \boldsymbol{x}_t - \bar{\boldsymbol{x}}. \tag{2.95}$$

Como consecuencia de la operación anterior, forzamos que el primer momento de la distribución de los datos observados, esto es, la media cepstral, sea nula. Como veremos a continuación, esta operación equivale a suprimir el efecto del ruido de canal sobre los datos observados. Para lograr este efecto, la distorsión del canal $\boldsymbol{h}$ debe ser estacionaria y su duración menor que la de la ventana de análisis empleada durante la extracción de características.

Consideremos la señal $y(n)$ obtenida tras aplicar el filtro $h(n)$ a la señal de voz $x(n)$. Bajo los supuestos anteriormente mencionados, en el dominio cepstral esta operación de filtrado se traduce en

$$\boldsymbol{y}_t = \boldsymbol{x}_t + \boldsymbol{h}. \tag{2.96}$$

En este caso, la media cepstral de la señal filtrada viene dada por

$$\bar{\boldsymbol{y}} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{y}_t = \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{x}_t + \boldsymbol{h}) = \bar{\boldsymbol{x}} + \boldsymbol{h}. \tag{2.97}$$

Tras normalizar la secuencia filtrada podemos ver cómo el efecto del canal $\boldsymbol{h}$ desaparece,

$$\hat{\boldsymbol{y}}_t = \boldsymbol{y}_t - \bar{\boldsymbol{y}} = (\boldsymbol{x}_t + \boldsymbol{h}) - (\bar{\boldsymbol{x}} + \boldsymbol{h}) = \boldsymbol{x}_t - \bar{\boldsymbol{x}} = \hat{\boldsymbol{x}}_t. \tag{2.98}$$

Para que la técnica CMN surta efecto se debe aplicar tanto en la fase de entrenamiento como en la de explotación del sistema. Asimismo, se ha verificado experimentalmente que ésta técnica funciona mejor para frases largas ($T \to \infty$) en las que la media muestral $\bar{\boldsymbol{x}}$ coincide aproximadamente con el desplazamiento debido al canal [215]. Para frases cortas, esta media puede contener información acústica relevante para el reconocimiento, resultando en una degradación del rendimiento del sistema. En la práctica se ha comprobado que CMN es muy útil para reducir la discrepancia debida al uso de diferentes micrófonos y canales telefónicos, pero no para combatir el efecto de la reverberación cuando el tiempo característico de la sala se aproxima al de la ventana de análisis [34].

MVN. Si CMN está pensada para normalizar el primer momento de la distribución de los datos observados, una extensión lógica a la misma consiste en normalizar los dos primeros momentos, esto es, la media y la varianza cepstral. Ésta es justamente la normalización que MVN implementa [268],

$$\hat{\boldsymbol{x}}_t = \frac{\boldsymbol{x}_t - \bar{\boldsymbol{x}}}{\boldsymbol{\sigma}_x}, \tag{2.99}$$

donde la operación de división se realiza elemento a elemento y la varianza de los datos se calcula de la siguiente forma;

$$\boldsymbol{\sigma}_x^2 = \frac{1}{T-1}\sum_{t=1}^{T}(\boldsymbol{x}_t - \bar{\boldsymbol{x}})^2. \tag{2.100}$$

Con esta transformación conseguimos que la media de la secuencia de vectores de características de la voz sea cero y su varianza unitaria. Al contrario que CMN, MVN no está diseñada específicamente para combatir una distorsión determinada, sino que experimentalmente ha demostrado ser efectiva para reducir las discrepancias debidas al canal, a la variabilidad del locutor y al ruido aditivo [34]. En este último caso debemos de recordar que, tal y como aparece reflejado en la figura 2.4, el efecto del ruido (aditivo) sobre los coeficientes cepstrales se traduce no sólo en un desplazamiento de la media, sino también en una compresión de la varianza de la distribución. De acuerdo con esta observación la normalización de estos dos momentos debe conllevar una reducción del desajuste provocado por el ruido en los datos y, con ello, un aumento de la tasa de reconocimiento del sistema.

HEQ. Siguiendo con la forma de discurrir iniciada con CMN y continuada en la técnica MVN, podemos pensar en normalizar los tres primeros momentos (media, varianza y asimetría) de la distribución de los datos observados [255], o incluso normalizar la curtosis (momento de cuarto orden) y otros momentos de orden superior [143]. Llevado hasta el límite, este razonamiento da lugar a la técnica de ecualización de histogramas, HEQ, cuya meta es normalizar todos los momentos de la distribución de los datos [73, 108, 258, 259].

En lugar de suprimir el ruido de la señal observada, la técnica HEQ, al igual que CMN y MVN, trabaja con la distribución de los datos observados como un conjunto. Como ya comentábamos anteriormente, la ventaja principal de esta forma de proceder radica en que no se hacen suposiciones a priori sobre las características del ruido, ni la forma en la que éste corrompe la voz. Simplemente, HEQ supone una transformación no lineal que modifica la distribución de los datos observados para que ésta concuerde con la distribución de los datos de entrenamiento. Para una variable aleatoria dada y cuya función de distribución de probabilidad es $p_y(y)$, la función $x = f(y)$ que transforma $p_y(y)$ en la PDF $p_x(x)$ viene dada por [259],

$$x = f(y) = C_x^{-1}(C_y(y)), \tag{2.101}$$

donde $C_y(y)$ denota la CDF asociada a $p_y(y)$ y $C_x^{-1}(x)$ representa la función de distribución inversa, esto es, $z = C_x^{-1}(\rho) \Leftrightarrow C_x(z) = \rho, \rho \in [0, 1]$.

A la hora de ecualizar los datos mediante HEQ, varias son las decisiones que el diseñador del sistema debe de tomar. En primer lugar nos encontramos con la decisión relativa al dominio de referencia al que se ecualizan los datos observados, es decir, la elección de C_x. Esta referencia puede ser una distribución estadística paramétrica, p.ej. una gaussiana de media cero y varianza unitaria [259], o una distribución empírica estimada a partir de los datos de entrenamiento [238]. Experimentalmente se ha demostrado que la referencia obtenida a partir del histograma de los datos de entrenamiento proporciona mejores resultados [108]. En segundo lugar, y no menos importante, debe decidirse en qué domino se aplica la técnica HEQ. Posibles dominios son el del banco de filtros [139] o el cepstral [106, 238, 259]. Pocos son los estudios que comparan ambas aproximaciones, aunque dado que la técnica HEQ se aplica por separado a cada componente de los vectores de características, parece razonable que esta técnica produzca mejores resultados en dominios menos correlacionados, como ocurre en el dominio de los MFCCs. De nuevo debemos de remarcar que, al igual que las otras técnicas de normalización estudiadas, HEQ se aplica tanto a los datos de entrenamiento como a los de evaluación.

La técnica HEQ presenta varias deficiencias que pueden limitar notablemente su rendimiento. En primer lugar, está la cuestión de cómo se estima el histograma de los

datos observados, C_y, de forma robusta. Esta cuestión está íntimamente relacionada con una suposición básica sobre la que HEQ se asienta: se asume que la distribución de los datos observados es la misma que la de los datos de entrenamiento, o dicho de otra forma, en cada frase se deben de observar todas las unidades acústicas en la misma proporción que ocurren durante el entrenamiento. Por supuesto esta presunción sólo es cierta para frases muy largas, de varios minutos de duración. Para suavizar este problema en aquellos casos en los que únicamente se disponga de unos pocos segundos de voz, en [139] se propone estimar C_y a partir de los cuantiles calculados para los datos observados en lugar de estimar el histograma de los mismos directamente. En [238] se propone una versión aún más eficaz y robusta de esta misma idea basada en la estadística ordenada de los datos. Dada la secuencia temporal $(y_1^i, y_2^i, \ldots, y_T^i)$ referida al i-ésimo coeficiente cepstral, la estadística ordenada para este coeficiente se obtiene tras ordenar dicha secuencia de forma ascendente

$$y_{(1)}^i \leq y_{(2)}^i \leq \ldots \leq y_{(r)}^i \leq \ldots \leq y_{(T)}^i, \tag{2.102}$$

donde el subíndice (r) indica la posición de cada elemento en el vector ordenado.

La probabilidad acumulada de un valor en particular $y_{(r)}^i$ se obtiene entonces como

$$C_y(y_{(r)}^i) = \frac{r - 0,5}{T} \quad \forall r = 1, \ldots, T. \tag{2.103}$$

Otra deficiencia presente en la técnica HEQ original se constata en la dependencia de la transformación aplicada por ésta respecto a la proporción de tramas de silencio en la frase a normalizar. En efecto, la función de distribución de los datos observados C_y se puede expresar como una combinación lineal de las funciones calculadas para los segmentos de voz, C_{sy}, y los segmentos de no voz C_{ny} [107],

$$C_y(y) = \alpha C_{ny}(y) + (1 - \alpha) C_{sy}(y), \tag{2.104}$$

donde α indica el ratio de tramas de no voz en la frase a reconocer.

De acuerdo con esta expresión, el número de tramas de silencio tendrá un impacto directo en el histograma de los datos y, por tanto, en la transformación aplicada por HEQ. Para evitar este efecto no deseado, se pueden eliminar de los datos a ecualizar aquellas tramas identificadas como silencio por un VAD. Otra opción considerada en [107] es disponer dos clases de referencia a la hora de ecualizar: una para la voz y otra para el silencio. Cada componente cepstral se ecualiza entonces respecto a estas dos clases por separado y, finalmente, se combinan linealmente ambos resultados en función de las probabilidades a posteriori de cada clase. Recientemente, la extensión lógica de esta idea al caso multi-clase (se consideran más de dos clases de referencia) ha sido explorada por varios autores con resultados prometedores [106, 254].

Como última deficiencia de la versión original de HEQ aquí resaltaremos su debilidad para compensar la distorsión introducida por ruidos altamente no estacionarios. Al contrario que en el caso de los ruidos estacionarios donde más observaciones suponen una mejor estimación de su espectro, el espectro de los ruidos no estacionarios o espontáneos puede variar en periodos de tiempo muy pequeños, por lo que aplicar una transformación global a todos los datos podría degradar la voz en esos instantes de tiempo. Para evitar esto, en [215, 238] se propone una versión adaptable de HEQ en la que se ecualizan las tramas dentro de una ventana deslizante. Esta versión permite, además, una implementación en tiempo real de HEQ. Los resultados de reconocimiento obtenidos por esta versión para la base de datos Aurora2 muestran que, con sólo un retardo de 50 ms, esta versión es capaz de obtener los mismos resultados que empleando la totalidad de los datos observados [238].

2.2.3.2. Realce de voz

Los distintos algoritmos de realce de voz persiguen la mejora de la calidad perceptiva de la voz ruidosa de cara a su posterior escucha por un ser humano. Debido a esto, la mayoría de algoritmos suelen trabajar en el dominio del tiempo o el dominio espectral, obteniéndose al final del proceso de realce una señal en el dominio del tiempo. Aunque es de esperar que los sistemas de reconocimiento también se beneficien de este realce, esto no siempre está garantizado. Las razones de ello son múltiples, desde el uso de bloques de procesado orientados a mejorar la calidad perceptiva de la voz que afectan negativamente al reconocimiento, hasta la mayor sensibilidad de estos sistemas a ciertas distorsiones que los humanos toleramos fácilmente.

Uno de las primera propuestas en el campo del realce de voz la encontramos en el algoritmo de sustracción espectral (SS, *Spectral Subtraction*) [37]. Esta técnica proporciona una estimación del espectro de la señal limpia sustrayendo al espectro ruidoso una estimación del ruido acústico:

$$|\hat{X}(f,t)|^{\alpha} = \text{máx}\left(|Y(f,t)|^{\alpha} - \beta|\hat{N}(f,t)|^{\alpha}, \gamma\right), \tag{2.105}$$

donde f indica el índice en frecuencia, t es el índice temporal y α indica el dominio donde se efectúa la sustracción: en el dominio de la magnitud espectral ($\alpha = 1$) o en el dominio de la potencia ($\alpha = 2$). Las estimas de ruido $\hat{N}(f,t)$ empleadas por el algoritmo SS se suelen calcular a partir de las partes de silencio de la señal, siendo común el uso de detectores de actividad de voz para identificar estas partes. Como puede darse el caso de que esta estima sea errónea y conduzca a valores negativos en magnitud o potencia del espectro, al espectro estimado de voz se le aplica entonces un umbral positivo γ que evita estos casos. Esta operación no lineal puede producir el

denominado ruido musical, que básicamente viene determinado por la incapacidad de estimar correctamente el espectro de ruidos poco estacionarios. Por último, también es común el uso de factores de sobresustracción, β, para compensar parcialmente las deficiencias en la estimación del ruido.

Como puede deducirse, la sustracción espectral es altamente dependiente de las estimas de ruido empleadas, que, a su vez, dependen de características como la estacionariedad del ruido, la SNR de la señal observada, etc. Para hacer frente a estas limitaciones, en la literatura se han propuesto otras técnicas de realce como la sustracción no lineal [181] o el filtrado de Wiener [34]. Otras técnicas de realce muy conocidas son las propuestas por Ephraim y Malah [81, 82, 83]. A diferencia de las anteriores, en estas técnicas el espectro de la voz limpia se estima siguiendo un criterio MMSE donde se asume que las componentes en frecuencia del ruido y la voz son variables independientes y normalmente distribuidas.

2.2.3.3. Compensación basada en modelos de distorsión

Al igual que las técnicas de realce de voz, las técnicas de compensación que presentaremos en este apartado permiten procesar las características de voz para mitigar la distorsión producida por el ruido. La mayor diferencia entre ambos enfoques reside en que las técnicas de compensación, al contrario que las de realce, fueron concebidas expresamente para robustecer los sistemas de reconocimiento frente al ruido. De esta forma, la mayoría suelen operar en el dominio cepstral o en el de los parámetros log-Mel. Asimismo, estas técnicas emplean, como su propio nombre indican, modelos analíticos para derivar las transformaciones que se aplican a las características ruidosas. El uso de modelos analíticos permite que estas técnicas sean muy eficientes en tiempo y memoria, requiriendo además pocos datos empíricos para estimar las transformaciones aplicadas. Como contrapartida, la mayor debilidad de estas técnicas reside en la correcta estimación de los parámetros involucrados en el cómputo de las transformaciones, como por ejemplo la estimación de la densidad espectral de potencia del ruido.

A fin de unificar la descripción de las técnicas de compensación consideraremos un estimador bayesiano MMSE que, partiendo de la observación ruidosa $\boldsymbol{y}$, estime las características relativas a la voz limpia $\hat{\boldsymbol{x}}$. El hecho de restringirnos al estimador MMSE no restará generalidad a la exposición, ya que, de hecho, la mayoría de técnicas que presentaremos se derivan de este estimador. Aparte de esto, durante la presentación asumiremos que la distorsión de canal es nula ($\boldsymbol{h} = \mathbf{0}$), simplificando con esto la exposición de dichas técnicas. Bajo las suposiciones anteriores, en la sección 3.1 se

mostrará que la estimación MMSE para las características de voz viene dada por

$$\hat{\boldsymbol{x}} = \sum_{k=1}^{M} P(k|\boldsymbol{y})\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, k], \tag{2.106}$$

donde suponemos que el estimador MMSE parte de un GMM de voz limpia $\mathcal{M}_x$ con M componentes.

Como puede verse, dos son los términos que intervienen en el cómputo de la estimación MMSE: las probabilidades a posteriori $P(k|\boldsymbol{y})$ y los valores esperados $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, k]$. Para calcular las probabilidades a posteriori, el modelo de voz limpia $\mathcal{M}_x$ se ha de adaptar a las condiciones acústicas de la frase que se reconoce, dando como resultado el modelo de voz ruidosa $\mathcal{M}_y$. La adaptación del modelo de voz se puede realizar mediante cualquiera de las técnicas estudiadas en la sección 2.2.2.2: PMC, VTS o Algonquin. Suponiendo que el modelo de ruido usado en la adaptación consta de una gaussiana, $\mathcal{M}_n = \mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$, cualquiera de estas tres técnicas obtendrá un GMM adaptado $\mathcal{M}_y$ con igual número de componentes que el de partida $\mathcal{M}_x$,

$$p(\boldsymbol{y}|\mathcal{M}_y) = \sum_{k=1}^{M} P(k)\mathcal{N}\left(\boldsymbol{y}; \boldsymbol{\mu}_y^{(k)}, \boldsymbol{\Sigma}_y^{(k)}\right). \tag{2.107}$$

Aparte de las tres técnicas mencionadas, en la literatura también podemos encontrar distintas aproximaciones numéricas que permiten una mejor adaptación de los modelos acústicos al salvar parte de las limitaciones de estas técnicas. Ejemplos representativos de estas aproximaciones numéricas son la transformada *unscented* [173] o los filtros de partículas [91]. Otra estrategia alternativa para estimar el modelo $\mathcal{M}_y$ que también emplea un modelo de distorsión, es la que propondremos en la sección 4.3. En este caso, como ya veremos, el modelo obtenido no será un GMM, pero sí que permitirá la evaluación analítica de las probabilidades $P(k|\boldsymbol{y})$ usadas por el estimador de la ecuación (2.106).

Por otro lado, los valores esperados $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, k]$ $(k = 1, \ldots, M)$ que aparecen en el estimador MMSE se pueden calcular de la siguiente forma,

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, k] = \int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}, k) d\boldsymbol{x}. \tag{2.108}$$

Al igual que en la adaptación del modelo de voz, para el cálculo de esta integral se hace necesario recurrir a ciertas aproximaciones debido a la no linealidad de los modelos de distorsión en los dominios log-Mel o MFCC (ver sección 2.1). Las aproximaciones más comunes vuelven a ser VTS [199, 239, 252, 258], Algonquin [94, 165], la transformada *unscented* [173] o el modelo de enmascaramiento que estudiaremos en el capítulo

4. Centrándonos en la aproximación VTS por ser la más difundida actualmente, podemos encontrar varias propuestas para el cálculo de $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y},k]$. En [199, 239, 258], por ejemplo, este valor se deduce del modelo de distorsión recogido en la ecuación (2.14), el cual reescribimos a continuación por comodidad,

$$\boldsymbol{y} \approx \boldsymbol{x} + \boldsymbol{C}\log\left(\boldsymbol{1} + e^{\boldsymbol{C}^{-1}(\boldsymbol{n}-\boldsymbol{x})}\right) = \boldsymbol{x} + \boldsymbol{g}(\boldsymbol{x},\boldsymbol{n}), \tag{2.109}$$

donde consideremos que las características empleadas son MFCCs.

A partir del modelo anterior, el valor esperado $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y},k]$ se puede expresar como,

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y},k] = \boldsymbol{y} - \mathbb{E}[\boldsymbol{g}(\boldsymbol{x},\boldsymbol{n})|\boldsymbol{y},k]. \tag{2.110}$$

En los trabajos [199, 239, 258] se asume que la función de discrepancia $\boldsymbol{g}(\boldsymbol{x},\boldsymbol{n})$ es suave dentro de cada región k del espacio, por lo que su valor esperado puede aproximarse por

$$\mathbb{E}[\boldsymbol{g}(\boldsymbol{x},\boldsymbol{n})|\boldsymbol{y},k] \approx \boldsymbol{g}(\boldsymbol{\mu}_x^{(k)},\boldsymbol{\mu}_n), \tag{2.111}$$

siendo $\boldsymbol{\mu}_x^{(k)}$ la media de la gaussiana k del GMM $\mathcal{M}_x$ y $\boldsymbol{\mu}_n$ la media de la gaussiana que modela el ruido (asumimos un modelo con una gaussiana).

En [252] se propone un cómputo alternativo de $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y},k]$. Como se puede apreciar, este valor esperado coincide con la media condicional de la distribución $p(\boldsymbol{x}|\boldsymbol{y},k)$. Asumiendo que la distribución conjunta $p(\boldsymbol{x},\boldsymbol{y}|k)$ es gaussiana, entonces $p(\boldsymbol{x}|\boldsymbol{y},k)$ también lo será y su media vendrá dada por

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y},k] \approx \boldsymbol{\mu}_x^{(k)} + \boldsymbol{\Sigma}_{xy}^{(k)}\boldsymbol{\Sigma}_y^{(k)^{-1}}\left(\boldsymbol{y} - \boldsymbol{\mu}_y^{(k)}\right). \tag{2.112}$$

En la ecuación anterior todos los términos son conocidos salvo la matriz de correlación cruzada $\boldsymbol{\Sigma}_{xy}^{(k)}$. Es posible demostrar [252] que bajo la aproximación VTS esta matriz se puede aproximar mediante

$$\boldsymbol{\Sigma}_{xy}^{(k)} \approx \boldsymbol{\Sigma}_x^{(k)}\boldsymbol{J}_x^{(k)^\top}, \tag{2.113}$$

siendo $\boldsymbol{J}_x^{(k)}$ la matriz jacobiana de la ecuación (2.53).

Aunque existen multitud de alternativas para el cómputo de los distintos términos que intervienen en el estimador MMSE, éstas quedan fuera del interés de esta tesis. Pasamos, por tanto, al estudio de la compensación basada en datos estéreo.

2.2.3.4. Compensación basada en datos estéreo

La tarea de compensación de las características de voz se ve simplificada si es posible conocer a priori las características de los entornos acústicos donde el sistema de RAH

se empleará. En particular, la magnitud de la distorsión de las características de voz es función de dos variables principalmente: (i) la densidad de potencia espectral del ruido presente en el entorno y (ii) el nivel de SNR. Aunque en la gran mayoría de las situaciones estos parámetros son desconocidos y, por tanto, deben de estimarse a partir de la propia frase que se reconoce, en ciertos contextos es posible conocer a priori, con una fiabilidad relativamente alta, las características de los ruidos que van a poder darse en el transcurso de la interacción entre el usuario y el sistema de reconocimiento. Por ejemplo, en los sistemas de navegación GPS (*Global Positioning System*, sistema de posicionamiento global) controlados por voz que se instalan en algunos automóviles, es asumible que el número de tipos de ruidos que pueden darse será limitado (p.ej. ruido del motor del coche, tráfico, pasajeros hablando, etc.). De forma similar, podemos considerar que en un ambiente de oficina la variabilidad acústica que puede darse estará, en cierta medida, limitada.

Bajo el supuesto anterior, esto es, disponer de un conocimiento a priori sobre el rango de distorsiones acústicas que pueden darse, en la literatura se han propuesto diversas técnicas que estiman los vectores de voz limpia (desconocidos) a partir de los vectores de voz contaminada (observados) y de un conjunto de transformaciones dependientes del tipo de ambiente acústico. Este conjunto de transformaciones se calculan durante la fase de entrenamiento del sistema para cada tipo de ruido y nivel de SNR usando bases de datos estéreo, que contienen grabaciones simultáneas de una señal de voz a una SNR alta y una o varias versiones de la misma señal degradadas con distintos tipos de ruido y SNRs[1]. Usando las grabaciones estéreo, es posible estimar de manera precisa (comparando las grabaciones de voz limpia con las de voz ruidosa) el conjunto de transformaciones que modelan la degradación de la voz en el entorno en cuestión.

Dado que el entorno acústico es conocido a priori y, por tanto, no hay necesidad de estimarlo, podemos considerar que las técnicas que trabajan con datos estéreo son más eficientes que aquellas en las que se deben estimar los parámetros del mismo (p.ej. VTS [199] o Algonquin [165]). Asimismo, las transformaciones que se estiman a partir de los datos estéreo pueden modelar distintos tipos de distorsiones como, por ejemplo, aquellas debidas al ruido aditivo, convolutivo, etc. Como inconvenientes de estas técnicas tenemos, en primer lugar, la necesidad de disponer de un conocimiento a priori sobre los ruidos y niveles de SNR que se van a dar durante la fase de explotación del sistema, si bien se han desarrollado versiones "ciegas" de estas técnicas donde los parámetros del modelo de ruido son estimados en línea [44, 199]. Por otro lado, dado que en la mayoría de los casos el entrenamiento de las técnicas que emplean datos

[1]Dada la gran dificultad práctica que conlleva la grabación de bases de datos estéreo con ruidos reales, éstas se suelen generar añadiendo artificialmente ruido a grabaciones de voz limpia. Un caso representativo de ello lo tenemos en la base de datos Aurora2 [141].

estéreo se hace con un conjunto limitado de tipos de ruido y SNRs, el rendimiento de estas técnicas para ruidos no considerados durante el entrenamiento tiende a ser pobre [122].

Vistas las características generales de las técnicas que emplean datos estéreo, en los siguientes apartados describiremos en detalle algunas de las técnicas más representativas dentro de esta categoría.

SPLICE. SPLICE (*Stereo-based Piecewise Linear Compensation for Environments*, compensación lineal por partes para entornos basada en datos estéreo) [67, 68, 78, 79] es una de las técnicas más conocidas y a la vez más eficientes para la compensación de las distorsiones producidas por el ruido en las características de voz (MFCCs, generalmente). Al contrario que otras técnicas (p.ej. VTS [199] o Algonquin [165]), SPLICE emplea un modelo de distorsión no paramétrico que se estima durante la fase de entrenamiento del sistema usando grabaciones estéreo. El objetivo de esta técnica es estimar el vector de voz limpia subyacente $\boldsymbol{x}$, supuesto que el vector de voz ruidosa $\boldsymbol{y}$ y el modelo de voz distorsionada $\mathcal{M}_y$ son dados. Para ello se recurre a un estimador del tipo MMSE (*Minimum Mean Square Error*, mínimo error cuadrático medio) cuyas expresiones se derivan empleando un modelo de voz conjunta $p(\boldsymbol{x}, \boldsymbol{y})$ y el modelo de distorsión no paramétrico mencionado anteriormente.

En primer lugar, SPLICE supone que la densidad de probabilidad conjunta $p(\boldsymbol{x}, \boldsymbol{y})$ puede representarse mediante un GMM con M componentes,

$$p(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{M} p(\boldsymbol{x}, \boldsymbol{y}, k) = \sum_{k=1}^{M} p(\boldsymbol{x}|\boldsymbol{y}, k)p(\boldsymbol{y}, k). \tag{2.114}$$

Esto modelo conjunto, como se verá, no se estima de forma explícita, sino que se deriva de las expresiones relativas al modelo de voz ruidosa $\mathcal{M}_y$ y de un conjunto de transformaciones de compensación. Por otra parte, los parámetros del GMM $\mathcal{M}_y$ se estiman usando voz contaminada con un tipo de ruido y SNR determinados. Luego, para cada condición acústica distinta (tipo de ruido y SNR), SPLICE estima un GMM diferente. Para simplificar el estudio de esta técnica, supondremos que cada frase de entrenamiento viene etiquetada con los parámetros del ambiente acústico con la que ha sido contaminada. En los casos en los que no se disponga de este etiquetado, será necesario emplear un clasificador de ruidos para estimar el tipo de ruido y nivel de SNR presentes en la señal [79]. Tras estimar la condición acústica embebida en cada frase de entrenamiento, el algoritmo EM [66] se emplea para obtener el GMM que modela los parámetros de voz en dicho ambiente,

$$p(\boldsymbol{y}) = \sum_{k=1}^{M} P(k)p(\boldsymbol{y}|k) = \sum_{k=1}^{M} P(k)\mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_y^{(k)}, \boldsymbol{\Sigma}_y^{(k)}). \tag{2.115}$$

El segundo punto clave de SPLICE es el modelo de distorsión que emplea. Recordemos que, en el domino cepstral, la relación entre el vector de voz distorsionada y los vectores de voz limpia y de ruido viene dada por (ver sección 2.1):

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{h} + \boldsymbol{C}\log(\boldsymbol{1} + e^{\boldsymbol{C}^{-1}(\boldsymbol{n}-\boldsymbol{x}-\boldsymbol{h})}) = \boldsymbol{x} + \boldsymbol{r}(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{h}). \tag{2.116}$$

Para simplificar el modelo de distorsión anterior, SPLICE hace dos suposiciones. En primer lugar, se supone que la función de discrepancia $\boldsymbol{r}(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{h})$ puede expresarse como una función del vector de voz ruidosa, es decir,

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{r}(\boldsymbol{y}). \tag{2.117}$$

Por otra parte, SPLICE supone que $\boldsymbol{r}(\boldsymbol{y})$ es una función suave de $\boldsymbol{y}$ que varía suavemente dentro de cada una de las regiones que define el GMM de voz ruidosa. Bajo este supuesto, $\boldsymbol{r}(\boldsymbol{y})$ se aproxima mediante la siguiente combinación lineal:

$$\boldsymbol{r}(\boldsymbol{y}) \approx \sum_{k=1}^{M} \boldsymbol{r}_k p(k|\boldsymbol{y}), \tag{2.118}$$

donde $\boldsymbol{r}_k$ define la transformación que se ha de aplicar a los vectores pertenecientes a la región k-ésima del GMM para compensar la distorsión introducida por el ruido. Aunque la técnica SPLICE básica supone una corrección aditiva por componente del GMM, una transformación afín más precisa del tipo $\boldsymbol{A}_k\boldsymbol{y} + \boldsymbol{r}_k$ por gaussiana también es posible.

Bajo la aproximación anterior del modelo de distorsión, SPLICE aproxima la probabilidad condicional $p(\boldsymbol{x}|\boldsymbol{y}, k)$ que aparece en (2.114) mediante una gaussiana con los siguientes parámetros:

$$p(\boldsymbol{x}|\boldsymbol{y}, k) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{y} - \boldsymbol{r}_k, \boldsymbol{\Gamma}_k), \tag{2.119}$$

donde la matriz de covarianza $\boldsymbol{\Gamma}_k$ modela el error en el que se incurre al aproximar $\boldsymbol{x}$ por $\boldsymbol{y} - \boldsymbol{r}_k$. Más adelante se verá cómo se obtiene esta matriz junto con los vectores de corrección $\boldsymbol{r}_k$.

Usando los elementos listados anteriormente, SPLICE obtiene una estimación del vector de voz limpia $\hat{\boldsymbol{x}}$ como el valor esperado de $\boldsymbol{x}$ dado el vector ruidoso observado $\boldsymbol{y}$,

$$\begin{aligned}
\hat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}] &= \int_{\boldsymbol{x}} \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x} \approx \int_{\boldsymbol{x}} (\boldsymbol{y} - \boldsymbol{r}(\boldsymbol{y}))\, p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x} \\
&\approx \boldsymbol{y} - \int_{\boldsymbol{x}} \sum_{k=1}^{M} \boldsymbol{r}_k p(\boldsymbol{x}, k|y) d\boldsymbol{x} = \boldsymbol{y} - \sum_{k=1}^{M} \boldsymbol{r}_k \int_{\boldsymbol{x}} p(\boldsymbol{x}, k|\boldsymbol{y}) d\boldsymbol{x} \\
&= \boldsymbol{y} - \sum_{k=1}^{M} P(k|\boldsymbol{y}) \boldsymbol{r}_k.
\end{aligned} \tag{2.120}$$

Para el cálculo de la probabilidad a posteriori $P(k|\boldsymbol{y})$ hemos de recurrir a la regla de Bayes,

$$P(k|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|k)P(k)}{\sum_{k'=1}^{M} p(\boldsymbol{y}|k')P(k')}. \tag{2.121}$$

Como se puede apreciar en la ecuación (2.120), la técnica SPLICE implementa una transformación aditiva para el cálculo del vector de voz limpia. A su vez, la transformación que se aplica a $\boldsymbol{y}$ se calcula mediante una combinación lineal de una serie de vectores de corrección previamente calculados $\boldsymbol{r}_k$ $(k = 1, \ldots, M)$ ponderados por sus correspondientes probabilidades $P(k|\boldsymbol{y})$.

En el estimador de la ecuación (2.120) se ha supuesto que la condición acústica presente en la frase a compensar es conocida de antemano. De nuevo, esta suposición no es realista, por lo que se ha de recurrir a un detector de ruidos que identifique el tipo de ruido y SNR que contamina a $\boldsymbol{y}$. En este sentido, una de las estrategias más simples consiste en usar los propios GMMs entrenados para cada uno de los entornos acústicos $\mathcal{M}_y^e$ $(e = 1, \ldots, E)$ para tal propósito. Notando como $\hat{\boldsymbol{x}}^e$ al vector de voz estimado con el modelo $\mathcal{M}_y^e$ en la ecuación (2.120), la estima final multientorno vendría dada por

$$\hat{\boldsymbol{x}} = \sum_{e=1}^{E} P(e|\boldsymbol{y})\hat{\boldsymbol{x}}^e. \tag{2.122}$$

Para finalizar con la presentación de la técnica SPLICE, describiremos el cálculo de $\boldsymbol{r}_k$ y $\boldsymbol{\Gamma}_k$ para cada entorno acústico. Los valores de estos dos términos deben ser calculados durante la fase de entrenamiento de acuerdo a cierto criterio de optimización previamente establecido usando las grabaciones estéreo disponibles. Tradicionalmente el criterio escogido ha sido el de máxima verosimilitud (ML), aunque recientemente se está extendiendo el uso de criterios basados en el entrenamiento discriminatorio [77]. De acuerdo al criterio ML, los valores de $\boldsymbol{r}_k$ y $\boldsymbol{\Gamma}_k$ son aquellos que maximizan la siguiente probabilidad:

$$\begin{aligned}\langle \boldsymbol{r}_k, \boldsymbol{\Gamma}_k \rangle &= \underset{\langle \boldsymbol{r}, \boldsymbol{\Gamma} \rangle}{\operatorname{argmax}} \sum_{t=1}^{T} P(k|\boldsymbol{y}_t) \log p(\boldsymbol{x}_t|\boldsymbol{y}_t, k) \\ &= \underset{\langle \boldsymbol{r}, \boldsymbol{\Gamma} \rangle}{\operatorname{argmax}} \sum_{t=1}^{T} P(k|\boldsymbol{y}_t) \log \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{y}_t - \boldsymbol{r}, \boldsymbol{\Gamma}).\end{aligned} \tag{2.123}$$

Derivando la ecuación anterior e igualando a cero, se obtienen las siguientes expresiones para el cálculo de $\boldsymbol{r}_k$ y $\boldsymbol{\Gamma}_{\boldsymbol{k}}$:

$$\boldsymbol{r}_k = \frac{\sum_{t=1}^{T} P(k|\boldsymbol{y}_t)(\boldsymbol{y}_t - \boldsymbol{x}_t)}{\sum_{t=1}^{T} P(k|\boldsymbol{y}_t)}, \tag{2.124}$$

$$\boldsymbol{\Gamma}_k = \frac{\sum_{t=1}^{T} P(k|\boldsymbol{y}_t)(\boldsymbol{y}_t - \boldsymbol{x}_t)^2}{\sum_{t=1}^{T} P(k|\boldsymbol{y}_t)} - \boldsymbol{r}_k^2. \tag{2.125}$$

SPLICE, tal y como se ha presentado en este apartado, puede considerarse como una versión mejorada de la técnica FCDCN (*Fixed Codeword Dependent Cepstral Normalization*, normalización cepstral dependiente de un símbolo fijo del diccionario) propuesta originalmente por Acero en [8, 10]. En lugar de emplear un GMM para modelar el espacio de características contaminadas, FCDCN utiliza un diccionario VQ para tal fin. Análogamente a la forma de proceder de SPLICE, a partir de este diccionario se estiman un conjunto de vectores de corrección, uno para cada celda del diccionario. La estima final se obtiene sustrayendo a $\boldsymbol{y}$ el vector de corrección $\boldsymbol{r}_{k^*}$ obtenido para la celda VQ a la cual $\boldsymbol{y}$ pertenece. Esta forma de proceder de FCDCN sería equivalente, en cierta forma, a seleccionar en SPLICE únicamente la corrección asociada a la gaussiana más probable $k^* = \text{argmax}_k\, p(k|\boldsymbol{y})$.

POF. La técnica POF (*Probilistic Optimum Filtering*, filtrado óptimo probabilístico) [209] puede considerarse como una versión extendida de SPLICE. Al igual que esta última, POF modela el espacio de características de voz distorsionadas mediante un GMM. No obstante, en lugar de asociar un vector de corrección $\boldsymbol{r}_k$ a cada región del espacio tal y como hace SPLICE, la transformación que POF implementa se deriva de un conjunto de filtros multidimensionales (uno por región) que se aplican sobre un conjunto de vectores alrededor de la observación actual,

$$\boldsymbol{Y}_t = \left[\boldsymbol{y}_{t-p}^\top, \ldots, \boldsymbol{y}_t^\top, \ldots, \boldsymbol{y}_{t+p}^\top, \boldsymbol{1}^\top\right]^\top, \tag{2.126}$$

siendo $\boldsymbol{1}$ un vector de unos del mismo tamaño que los observados.

El vector de voz limpia se estima siguiendo un criterio MMSE a partir de la observación actual $\boldsymbol{y}_t$ y del supervector $\boldsymbol{Y}_t$,

$$\hat{\boldsymbol{x}}_t = \sum_{k=1}^{M} P(k|\boldsymbol{y}_t)\boldsymbol{W}_k^\top \boldsymbol{Y}_t = \left(\sum_{k=1}^{M} P(k|\boldsymbol{y}_t)\boldsymbol{W}_k^\top\right)\boldsymbol{Y}_t, \tag{2.127}$$

donde $\boldsymbol{W}_k$ es la matriz con los coeficientes del filtro. Los valores de esta matriz se obtienen usando datos estéreo y minimizando el siguiente error cuadrático:

$$E_k = \sum_{t=p}^{N-1-p} P(k|\boldsymbol{y}_t)\left\|\boldsymbol{x}_t - \boldsymbol{W}_k^\top \boldsymbol{Y}_t\right\|^2. \tag{2.128}$$

En comparación con SPLICE, la técnica POF cuenta con la ventaja de permitir explotar la información relativa a la evolución temporal de la voz durante el proceso de compensación. Esto, tal y como se muestra en [209], es importante, ya que produce mejores estimadores de características y, por consiguiente, incrementa la precisión del reconocedor. Debida a ésta y otras razones, en el capítulo 5 se analizará en profundidad el modelado temporal de la voz en el contexto de las técnicas de compensación propuestas en este trabajo.

RATZ. RATZ (*Multivariate-Gaussian-based cepstral normalization*, normalización cepstral basada en gaussianas multivariantes) [199, 200] es otra técnica de compensación que emplea datos estéreo. Al contrario que las técnicas revisadas hasta el momento (SPLICE y POF), la compensación de la técnica RATZ se construye sobre un modelo de voz limpia consistente en un GMM con M componentes. Por otra parte, RATZ asume que el efecto del ruido sobre cada componente del modelo se traduce en un desplazamiento de la media y un incremento de la varianza[1], por lo que la estadística de la voz contaminada se puede calcular de la siguiente forma

$$p(\boldsymbol{y}) = \sum_{k=1}^{M} P(k)p(\boldsymbol{y}|k) = \sum_{k=1}^{M} P(k)\mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_x^{(k)} + \boldsymbol{r}_k, \boldsymbol{\Sigma}_x^{(k)} + \boldsymbol{\Gamma}_k), \tag{2.129}$$

donde $\boldsymbol{r}_k$ y $\boldsymbol{\Gamma}_k$ modelan el efecto del ruido sobre la gaussiana k. El cálculo de estos factores de corrección variará en función de si durante la fase de entrenamiento se disponen de datos estéreo que puedan ser utilizados para estimarlos. Si no existen tales datos, en [199] se propone un método iterativo denominado RATZ ciego (*Blind RATZ*) que permite calcular los factores anteriores. En caso contrario, es decir, si hay datos estéreo disponibles, los valores de $\boldsymbol{r}_k$ y $\boldsymbol{\Gamma}_k$ se estiman siguiendo un criterio ML,

$$\boldsymbol{r}_k = \frac{\sum_{t=1}^{T} P(k|\boldsymbol{x}_t)(\boldsymbol{y}_t - \boldsymbol{x}_t)}{\sum_{t=1}^{T} P(k|\boldsymbol{x}_t)}, \tag{2.130}$$

$$\boldsymbol{\Gamma}_k = \frac{\sum_{t=1}^{T} P(k|\boldsymbol{x}_t)(\boldsymbol{y}_t - \boldsymbol{\mu}_x^{(k)} - \boldsymbol{r}_k)(\boldsymbol{y}_t - \boldsymbol{\mu}_x^{(k)} - \boldsymbol{r}_k)^\top}{\sum_{t=1}^{T} P(k|\boldsymbol{x}_t)} - \boldsymbol{\Sigma}_x^{(k)}. \tag{2.131}$$

Por último, el vector de voz limpia se estima siguiendo un criterio MMSE en base al modelo de voz contaminada dado por las ecuaciones (2.129)-(2.131),

$$\begin{aligned}
\hat{\boldsymbol{x}} &= \int_{\boldsymbol{x}} \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{y} \approx \boldsymbol{y} - \int_{\boldsymbol{x}} \boldsymbol{r}(\boldsymbol{x}) p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x} \\
&\approx \boldsymbol{y} - \sum_{k=1}^{M} P(k|\boldsymbol{y}) \boldsymbol{r}_k \int_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}, k) d\boldsymbol{x} \\
&\approx \boldsymbol{y} - \sum_{k=1}^{M} P(k|\boldsymbol{y}) \boldsymbol{r}_k.
\end{aligned} \tag{2.132}$$

Como se puede apreciar, la expresión obtenida en (2.132) para la técnica RATZ es equivalente a la del estimador SPLICE de la ecuación (2.120). Aunque la expresión obtenida es la misma, ambas técnicas parten de supuestos diferentes: SPLICE modela a priori el espacio de características contaminadas $p(\boldsymbol{y})$ y de ahí deriva las transformaciones oportunas, mientras que RATZ representa $p(\boldsymbol{x})$, obteniendo de este modelo las

[1] Se asume que los pesos de las gaussianas, $P(k)$ $(k = 1, \ldots, M)$, no se ven modificados por la adición del ruido.

transformaciones oportunas. En [235] se puede encontrar una comparativa de estas dos técnicas donde se muestra que el rendimiento de SPLICE es superior al de RATZ. La razón de ello es, según los autores, la mayor precisión en el modelado de $p(\boldsymbol{y})$ por parte de SPLICE, lo que supone un menor error en la estimación de la señal de voz limpia.

MEMLIN. MEMLIN (*Multi-Environment Model-based Linear Normalization*, normalización lineal basada en modelos multi-entorno) [44] es otra técnica no parámetrica de compensación de características que emplea datos estéreo recolectados de varios entornos acústicos ($e = 1, \ldots, E$). Para cada uno de los entornos, MEMLIN entrena un GMM $\mathcal{M}_y^e$ que modela la voz contaminada con el ruido propio de dicho entorno,

$$p(\boldsymbol{y}|e) = \sum_{k_y^e}^{M_y^e} P(k_y^e|e)\mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_{k_y^e}, \boldsymbol{\Sigma}_{k_y^e}), \tag{2.133}$$

donde k_y^e es el índice utilizado para referenciar las gaussianas de $\mathcal{M}_y^e$ y M_y^e es el número de gaussianas de este GMM.

La motivación detrás de esta técnica es obtener una mejor representación de la transformación que se produce en la voz a causa del ruido. Para conseguir este propósito, MEMLIN modela no sólo el espacio de voz ruidosa (para cada entorno), sino también el espacio de características limpias $p(\boldsymbol{x})$. Para representar este espacio de nuevo se emplea un GMM cuyas componentes indicaremos mediante el índice k_x. A partir de estos modelos de voz, MEMLIN estima un vector de corrección $\boldsymbol{r}_{k_x,k_y^e}$ con la distorsión promedio que se produce entre cada par de componentes (k_x, k_y^e),

$$\boldsymbol{r}_{k_x,k_y^e} = \frac{\sum_{t=1}^{T} p(k_x|\boldsymbol{x}_t)p(k_y^e|\boldsymbol{y}_t)(\boldsymbol{y}_t - \boldsymbol{x}_t)}{\sum_{t=1}^{T} p(k_x|\boldsymbol{x}_t)p(k_y^e|\boldsymbol{y}_t)} \tag{2.134}$$

Como se aprecia en la figura 2.11, estos vectores de corrección permiten un modelado más preciso de la distorsión que se produce debido al entorno, entendiendo por entorno acústico cualquier distorsión que afecte a las características.

En base al modelado dual anterior, MEMLIN propone el siguiente estimador:

$$\begin{aligned}\hat{\boldsymbol{x}} &= \boldsymbol{y} - \int_{\boldsymbol{x}} \sum_{e}^{E} \sum_{k_y^e}^{M_y^e} \sum_{k_x}^{M_x} \boldsymbol{r}_{k_x,k_y^e} p(\boldsymbol{x}, k_x, e, k_y^e|\boldsymbol{y}) d\boldsymbol{x} \\ &= \boldsymbol{y} - \sum_{e}^{E} \sum_{k_y^e}^{M_y^e} \sum_{k_x}^{M_x} \boldsymbol{r}_{k_x,k_y^e} p(e|\boldsymbol{y}) p(k_y^e|\boldsymbol{y}, e) p(k_x|\boldsymbol{y}, e, k_y^e),\end{aligned} \tag{2.135}$$

donde la probabilidad a posteriori del entorno e-ésimo $p(e|\boldsymbol{y})$ se calcula recursivamente de acuerdo a la siguiente expresión,

$$p(e|\boldsymbol{y}_t) = \beta \cdot p(e|\boldsymbol{y}_{t-1}) + (1 - \beta)\frac{p(\boldsymbol{y}_t|e)}{\sum_{e'}^{E} p(\boldsymbol{y}_t|e')} \tag{2.136}$$

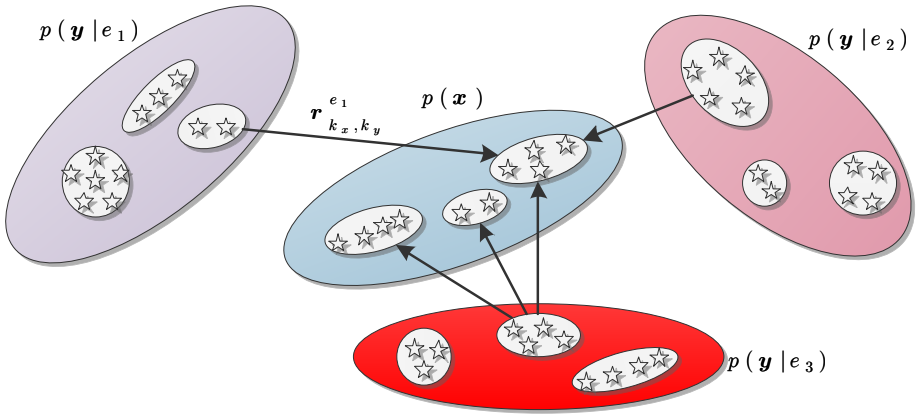

Figura 2.11: Representación esquemática de la transformación cepstral implementada por MEMLIN.

donde $\beta = 0,98$ y la probabilidad $p(k_x|\boldsymbol{y}, e, k_y^e) \approx p(k_x|k_y^e, e)$ define un modelo cruzado que representa las transformaciones que se producen entre las gaussianas del espacio sucio de cada entorno y las gaussianas del entorno limpio de referencia. Esta probabilidad se calcula como

$$p(k_x|k_y^e, e) = \frac{\sum_{t=1}^{T} p(\boldsymbol{x}_t|k_x)p(\boldsymbol{y}_t|k_y^e)p(k_x)p(k_y^e|e)}{\sum_{t=1}^{T}\sum_{k'_x} p(\boldsymbol{x}_t|k'_x)p(\boldsymbol{y}_t|k_y^e)p(k'_x)p(k_y^e|e)}. \tag{2.137}$$

Además de la técnica MEMLIN, se proponen otras tres técnicas de compensación derivadas que emplean grabaciones estéreo [44]: P-MEMLIN (*Polynomial MEMLIN*, MEMLIN polinómico), MEMHIN (*Multi-environvent Model-based Histogram Normalization*, normalización de histograma basada en modelos multientorno) y PD-MEMLIN (*Phoneme-Dependent MEMLIN*, MEMLIN dependiente del fonema). La principal diferencia entre estas técnicas se debe al tipo de transformación usada para representar la degradación entre cada par de componentes. Si en la técnica MEMLIN básica esta transformación se modela mediante un vector de corrección, esto es, $\boldsymbol{x} \approx \Psi(\boldsymbol{y}, k_x, k_y^e) = \boldsymbol{y} - \boldsymbol{r}_{k_x,k_y^e}$, P-MEMLIN implementa una transformación afín $\Psi(\boldsymbol{y}, k_x, k_y^e) = \boldsymbol{A}_{k_x,k_y^e}\boldsymbol{y} + \boldsymbol{b}_{k_x,k_y^e}$ y MEMHIN propone una transformación no lineal basada en la técnica de ecualización de histogramas. En cuanto a PD-MEMLIN, esta técnica introduce un mejor modelado de la voz al considerar la degradación de las distintas unidades acústicas (fonemas) debido al ruido. Este modelado más complejo permite a la técnica PD-MEMLIN obtener mejores tasas de reconocimiento en comparación con el resto de técnicas anteriores [43].

SSM y FE-Joint. Para terminar el estudio de las técnicas de compensación que estiman sus transformaciones a partir de datos estéreo, en este apartado revisaremos dos técnicas que, al igual que MEMLIN, modelan ambos espacios de características: el espacio de características limpias $p(\boldsymbol{x})$ y el de voz distorsionada $p(\boldsymbol{y})$. No obstante, este modelado no se realiza de forma independiente como en MEMLIN, sino de forma conjunta $p(\boldsymbol{x}, \boldsymbol{y})$. Estas técnicas son conocidas como SSM (*Stereo-based Stochastic Mapping*, transformación estocástica basada en datos estéreo) [11, 61, 62] y FE-Joint (*Front-end Joint uncertainty decoding*, descodificación con incertidumbre usando un modelado conjunto de la voz en el extractor de características) [175, 177]. La mayor diferencia entre ambas radica en que FE-Joint incluye, además, un modulo que permite explotar la incertidumbre de los vectores de voz compensados en el reconocedor. De esta forma, los vectores de características cuya fiabilidad es menor tienen menos peso durante al etapa de reconocimiento. Este tipo de estrategias híbridas que combinan estimación de las características de voz limpia y propagación de la incertidumbre al reconocedor serán analizadas con mayor detalle en la sección 2.2.4.

Como viene siendo habitual, ambas técnicas suponen la existencia de una base de datos estéreo que contiene grabaciones de voz limpia y contaminada bajo un entorno acústico determinado. A partir del conjunto de vectores de características $\{(\boldsymbol{x}_t, \boldsymbol{y}_t)\}$ $(t = 1, \ldots, T)$ presentes en la base de datos, es posible definir el supervector $\boldsymbol{z}_t = (\boldsymbol{x}_{t-p}, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_{t+p}, \boldsymbol{y}_{t-l}, \ldots, \boldsymbol{y}_t, \ldots, \boldsymbol{y}_{t+l})$ como la concatenación de $2 \cdot p + 1$ vectores del canal de voz limpia y $2 \cdot l + 1$ vectores contaminados. Con objeto de simplificar la presentación de ambas técnicas, supondremos que $p = l = 0$, luego $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$, donde además hemos omitido el subíndice de tiempo para mayor claridad. La extensión de SSM y FE-Joint para los casos en los que $p, l > 0$ es directa. Usando las grabaciones estéreo, ambas técnicas construyen un GMM que modela la probabilidad conjunta,

$$p(\boldsymbol{z}) = \sum_{k=1}^{M} P(k) \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_z^{(k)}, \boldsymbol{\Sigma}_z^{(k)}). \tag{2.138}$$

A partir del modelo anterior, es posible derivar el siguiente estimador MMSE para el vector de voz limpia

$$\begin{aligned} \hat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}] &= \int_{\boldsymbol{x}} \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x} = \sum_{k=1}^{M} \int_{\boldsymbol{x}} \boldsymbol{x} p(\boldsymbol{x}, k|\boldsymbol{y}) d\boldsymbol{x} \\ &= \sum_{k=1}^{M} P(k|\boldsymbol{y}) \underbrace{\int_{\boldsymbol{x}} \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}, k) d\boldsymbol{x}}_{\hat{\boldsymbol{x}}^{(k)}}, \end{aligned} \tag{2.139}$$

donde los términos $P(k|\boldsymbol{y})$ y $\hat{\boldsymbol{x}}^{(k)}$ que aparecen en la ecuación anterior se derivan del GMM de la ecuación (2.138), tal y como se verá a continuación.

En primer lugar, usando la regla de Bayes, $P(k|\boldsymbol{y})$ puede reescribirse de la siguiente forma:

$$P(k|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|k)P(k)}{\sum_{k'=1}^{M} P(k')p(\boldsymbol{y}|k')}. \tag{2.140}$$

SSM y FE-Joint suponen que las probabilidades a priori de la gaussianas, $P(k)$, no se ven modificadas por la adición de ruido. Esto es comprensible si el ruido no incluye reverberaciones que puedan modificar estas estadísticas a priori. Por otro lado, $p(\boldsymbol{y}|k)$ es la probabilidad marginal de $\boldsymbol{y}$ para la gaussiana k-ésima. Para obtener los parámetros de esta PDF, primero se dividen los parámetros de $p(\boldsymbol{z}|k)$ en sus componentes $\boldsymbol{x}$ y $\boldsymbol{y}$:

$$\boldsymbol{\mu}_z^{(k)} = \begin{bmatrix} \boldsymbol{\mu}_x^{(k)} \\ \boldsymbol{\mu}_y^{(k)} \end{bmatrix}, \tag{2.141}$$

$$\boldsymbol{\Sigma}_z^{(k)} = \begin{bmatrix} \boldsymbol{\Sigma}_x^{(k)} & \boldsymbol{\Sigma}_{xy}^{(k)} \\ \boldsymbol{\Sigma}_{yx}^{(k)} & \boldsymbol{\Sigma}_y^{(k)} \end{bmatrix} \tag{2.142}$$

y, por tanto, $p(\boldsymbol{y}|k) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_y^{(k)}, \boldsymbol{\Sigma}_y^{(k)})$.

En cuanto al valor de $\hat{\boldsymbol{x}}^{(k)}$ en (2.139), éste se corresponde con la media de la densidad de probabilidad condicional $p(\boldsymbol{x}|\boldsymbol{y}, k)$. Es bien conocido que si $p(\boldsymbol{x}, \boldsymbol{y}|k)$ es una gaussiana multivariante, entonces $p(\boldsymbol{x}|\boldsymbol{y}, k) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{x|y}^{(k)}, \boldsymbol{\Sigma}_{x|y}^{(k)})$ es también una normal con los siguientes parámetros [211]:

$$\boldsymbol{\mu}_{x|y}^{(k)} = \boldsymbol{\mu}_x^{(k)} + \boldsymbol{\Sigma}_{xy}^{(k)}\boldsymbol{\Sigma}_y^{(k)^{-1}}(\boldsymbol{y} - \boldsymbol{\mu}_y^{(k)}), \tag{2.143}$$

$$\boldsymbol{\Sigma}_{x|y}^{(k)} = \boldsymbol{\Sigma}_x^{(k)} - \boldsymbol{\Sigma}_{xy}^{(k)}\boldsymbol{\Sigma}_y^{(k)^{-1}}\boldsymbol{\Sigma}_{yx}^{(k)} \tag{2.144}$$

y $\hat{\boldsymbol{x}}^{(k)} = \boldsymbol{\mu}_{x|y}^{(k)}$.

Así, podemos ver que las técnicas SSM y FE-Joint implementan una transformación afín

$$\hat{\boldsymbol{x}}^{(k)} = \boldsymbol{A}_k\boldsymbol{y} + \boldsymbol{b}_k, \tag{2.145}$$

por cada componente k, donde los parámetros de la transformación afín vienen dados por:

$$\boldsymbol{A}_k = \boldsymbol{\Sigma}_{xy}^{(k)}\boldsymbol{\Sigma}_y^{(k)^{-1}}, \tag{2.146}$$

$$\boldsymbol{b}_k = \boldsymbol{\mu}_x^{(k)} - \boldsymbol{\Sigma}_{xy}^{(k)}\boldsymbol{\Sigma}_y^{(k)^{-1}}\boldsymbol{\mu}_y^{(k)}. \tag{2.147}$$

Desde esta perspectiva, podemos hacer una analogía directa entre estas dos técnicas y la técnica P-MEMLIN mencionada anteriormente.

2.2.4. Reconocimiento con incertidumbre

En las técnicas de compensación que se han presentado hasta ahora, se ha supuesto que el proceso de compensación del que se derivan las estimas de voz limpia no presenta ningún error, sin embargo en la práctica esto raramente es así. En efecto, existen numerosos aspectos que pueden influir en la calidad de las estimas como son la SNR de la observación, la aleatoriedad del ruido y la bondad de los modelos de voz usados en la estimación, entre otros. En base a esto, parece razonable pensar que el reconocedor de voz se beneficiaría del conocimiento de la incertidumbre residual del proceso de compensación, de forma que las estimaciones menos fiables tuviesen un peso menor en el proceso de descodificación. En este apartado estudiaremos una estrategia eficiente para llevar a cabo esto: el reconocimiento con incertidumbre.

Por reconocimiento con incertidumbre se ha denominado en la literatura a un conjunto de técnicas distintas que manejan términos como estimación, incertidumbre de la estimación y propagación de esa incertidumbre al reconocedor. Aquí intentaremos categorizar esas técnicas en tres grandes grupos: (i) esquemas basados en la propagación de la incertidumbre del proceso de realce, (ii) enfoques basados en el paradigma de datos perdidos y (iii) propagación de la probabilidad a posteriori de la voz ruidosa. Aunque se comentarán de forma breve aquí, los dos primeros enfoques serán analizados en mayor profundidad en las secciones 5.2 y 2.2.5, respectivamente.

En lugar de considerar las estimas de voz $\hat{\boldsymbol{x}}_t$ como deterministas, los esquemas basados en la propagación de la incertidumbre de la estima consideran que el proceso de realce devuelve una PDF de evidencia $\boldsymbol{x}_t \sim p(\boldsymbol{x}_t|\boldsymbol{y}_t)$. Se debe notar que en el caso determinista esta PDF degenerará a una delta de Dirac centrada en $\hat{\boldsymbol{x}}_t$, es decir, $p(\boldsymbol{x}_t|\boldsymbol{y}_t) = \delta_{\hat{\boldsymbol{x}}_t}(\boldsymbol{x}_t)$. Dos son las modificaciones que se han de realizar al reconocedor para adaptarlo a este esquema de reconocimiento con incertidumbre: estimación de la PDF de evidencia y modificación del cómputo de las probabilidades de observación del reconocedor para aceptar entradas inciertas. En relación al primer punto, varias han sido las estrategias que se han propuesto para estimar los parámetros de la PDF de evidencia (ver sección 5.2), aunque en la mayoría de ellas esta PDF se deriva a partir del propio proceso de realce, estimando, por ejemplo, la varianza de la estimación MMSE [69, 72, 122, 126, 252]. En cuanto al cómputo modificado de las probabilidades de observación, hay dos estrategias principales para llevar a cabo esta tarea: reconocimiento *soft-data* y algoritmo ponderado de Viterbi. Estas estrategias serán descritas más adelante en las secciones 5.2.1 y 5.2.2, respectivamente.

El paradigma de reconocimiento con datos perdidos que describiremos en la siguiente sección puede considerarse como un caso especial de reconocimiento con incertidumbre. El primer paso que llevan a cabo las técnicas que se encuadran dentro de

este paradigma, consiste en estimar la fiabilidad de las características observadas. Por fiabilidad aquí se entiende el que dichas características estén dominadas por la energía del ruido (características no fiables o perdidas) o por la energía de la voz (características fiables). A partir de esta clasificación a priori del espectro observado, se puede proceder de dos formas alternativas: bien estimando las características perdidas o bien teniendo en cuenta la distinción entre características fiables y perdidas durante el proceso de descodificación. La primera estrategia conduce a las técnicas de imputación (en realidad, una forma de compensación de características) que estudiaremos en la sección 2.2.5.4. En cuanto a la segunda estrategia, que podemos incluir más claramente dentro de las técnicas de reconocimiento con incertidumbre que estamos analizando aquí, será estudiada en la sección 2.2.5.3. De forma resumida, las técnicas que caen dentro de este segundo grupo modifican el cómputo de las probabilidades de observación del reconocedor para que se tenga en cuenta la fiabilidad de cada característica. Esta modificación puede ir desde ignorar completamente en dicho cómputo las características perdidas, o llevar a cabo una marginalización acotada de las mismas.

La última de las estrategias de reconocimiento con incertidumbre que estudiaremos en este apartado se basa en propagar al reconocedor la probabilidad a posteriori de la voz ruidosa $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$. Esto contrasta con el esquema de propagación de la incertidumbre de la estima que, como hemos visto, pasa el reconocedor la PDF de evidencia $p(\boldsymbol{x}_t|\boldsymbol{y}_t)$. El objetivo que se persigue al propagar $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$ es el de poder reconocer con la propia voz ruidosa. Para ello consideremos la probabilidad de observación $p(\boldsymbol{y}_t|s)$, donde s es un estado dado del modelo acústico. Esta probabilidad se puede expresar de la siguiente forma

$$\begin{aligned} p(\boldsymbol{y}_t|s) &= \iint p(\boldsymbol{x}_t, \boldsymbol{n}_t, \boldsymbol{y}_t|s) d\boldsymbol{x}_t d\boldsymbol{n}_t \\ &= \iint p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{n}_t) p(\boldsymbol{x}_t|s) p(\boldsymbol{n}_t) d\boldsymbol{x}_t d\boldsymbol{n}_t. \end{aligned} \tag{2.148}$$

Esta ecuación coincide con la obtenida en (2.34) para la estrategia de descomposición de modelos. Como puede observarse, en la integral intervienen tres términos: las probabilidades de observación de la voz limpia $p(\boldsymbol{x}_t|s)$, las probabilidades del ruido $p(\boldsymbol{n}_t)$ y el modelo de distorsión $p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{n}_t)$. Para simplificar la ecuación anterior, la integración sobre la variable oculta $\boldsymbol{n}_t$ se efectúa de manera independiente del modelo de voz,

$$p(\boldsymbol{y}_t|\boldsymbol{x}_t) = \int p(\boldsymbol{y}_t|\boldsymbol{x}_t, \boldsymbol{n}_t) p(\boldsymbol{n}_t) d\boldsymbol{n}_t. \tag{2.149}$$

Por tanto, la probabilidad de observación de la ecuación (2.148) queda expresada como,

$$p(\boldsymbol{y}_t|s) = \int p(\boldsymbol{y}_t|\boldsymbol{x}_t) p(\boldsymbol{x}_t|s) d\boldsymbol{x}_t. \tag{2.150}$$

El problema que se plantea ahora es la elección de una distribución $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$ que resulte en una solución analítica de la integral anterior y que, además, sea eficiente de calcular. En este sentido, varias son las propuestas que podemos encontrar en la literatura. Por ejemplo, la técnica SPLICE con incertidumbre [78] deriva $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$ mediante la aplicación de la regla de Bayes sobre la distribución $p(\boldsymbol{x}_t|\boldsymbol{y}_t)$ que esta técnica estima (ver sección 2.2.3.4). Dado que la propagación de esta PDF al reconocedor supondría multiplicar por M_y el número de componentes del modelo acústico, donde M_y es el número de gaussianas del GMM utilizado por SPLICE, únicamente se selecciona la gaussiana más probable k^* en cada instante de tiempo. Bajo esta simplificación, es posible demostrar que la probabilidad de la ecuación (2.150) equivale a [78, 175],

$$p(\boldsymbol{y}_t|s) = \sum_{m=1}^{M} P(m|s) \left|\boldsymbol{A}^{(k^*)}\right| \mathcal{N}\left(\boldsymbol{A}^{(k^*)}\boldsymbol{y}_t + \boldsymbol{b}^{(k^*)}; \boldsymbol{\mu}_s^{(m)}, \boldsymbol{\Sigma}_s^{(m)} + \boldsymbol{\Sigma}^{(k^*)}\right), \tag{2.151}$$

donde $P(m|s)$, $\boldsymbol{\mu}_s^{(m)}$ y $\boldsymbol{\Sigma}_s^{(m)}$ denotan los parámetros de la gaussiana m en el estado s del HMM, mientras que $\boldsymbol{A}^{(k^*)}$, $\boldsymbol{b}^{(k^*)}$ y $\boldsymbol{\Sigma}^{(k^*)}$ se derivan del proceso de compensación que SPLICE lleva a cabo.

Otra propuesta alternativa para el cómputo de $p(\boldsymbol{y}_t|\boldsymbol{x}_t)$ es la basada en la técnica FE-Joint [175, 177] estudiada anteriormente. En este caso, la PDF anterior se deriva del modelo conjunto $p(\boldsymbol{x}_t, \boldsymbol{y}_t)$ que esta técnica estima explícitamente. Aunque el origen de partida sea diferente al de SPLICE, ambas técnicas resultan en expresiones muy parecidas.

2.2.5. El paradigma de datos perdidos en el reconocimiento robusto de voz

El paradigma de datos perdidos (MD, *missing data*) es una aproximación al reconocimiento robusto en condiciones de ruido acústico propuesta por varios investigadores de la Universidad de Sheffield a mediados de los años 1990. En lugar de caracterizar explícitamente el ruido acústico como hacen otras técnicas (p.ej. VTS), las técnicas que se encuadran dentro de este paradigma explotan la redundancia inherente de la señal de voz para reconocerla incluso en presencia de distorsiones severas. La idea en la que se basa este paradigma es que la degradación producida por el ruido en la señal de voz no afecta a todas las regiones del espectro por igual, sino que hay regiones que apenas se ven afectadas (regiones fiables), mientras que otras se encuentran completamente enmascaradas por la energía del ruido (regiones perdidas o no fiables). Las regiones fiables cuentan con una SNR alta, por lo que pueden ser utilizadas directamente por el reconocedor. Por contra, las regiones perdidas requieren un procesamiento especial, ya

que su SNR suele ser baja (inferior a 0 dB) y producirían una merma en la precisión del reconocedor si éste trabajase con ellas directamente.

Partiendo de la observación anterior, el objetivo que se plantean las técnicas MD es el diseño de un sistema de reconocimiento del habla que sea robusto a las pérdidas provocadas por el ruido. Para alcanzar este objetivo, en la literatura se han planteado dos enfoques principalmente: (i) modificar el reconocedor para que ignore los valores perdidos durante el reconocimiento y (ii) imputar (estimar) estos valores antes de descodificar la voz. Ambos enfoques serán analizados en profundidad en la sección 2.2.5.2, pero antes de ello, dada la gran importancia del paradigma de datos perdidos en esta tesis, en el siguiente apartado presentaremos sus los fundamentos desde el punto de vista psicoacústico.

2.2.5.1. Fundamentos psicoacústicos

La falta de robustez ha limitado el uso y difusión del reconocimiento automático del habla fuera de ambientes donde la voz posee una SNR alta. Por contra, el oído humano demuestra ser más robusto ante estas distorsiones sin necesidad de un entrenamiento específico para cada condición acústica. La razón de esta mayor robustez en la tarea de reconocimiento realizada por seres humanos la podemos encontrar, entre otros factores, en dos propiedades claves de la propia señal de voz [55, 183]: su distribución poco densa en el dominio de la frecuencia y su redundancia espectro-temporal.

Cuando la señal de voz se expresa en el dominio de la frecuencia, se observa cómo la energía de la misma se tiende a agrupar en torno a ciertas de frecuencias de interés como son los formantes en el caso de las vocales, o las altas frecuencias para las fricativas. La distribución dispersa de la voz en el dominio de la frecuencia implica que cuando ésta se ve distorsionada por una fuente de ruido, éste tenderá a colmar los valles del espectro (regiones donde la voz tiene una energía baja), dejando prácticamente inalteradas aquellas regiones donde la voz posee una energía alta. Un ejemplo ilustrativo de este efecto se muestra en la figura 2.12. En ella se aprecia cómo la distorsión producida por el ruido en el espectro de la voz se traduce en un enmascaramiento de ciertas frecuencias dejando otras regiones intactas. La dominancia de una fuente sonora u otra se representa mediante una máscara binaria (también llamada máscara *missing data*) en donde las regiones correspondientes a la voz se identifican con un uno, mientras que las regiones en la que ésta ha sido enmascarada y, por tanto, el ruido es dominante, se marcan con un cero. Como se aprecia en la figura, aunque gran parte de la energía de la voz se ve enmascarada por el ruido a SNRs bajas, ciertas regiones del espectro de la voz no se ven afectadas por esta distorsión: esta observación clave constituye el fundamento de todas las técnicas MD.

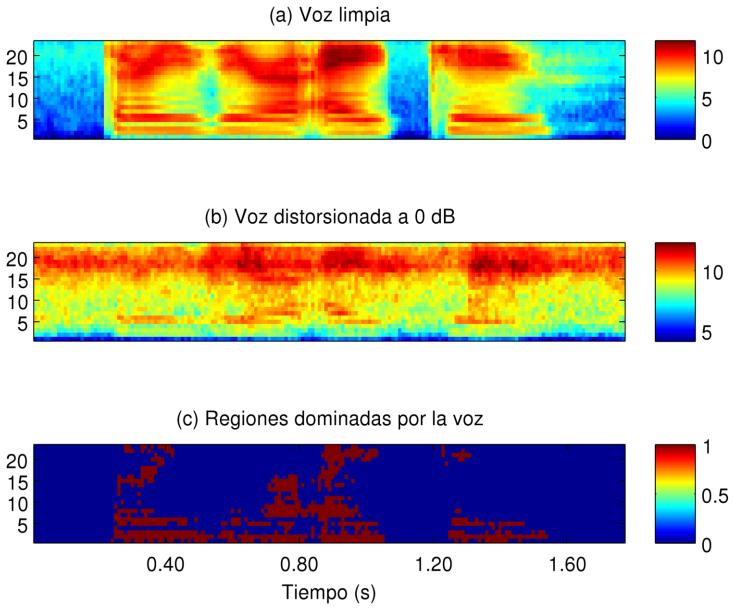

Figura 2.12: Ejemplo del enmascaramiento de dos fuentes sonoras simultáneas: (a) Espectrograma log-Mel de la frase *three zero eight two* (tres cero ocho dos) extraída de la base de datos Aurora2, (b) frase distorsionada con ruido aditivo de tipo *subway* a 0 dB y (c) máscara binaria con las regiones donde domina la energía de la voz.

El efecto de enmascaramiento que acabamos de describir informalmente para el caso de dos fuentes sonoras, voz y ruido, es bien conocido en el campo de la psicoacústica. Así, es conocido que en la percepción de un determinado sonido puede influir la presencia de otro sonido cuya energía sea mayor y, por tanto, el primero no sea audible [85]. El efecto de enmascaramiento puede ser tanto simultáneo, esto es, los dos sonidos se producen en el mismo instante de tiempo, o no simultáneo cuando ocurren en instantes diferentes. Suponiendo tonos puros, el efecto del enmascaramiento simultáneo se traduce en que ciertos sonidos son inaudibles por el oído humano (no se produce un disparo en las neuronas auditivas [193]) en presencia de sonidos con mayor energía en frecuencias adyacentes. Esta misma idea se extiende al caso del enmascaramiento no simultáneo en el que también influye la componente de tiempo: un tono puede estar enmascarado tanto por tonos "cercanos" en frecuencia como en tiempo. Este efecto de enmascaramiento ya ha sido explotado con anterioridad para enmascarar el error de cuantificación en los codificadores MP3 [217].

La segunda propiedad clave de la señal de voz que permite su inteligibilidad incluso en situaciones acústicamente adversas es su redundancia. En este sentido los experimentos de inteligibilidad reportados por Fletcher en [93] prueban que la inteligibilidad de

la voz no se ve mermada gravemente en presencia de distorsiones. Así, Fletcher realiza experimentos en los que se suprimen bandas de frecuencias exclusivas del espectro (frecuencias superiores e inferiores a 1800 Hz), concluyendo que ninguna frecuencia es clave para el reconocimiento humano. Otros experimentos de inteligibilidad también avalan esta conclusión. En [15, 273, 274], por ejemplo, se presentan una serie de experimentos subjetivos en los que se filtra la voz pasa-banda en bandas mutuamente exclusivas y se mide la inteligibilidad de la voz filtrada. Incluso con bandas muy estrechas (1/3 de octava) la voz filtrada posee una inteligibilidad alta.

Expuestas las dos propiedades anteriores de la voz, se plantean dos preguntas sobre el reconocimiento humano en presencia de ruido: ¿cómo es posible distinguir las regiones del espectro dominadas por la energía de la voz de las dominadas por el ruido? Y, una vez identificadas ambos tipos de regiones, ¿cómo se lleva a cabo el reconocimiento con espectros parcialmente observados? Estas dos cuestiones nos llevan al estudio del *análisis de la escena auditiva* (ASA, *Auditory Scene Analysis*) [41, 42], a saber, la forma en la que los seres humanos perciben y estructuran una escena acústica compuesta de varias fuentes sonoras. Cuando varias fuentes sonoras se superponen en el tiempo, los seres humanos somos capaces de segregar las componentes espectrales que se corresponden a cada una de ellas y centrar nuestra atención en la fuente de interés (voz por ejemplo). En este proceso de segregación intervienen tanto propiedades de bajo nivel de cada fuente como, por ejemplo, continuidad temporal, *onset/offset* común compartido por varias frecuencias, frecuencia fundamental, dirección desde la que llega el sonido de la fuente, etc., como información de alto nivel que indica cómo de parecido es cierto atributo a un patrón conocido a priori. En resumidas cuentas, la segregación es un proceso complejo guiado tanto por propiedades físicas de las señales (enfoque abajo-arriba) como por modelos obtenidos a través de la experiencia (enfoque arriba-abajo).

Para responder a la segunda pregunta formulada anteriormente, nos apoyaremos en los experimentos presentados por Cooke en [55] en los que compara el rendimiento de un reconocedor de voz *missing data* basado en marginalización (los detalles de este reconocedor se estudiarán más adelante) con el de un humano en una tarea de identificación de consonantes con ruido aditivo. Supuesto que se le proporciona al reconocedor una segmentación a priori entre voz/ruido, el resultado de éste es comparable al de un ser humano. Usando análisis de regresión entre los resultados alcanzados por el humano y la parte visible del espectro presentada al reconocedor automático, se hipotetiza que, en el ser humano, la tarea de descodificación de la voz en presencia de ruido se lleva a cabo comparando los espectros de voz parcialmente observados con patrones de voz conocidos previamente.

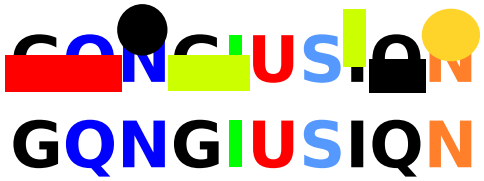

Figura 2.13: Analogía visual del proceso de análisis de la escena y la capacidad humana para identificar patrones usando observaciones parciales.

Como nota aclaratoria debemos de indicar que la capacidad para estructurar e interpretar una escena compleja en la que ciertos objetos se enmascaran no es exclusiva del sistema auditivo, sino que también está presente en la visión [13]. Una analogía visual de la capacidad humana para interpretar objetos parcialmente ocultos se muestra en la figura 2.13. La parte superior de la figura contiene una escena visual compleja compuesta de un texto junto con algunas figuras geométricas que ocultan ciertos trazos de los caracteres del texto. A partir de esa escena visual, el ser humano es capaz de segregar la escena e identificar qué partes pertenecen al objeto de interés (texto) y cuales no son de interés. Para ello se emplean propiedades de bajo nivel de la imagen como el color y la continuidad espacial de las formas, así como patrones de alto nivel como la forma de las letras, modelos lingüísticos de un lenguaje dado (en este caso el español), reglas ortográficas, etc. Como resultado de este proceso, el cerebro interpreta que la gráfica de arriba se refiere a la palabra "CONCLUSION", aunque otra solución factible desde el punto de vista gráfico, pero no gramatical, sería la pseudopalabra "GQNGIUSIQN" que aparece en la parte inferior de la figura.

Inspirados por todas estas ideas y observaciones, diversos investigadores de la Universidad de Sheffield (Reino Unido) propusieron durante la segunda mitad de la década de los 90 diversos modelos computacionales del análisis de la escena auditiva (CASA, *Computational* ASA), que dieron lugar a las técnicas *missing data* [56, 57, 59, 129, 204, 270]. Como en el ASA, las técnicas MD constan de dos fases bien diferenciadas: (i) segregación de la escena auditiva y (ii) descodificación de la voz con datos incompletos. En las técnicas MD, el proceso de segregación se traduce en la estimación de una máscara (binaria o continua) que identifica qué elementos tiempo-frecuencia del espectrograma se asocian a la voz y cuáles son considerados como ruido. Después del proceso de segregación, un reconocedor modificado procede a descodificar la voz usando la información proveniente de los elementos identificados como voz al mismo tiempo que impone ciertas restricciones en los identificados como ruido.

Las técnicas MD han demostrado su potencial en distintas tareas de reconocimiento automático de voz, probando que es posible obtener una tasa de reconocimiento alta

incluso si gran parte del espectro de la voz se encuentra enmascarado [59]. Para lograr este rendimiento, estas técnicas precisan tanto de una segregación precisa como de un reconocedor capaz de trabajar con datos perdidos. En el siguiente apartado revisaremos las distintas estrategias propuestas en la literatura para el reconocimiento de la voz en presencia de datos perdidos. El estudio de las técnicas propuestas para el cálculo de la máscara de segregación se hará, por otra parte, en la sección 2.2.5.5.

2.2.5.2. Aplicación al reconocimiento robusto de voz

Consideremos las variables $\boldsymbol{x}$, $\boldsymbol{y}$ y $\boldsymbol{n}$ relativas a las características log-Mel de la voz limpia, voz ruidosa y el ruido aditivo, respectivamente. En la sección 4.1 veremos que la idea de enmascaramiento entre voz y ruido en la que se basan las técnicas MD puede derivarse del modelo de distorsión expuesto en la sección 2.1. Esta idea se plasmará en un modelo de distorsión alternativo que relaciona analíticamente las variables anteriores. Dejando para la sección 4.1 el análisis detallado de esta derivación, a continuación presentamos la expresión final para dicho modelo:

$$\boldsymbol{y} \approx \max(\boldsymbol{x}, \boldsymbol{n}), \tag{2.152}$$

donde la operación máx se aplica elemento a elemento.

De acuerdo a este modelo, al cual denominaremos como *modelo de enmascaramiento de la voz* [90, 125] o, de forma alternativa, aproximación *log-max* [207, 233, 267], la observación $\boldsymbol{y}$ puede dividirse en dos subvectores $\boldsymbol{y} = (\boldsymbol{y}_r, \boldsymbol{y}_u)$. El vector $\boldsymbol{y}_r$ contiene los elementos fiables de la observación, es decir, aquellos donde la energía de la voz domina a la del ruido, por lo que $\boldsymbol{y}_r = \boldsymbol{x}_r$ y $\boldsymbol{n}_r \leq \boldsymbol{x}_r$, siendo $\boldsymbol{x}_r$ y $\boldsymbol{n}_r$ los elementos de $\boldsymbol{x}$ y $\boldsymbol{n}$ asociados al vector $\boldsymbol{y}_r$. Por otro lado, $\boldsymbol{y}_u$ contiene aquellos elementos donde domina la energía del ruido, de esta forma se cumple que $\boldsymbol{y}_u = \boldsymbol{n}_u$ y $\boldsymbol{x}_u < \boldsymbol{n}_u$. En este caso la única información que se dispone sobre las características enmascaradas de voz es el rango en el que se encuentran: este rango está acotado inferiormente por $-\infty$ (o cualquier otro umbral estimado a partir de los datos de entrenamiento) y superiormente por los propios valores de voz contaminada observados $\boldsymbol{y}_u$.

Puesto que $\boldsymbol{y}_r$ constituye una buena aproximación a las características correspondientes de voz, este conjunto de valores puede ser directamente usado por el reconocedor. Por contra, el reconocedor debe ser modificado para poder trabajar con el conjunto de datos perdidos $\boldsymbol{y}_u$. Dos son los principales enfoques que se han planteado para este fin: (i) reconocimiento de espectros incompletos de voz y (ii) imputación (estimación) de los valores perdidos previa a la etapa de reconocimiento. La primera categoría se centra en la modificación de la etapa de reconocimiento para aceptar espectros de voz ruidosos (incompletos desde el punto de vista del paradigma de datos perdidos). Desde

este punto de vista, podemos hacer una analogía entre este grupo de técnicas y las de adaptación expuestas en la sección 2.2.2, ya que ambas trabajan sobre los modelos del reconocedor. Por otra parte, las técnicas de imputación, al igual que las de compensación estudiadas en la sección 2.2.3, procesan las características de voz antes de ser usadas por el motor de reconocimiento.

Aunque no aparezca recogido de forma explícita en la ecuación (2.152), el paradigma de datos perdidos también se ha empleado con éxito para combatir el ruido convolutivo [210, 265]. En general, la estrategia para abordar este ruido consiste en estimar (en el dominio espectral logarítmico) un vector de desplazamiento global que compense el desplazamiento sufrido por las características de la voz como consecuencia del ruido convolutivo. Este vector se suele estimar a partir de las regiones del espectro identificadas como voz (regiones fiables), descartando las regiones identificadas como ruido. Con ello se pretende evitar que las regiones donde el ruido domina alteren la estimación del factor de compensación y se produzca una sobre-sustracción de ruido.

En los siguientes apartados se describen los dos enfoques considerados para el reconocimiento de voz con datos perdidos.

2.2.5.3. Reconocimiento de espectros incompletos

Las técnicas que se encuadran en esta categoría pretenden la modificación del reconocedor de voz para que acepte señales de voz ruidosa. En vez de modificar los parámetros del modelo acústico, tal y como se hace en la adaptación de modelos, las técnicas que estudiaremos en este apartado modifican el cómputo de las probabilidades de observación en el reconocedor, consiguiendo con ello que éste tenga constancia de cuáles son los elementos fiables y no fiables del espectro.

En este apartado estudiaremos dos enfoques alternativos para reconocer con espectros de voz incompletos: el reconocimiento mediante marginalización [57, 59] y la aproximación basada en la descodificación de fragmentos de voz (SFD, Speech Fragment Decoding) [25, 26, 27, 30, 32]. Ambos enfoques son descritos en los siguientes puntos.

Marginalización de los datos perdidos. Consideremos un reconocedor estándar basado en HMMs en el que cada estado se modela mediante una mezcla de gaussianas. De forma estándar, la probabilidad de observación de los vectores de voz limpia bajo el estado s-ésimo del modelo se calcula de la siguiente forma:

$$p(\boldsymbol{x}|s) = \sum_{k=1}^{M} P(k|s)p(\boldsymbol{x}|k,s), \tag{2.153}$$

donde $p(\boldsymbol{x}|k,s)$ representa una función de densidad de probabilidad gaussiana en varias variables.

Suponiendo que el modelo acústico del reconocedor se ha entrenado con voz limpia, para el cálculo de la probabilidad de observación anterior se requiere que todas las componentes de los vectores de voz observados sean fiables, o dicho con otras palabras, en presencia de ruido el uso de la ecuación anterior conlleva una discrepancia. Para abordar el problema de datos perdidos, Cooke propone en [57] descodificar la voz en base a los datos fiables observados. Matemáticamente esto implica evaluar la siguiente probabilidad marginal de los datos fiables,

$$\begin{aligned} p(\boldsymbol{y}|s) \equiv p(\boldsymbol{y}_r, \boldsymbol{y}_u|s) &= \int p(\boldsymbol{y}_r, \boldsymbol{x}_u|s) d\boldsymbol{x}_u \\ &\equiv \sum_{k=1}^{M} \int p(\boldsymbol{y}_r, \boldsymbol{x}_u, k|s) d\boldsymbol{x}_u \\ &= \sum_{k=1}^{M} P(k|s) p(\boldsymbol{y}_r|k,s) \int p(\boldsymbol{x}_u|\boldsymbol{y}_r, k, s) d\boldsymbol{x}_u. \end{aligned} \tag{2.154}$$

En la ecuación anterior vemos que intervienen varias PDFs que involucran los datos fiables, no fiables y las gaussianas del GMM: $p(\boldsymbol{y}_r|k,s)$ es la PDF marginal de los datos fiables para la componente k-ésima del GMM, mientras que $p(\boldsymbol{x}_u|\boldsymbol{y}_r,k,s)$ es la probabilidad condicional de los datos perdidos dados los fiables. Puesto que $p(\boldsymbol{x}|k,s)$ es gaussiana, se tiene que estas PDFs también son gaussianas. Omitiendo el índice s del estado por claridad, los parámetros de las distribuciones marginales de los datos fiables $p(\boldsymbol{x}_r|k) = \mathcal{N}(\boldsymbol{x}_r; \boldsymbol{\mu}_r^{(k)}, \boldsymbol{\Sigma}_r^{(k)})$ y no fiables $p(\boldsymbol{x}_u|k) = \mathcal{N}(\boldsymbol{x}_u; \boldsymbol{\mu}_u^{(k)}, \boldsymbol{\Sigma}_u^{(k)})$ se pueden calcular reordenando los parámetros de $p(\boldsymbol{x}|k)$ según la información recogida en la máscara de segregación,

$$\boldsymbol{\mu}^{(k)} = \begin{pmatrix} \boldsymbol{\mu}_r^{(k)} \\ \boldsymbol{\mu}_u^{(k)} \end{pmatrix}, \tag{2.155}$$

$$\boldsymbol{\Sigma}^{(k)} = \begin{pmatrix} \boldsymbol{\Sigma}_{rr}^{(k)} & \boldsymbol{\Sigma}_{ru}^{(k)} \\ \boldsymbol{\Sigma}_{ur}^{(k)} & \boldsymbol{\Sigma}_{uu}^{(k)} \end{pmatrix}. \tag{2.156}$$

Por otra parte, la media y covarianza de la PDF condicional $p(\boldsymbol{x}_u|\boldsymbol{y}_r,k) = \mathcal{N}(\boldsymbol{x}_u; \boldsymbol{\mu}_{u|r}^{(k)}, \boldsymbol{\Sigma}_{u|r}^{(k)})$ se calculan de la siguiente forma:

$$\boldsymbol{\mu}_{u|r}^{(k)} = \boldsymbol{\mu}_u^{(k)} + \boldsymbol{\Sigma}_{ur}^{(k)} \boldsymbol{\Sigma}_{rr}^{(k)^{-1}} (\boldsymbol{y}_r - \boldsymbol{\mu}_r^{(k)}), \tag{2.157}$$

$$\boldsymbol{\Sigma}_{u|r}^{(k)} = \boldsymbol{\Sigma}_{uu}^{(k)} - \boldsymbol{\Sigma}_{ur}^{(k)} \boldsymbol{\Sigma}_{rr}^{(k)^{-1}} \boldsymbol{\Sigma}_{ru}^{(k)}, \tag{2.158}$$

coincidiendo estos parámetros con los de la distribución marginal $p(\boldsymbol{x}_u|k)$ si se usan matrices de covarianza diagonales en el GMM.

Volviendo al cálculo de la ecuación (2.154), notaremos como $\boldsymbol{s}_r$ y $\boldsymbol{s}_u$ a los conjuntos con los índices de las componentes fiables y no fiables, respectivamente, del vector $\boldsymbol{y}$. Bajo la suposición de diagonalidad de las matrices de covarianza, esta ecuación se puede reescribir como

$$\begin{aligned} p(\boldsymbol{y}|s) &= \sum_{k=1}^{M} P(k|s) \prod_{i \in \boldsymbol{s}_r} p(y_{r,i}|k,s) \prod_{j \in \boldsymbol{s}_u} \int p(x_j|k,s) dx_j \\ &= \sum_{k=1}^{M} P(k|s) \prod_{i \in \boldsymbol{s}_r} p(y_{r,i}|k,s), \end{aligned} \tag{2.159}$$

es decir, el reconocimiento se efectúa teniendo en cuenta únicamente las componentes fiables del espectrograma. En el caso límite en el que no se haya identificado ninguna componente fiable, las probabilidades de observación para cada estado serán equiprobables, viniendo guiado el proceso de reconocimiento únicamente por las probabilidades de transición entre estados. Desde el punto de vista bayesiano, ésta es la estrategia óptima de reconocimiento en caso de que el ruido enmascare completamente a la señal de voz.

Debemos de remarcar que, para proceder a calcular la probabilidad de observación definida en (2.159), las características deben expresarse en el dominio espectral (lineal o logarítmico): la segregación en componentes perdidas y fiables llevada a cabo de forma natural en este dominio se difumina en el cepstrum (o cualquier otro dominio que suponga una transformación del espectro de la voz). Visto desde otra perspectiva, un ruido localizado en cierta banda de frecuencias del espectro afectará a todas las componentes cepstrales tras aplicar la DCT. El corolario de este razonamiento es que, para aplicar marginalización, los modelos acústicos del reconocedor se entrenan con características espectrales en lugar del cepstrum empleado en la mayoría de reconocedores actuales.

Además, el uso de características espectrales en el reconocedor lleva asociados otros problemas. En primer lugar, es bien sabido que el uso de estas características supone un deterioro del rendimiento del reconocedor en comparación con los parámetros MFCC [64]. Dos propiedades fundamentales de la DCT justifican este mayor rendimiento del cepstrum: (i) su capacidad para compactar la información relevante en pocos parámetros y (ii) el eliminar la correlación entre los elementos.

La primera propiedad de la DCT tiene como consecuencia que, dada la mayor dimensión de los parámetros espectrales que la de los cepstrales, para entrenar de forma robusta los parámetros del modelo acústico se requerirá un mayor volumen de datos en el caso de los parámetros espectrales.

La segunda propiedad de la DCT implica que, al contrario que los MFCCs, los parámetros log-Mel exhiben una gran correlación entre las distintas frecuencias. La

solución más rápida a este problema sería utilizar matrices de covarianza completas en el reconocedor, de tal forma que la correlación entre las característica pueda ser modelada de forma efectiva. Esta solución, no obstante, no está exenta de problemas. Por un lado, tenemos que el uso de matrices de covarianza no diagonales incrementa excesivamente el tiempo de cómputo, el cual puede llegar a ser prohibitivo para ciertos sistemas (p.ej. sistemas de reconocimiento de gran vocabulario). Por otro lado, y quizás más acuciante, es la no existencia de una expresión analítica para el cálculo de la integral de la ecuación (2.154) para matrices de covarianza no diagonales. Además nos encontramos que al usar matrices no diagonales se requerirá un mayor volumen de datos de entrenamiento. Por estas razones, el uso de reconocedores que emplean matrices de covarianza diagonales se ha impuesto en el enfoque basado en marginalización. No obstante, con el fin de representar de forma precisa las correlaciones entre componentes espectrales, estos reconocedores emplean un número de gaussianas por estado mayor del que es habitual en los reconocedores basados en coeficientes MFCC [1].

Como se puede apreciar, en la definición de la probabilidad de observación en presencia de datos perdidos de la ecuación (2.154) se ha descartado cualquier información proveniente de los datos considerados como no fiables. No obstante, a pesar de estar distorsionados por el ruido, estos datos nos proporcionan cierta información relevante sobre la voz enmascarada: la aproximación *log-max* de la ecuación (2.152) nos indica que la energía de la voz se va a encontrar en el intervalo $(-\infty, y]$ para cierta componente no fiable y observada. En consecuencia, podemos aplicar estos límites a la integral de la ecuación (2.154) dando lugar a lo que se conoce como marginalización acotada (*bounded marginalization*),

$$p(\boldsymbol{y}|s) = \int_{-\infty}^{\boldsymbol{y}_u} p(\boldsymbol{y}_r, \boldsymbol{x}_u|s) d\boldsymbol{x}_u = \sum_{k=1}^{M} P(k|s) p(\boldsymbol{y}_r, \boldsymbol{y}_u|k, s), \tag{2.160}$$

donde (asumiendo de nuevo matrices de covarianza diagonales),

$$p(\boldsymbol{y}_r, \boldsymbol{y}_u|k, s) = \prod_{i \in \boldsymbol{s}_r} p(y_{r,i}|k, s) \prod_{j \in \boldsymbol{s}_u} \int_{-\infty}^{y_{u,j}} p(x_j|k, s) dx_j. \tag{2.161}$$

El uso de límites en la integral de marginalización resulta en un mejor rendimiento del reconocedor frente a no usarlos [57]. En particular, el producto sobre los elementos no fiables en la ecuación (2.161) refleja lo que se denomina en la literatura como contra-evidencia a una componente del GMM. Es esta ecuación se puede observar que las

[1] El uso de un elevado número de componentes por estado no supone más problema que el de requerir un mayor volumen de datos para su entrenamiento. Desde el punto de vista de la capacidad de modelado, debemos de notar que cualquier función de densidad de probabilidad se puede aproximar mediante un GMM con matrices de covarianza diagonales [230].

integrales en este productorio se aproximarán a uno en las componentes cuyas medias tengan una energía baja y varianzas pequeñas. Dicho con otras palabras, si un elemento del vector de entrada es identificado como no fiable, el modelo de enmascaramiento forzará a que el valor desconocido de la voz sea mucho menor que el valor observado y, por esta razón, las componentes del GMM que modelen silencios tendrán más peso en el cómputo de la probabilidad de observación.

Existen otros indicios psicoacústicos que refuerzan el uso de los límites en la integral de marginalización. Por dar un ejemplo, en [55] se describen ciertos experimentos en donde se compara la inteligibilidad de la voz distorsionada por ruido acústico en dos contexto diferentes. En el primero de ellos, se mide simplemente la inteligibilidad de la voz distorsionada. Para el segundo caso, se suprimen (filtran) del espectro aquellas componentes distorsionadas por el ruido. Al medir la inteligibilidad en ambos casos se comprueba que la presencia del ruido, es decir, la no supresión de éste del espectro, ayuda al oído humano a entender mejor el mensaje. En [63] se presentan unos experimentos similares, pero esta vez en lugar de ruido acústico la voz se distorsiona filtrándola pasa-banda. De nuevo la adición artificial de ruido en las bandas filtradas ayuda a mejorar la inteligibilidad de la voz.

En las ecuaciones anteriores se ha supuesto que el proceso de segregación que identifica la voz del resto de fuentes es discreto, es decir, cada elemento del espectrograma se etiqueta inequívocamente como voz o ruido. Aunque teóricamente ésta es la forma correcta de proceder, en situaciones realistas donde el proceso de segregación involucra la estimación de la máscara *missing data* (p.ej. usando una estimación de la densidad de potencia espectral del ruido), se ha comprobado que el uso de máscaras continuas que toman valores en el intervalo $m_i \in [0, 1]$ proporciona mejores resultados [29, 31]. Valores cercanos a 1 en esta máscara indican una alta probabilidad de que el elemento en cuestión sea voz, por otra parte valores en torno a 0 indican presencia de ruido. Usando la información de esta máscara podemos extender el cómputo de la probabilidad dada por la ecuación (2.160) de la siguiente forma

$$p(\boldsymbol{y}|s) = \sum_{k=1}^{M} P(k|s) \prod_{i=1}^{D} \left(m_i p(y_i|k,s) + (1 - m_i) \int_{-\infty}^{y_i} p(x|k,s)dx \right). \tag{2.162}$$

Como se observa, la probabilidad para cada elemento del vector se calcula como una combinación lineal de la probabilidad para el caso de que el elemento sea voz y de la probabilidad marginal para el caso de que dicho elemento sea ruido.

Barker propone en [31] una ecuación alternativa a (2.162) en la que se asume que la probabilidad a priori del ruido es uniforme en el intervalo $[\theta_i, y_i]$ para cierto valor mínimo θ_i observado durante la fase de entrenamiento. Bajo estas consideraciones, la

probabilidad de observación para cada estado viene dada por

$$p(\boldsymbol{y}|s) = \sum_{k=1}^{M} P(k|s) \prod_{i=1}^{D} \left(m_i p(y_i|k,s) + (1-m_i) \frac{1}{y_i - \theta_i} \int_{\theta_i}^{y_i} p(x|k,s)dx \right). \quad (2.163)$$

Los resultados de reconocimiento obtenidos por las técnicas de marginalización en distintas bases de datos y publicados a lo largo de la última década, han demostrado el potencial de esta técnica para combatir de forma efectiva la degradación debida al ruido. De hecho, desde el punto de vista de la teoría bayesiana, el reconocimiento mediante marginalización se corresponde con el algoritmo óptimo de descodificación en presencia de datos perdidos, siempre y cuando la máscara de segregación no contenga errores [224]. A pesar de sus bondades, dos son las desventajas que se le pueden achacar a esta técnica. En primer lugar, y como comentábamos anteriormente, esta técnica requiere que los modelos acústicos se entrenen con características espectrales (p.ej. log-Mel). Esta es quizás una de las causas por las que, al no emplear un conjunto óptimo de características para el reconocimiento, el rendimiento de las técnicas de marginalización no sea competitivo en sistemas de gran vocabulario [65]. En segundo lugar, las técnicas de marginalización requieren para operar una máscara que diferencie la voz del ruido. En la mayoría de las situaciones la estimación de esta máscara es un problema difícil, más si consideramos que lo que de cara al reconocedor es ruido puede tratarse de otra persona hablando.

En la literatura se han propuesto distintas soluciones para abordar los dos problemas anteriores. Para evitar tener que operar en el reconocedor con características espectrales, las técnicas de imputación descritas en la sección 2.2.5.4 estiman los elementos perdidos del espectro para, posteriormente, proceder a calcular los MFCCs utilizados por un reconocedor tradicional. En cuanto al segundo problema, la técnica SFD [27] descrita en el siguiente apartado puede considerarse como una extensión de las técnicas de marginalización, ya que no requiere de una segregación a priori entre voz y ruido, sino que opera sobre un conjunto de fragmentos sonoros.

Descodificación de fragmentos de voz. La técnica de descodificación de fragmentos de voz (SFD) [25, 26, 27, 30, 32] supone ahondar más en el problema del reconocimiento automático del habla en presencia de datos perdidos. En el apartado anterior hemos estudiado cómo el enfoque basado en marginalización requería de una segregación a priori que diferenciase voz y ruido y, en base a esa segregación, proponía un cómputo modificado de las probabilidades de observación de los estados del HMM. Comentábamos también la dificultad en el diseño de un algoritmo de segregación que sea robusto a una gran variedad de tipos de ruido y SNRs, siendo éste uno de los principales problemas de las técnicas de marginalización. En este apartado estudiaremos

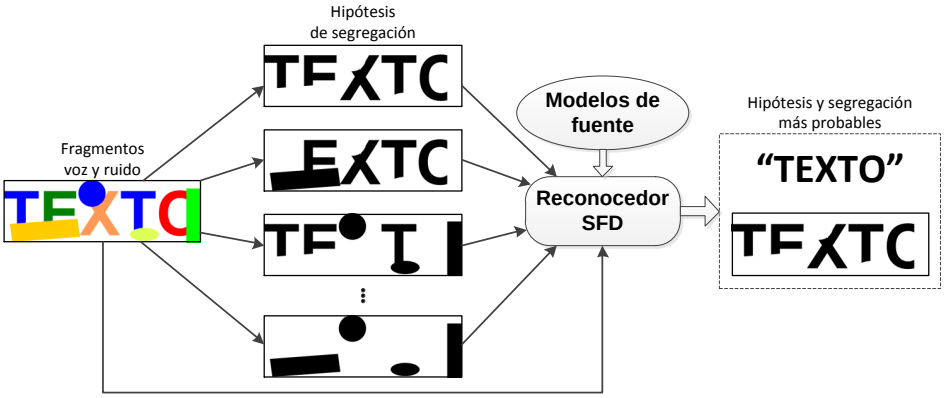

Figura 2.14: Esquema gráfico de un reconocedor SFD.

una aproximación alternativa para, de forma conjunta, obtener la máscara de segregación y la hipótesis del reconocedor (texto) que maximizan la probabilidad de los datos observados: la técnica SFD.

La técnica SFD puede considerarse como una implementación eficiente del CASA, esto es, del análisis computacional de la escena auditiva [42, 271]. Al contrario que el reconocimiento basado en marginalización de los datos observados, el reconocedor SFD relaja la necesidad de disponer de una segmentación a priori del espectro de entrada, necesitando, por contra, la identificación de una serie de fragmentos sonoros en dicho espectro. Dichos fragmentos son regiones conexas en el espacio frecuencia-tiempo del espectro cuya energía se encuentra dominada por una única fuente sonora como por ejemplo la voz del locutor, la voz otra persona o, posiblemente, un ruido de fondo como el de un ventilador. Identificados los fragmentos sonoros que pueden ser de voz o de ruido, un algoritmo de descodificación modificado al que llamaremos reconocedor SFD se encarga de seleccionar aquellos que mejor explican la señal observada dados los modelos de la voz disponibles. De forma suscinta, tal y como se muestra en la figura 2.14, la técnica SFD consta de dos componentes bien diferenciados: (i) el generador de fragmentos y (ii) el algoritmo de descodificación de fragmentos o reconocedor SFD.

El generador de fragmentos se encarga de segmentar la señal de entrada en regiones continuas cuya energía viene dominada por una única fuente. Para ello se explotan primitivas de bajo nivel del ASA para agrupar en fragmentos los elementos del espectrograma. Primitivas comunes que se suelen emplear para agrupar distintas frecuencias son su relación armónica (*pitch* común), *onset*/*offset* común y la localización espacial de la fuente que las genera [133, 186]. La analogía visual que se emplea en el ejemplo de la figura 2.14 sería segmentar una imagen usando primitivas como la forma geométrica de los objetos que la componen, el color, la continuidad de las formas o el tipo de tex-

tura. Como resultado del proceso de generación de fragmentos se obtiene una máscara discreta, del mismo tamaño que el del espectrograma, que recoge la identidad de los distintos fragmentos obtenidos. De forma opcional el generador de fragmentos también puede producir una máscara adicional con la confianza de la segmentación obtenida [32].

A partir de los fragmentos identificados, el reconocedor SFD genera todas las posibles hipótesis de segregación posibles. Estas hipótesis se evalúan una por una por el reconocedor SFD de forma eficiente usando para ello los modelos acústicos y de lenguaje entrenados con voz limpia. Aquí aparece de nuevo el problema de descodificación de la voz usando espectrogramas incompletos. Es por ello que, en el núcleo del reconocedor SFD, nos encontramos un reconocedor MD basado en marginalización, el cual evalúa la probabilidad de las distintas hipótesis de segregación generadas anteriormente. Finalmente, tal y como se refleja en la figura 2.14, el reconocedor SFD devuelve la secuencia de texto más probable y, opcionalmente, la hipótesis de segregación que ha generado.

La flexibilidad que proporciona la técnica SFD al identificar distintas regiones del espectro dominadas por la energía de una fuente sonora y, posteriormente, evaluarlas en función de unos modelos de fuente dados, ha resultado en la aplicación de esta técnica a distintos problemas del procesado de voz. En [32], por ejemplo, se utiliza la técnica SFD para el reconocimiento simultáneo del habla y del locutor en un entorno en el que existen varias personas hablando al mismo tiempo. Para ello se utiliza un estimador *multi-pitch*, esto es, un algoritmo que puede seguir la pista a varias frecuencias fundamentales al mismo tiempo, para generar los fragmentos asociados a cada una de las frecuencias fundamentales. Es tarea del reconocedor SFD seleccionar aquellos fragmentos provenientes del locutor objetivo. Otro problema relacionado es el de la separación de las voces de múltiples locutores grabadas usando un único micrófono, tarea en la que la técnica SFD ha demostrado también obtener resultados bastante prometedores [58].

2.2.5.4. Imputación de las características perdidas

Un enfoque alternativo para abordar el problema de datos perdidos en el RAH consiste en, a semejanza de las técnicas de compensación de características estudiadas en la sección 2.2.3, estimar las características perdidas en una fase previa a la etapa de reconocimiento propiamente dicha. Para llevar a buen puerto este cometido, las técnicas de imputación explotan la redundancia de la voz para estimar de forma precisa la energía de la voz en aquellos elementos identificados como perdidos. Para modelar esta redundancia, se suelen emplear modelos estadísticos de la voz en los que se recoge de

forma explícita las correlaciones entre las características perdidas y las fiables. Asimismo, la mayoría de las técnicas de imputación también explotan, de una u otra forma, las restricciones impuestas por el modelo de enmascaramiento de la voz de la ecuación (2.152), es decir, las características perdidas $\boldsymbol{x}_u$ se sitúan en el intervalo $(-\infty, \boldsymbol{y}_u]$.

Si comparamos las técnicas de imputación con el reconocimiento basado en marginalización, encontramos con las mismas ventajas e inconvenientes que comentábamos cuando comparábamos las técnicas de compensación con las de adaptación. En pocas palabras, la estimación de las componentes perdidas del espectro previa a la etapa de descodificación permite, por un lado, mantener el motor de reconocimiento intacto sin que éste tenga que ser modificado para aceptar datos perdidos y, por otro, trabajar con un conjunto óptimo de características para el reconocimiento, no necesariamente con características espectrales. Por contra, las técnicas de imputación suelen acarrear varias desventajas inherentes a su naturaleza. En primer lugar, las técnicas de imputación suelen devolver estimaciones puntuales de los datos perdidos, mientras que en marginalización se trabaja con la distribución a posteriori al completo de los datos perdidos, lo cual, desde el punto de vista bayesiano, es la óptimo. Para paliar este problema, diversos trabajos [113, 126, 248] proponen extensiones a las técnicas de imputación usando el paradigma de descodificación con incertidumbre (ver sección 2.2.4) con resultados muy prometedores. Otra limitación de las técnicas de imputación es que éstas suelen emplear, por razones de eficiencia, modelos de fuente de poca resolución (GMMs generalmente), mientras que las técnicas de marginalización emplean el propio modelo acústico del reconocedor.

Después de esta introducción, en los siguientes puntos se presentarán algunas de las técnicas de imputación propuestas en la literatura.

CBR. Uno de los primeros intentos para resolver el problema de estimación de los datos perdidos del espectro fue propuesto por Raj en [224]. En este artículo se detallan dos técnicas para la reconstrucción (estimación) de los elementos perdidos del espectrograma. Partiendo del espectrograma de una señal de voz distorsionada por ruido aditivo y de una máscara binaria que identifica los elementos fiables y los perdidos, estas técnicas devuelven a su salida un espectrograma completo que coincide con el espectrograma original en los elementos fiables, pero donde los elementos no fiables se reemplaza por estimaciones oportunas. Posteriormente este espectrograma puede ser utilizado para calcular los parámetros MFCCs empleados por un reconocedor común.

La primera técnica propuesta por Raj en [224] se denomina reconstrucción basada en la correlación. En esta técnica se asume que la señal de voz limpia es un proceso aleatorio estacionario de sentido amplio. El anterior calificativo implica que, para para

cada componente en frecuencia k del espectro, la media de dicha componente es independiente del tiempo $\mu(t,k) = \mu(k)$. Asimismo, la suposición de estacionario de sentido amplio también implica que la covarianza entre los elementos del espectrograma que ocupan las posiciones (k_1, t_1) y (k_2, t_2), siendo k el índice de frecuencia y t el de tiempo, es independiente de sus posiciones absolutas en tiempo, dependiendo la covarianza únicamente de la distancia temporal relativa que los separa $\tau = t_2 - t_1$. Teniendo en cuenta estas dos suposiciones, esta técnica estima el valor de los elementos perdidos del espectro usando para ello las correlaciones conocidas con los elementos fiables que se encuentran en la vecindad. Aunque esta técnica consigue importantes mejoras respecto a reconocer con voz ruidosa, no se muestra lo suficientemente competitiva respecto a una segunda técnica basada en GMMs también propuesta por Raj en [224]. La razón principal de ellos es la falta de veracidad de las suposiciones sobre las que se asienta la técnica de reconstrucción basada en correlaciones.

La segunda técnica de reconstrucción propuesta en [224] se denomina reconstrucción basada en regiones (CBR, *Cluster-Based Reconstruction*). Esta técnica asume que las características de la voz se modelan mediante un GMM con M componentes. Usando este GMM la técnica CBR desarrolla la siguiente estimación MAP de las características perdidas del espectro,

$$\hat{\boldsymbol{x}}_u = \underset{\boldsymbol{x}_u}{\operatorname{argmax}}\, p(\boldsymbol{x}_u | \boldsymbol{x}_u \leq \boldsymbol{y}_u, \boldsymbol{x}_r = \boldsymbol{y}_r). \tag{2.164}$$

La estimación MAP acotada definida en la ecuación anterior resulta, no obstante, difícil de implementar cuando la densidad de probabilidad de la variable $\boldsymbol{x}$ se distribuye según un GMM. Para simplificar esta tarea, Raj propone aproximar el estimador anterior mediante una combinación lineal de estimaciones parciales obtenidas para cada gaussiana del GMM,

$$\hat{\boldsymbol{x}}_u = \sum_{k=1}^{M} P(k | \boldsymbol{y}_r, \boldsymbol{x}_u \leq \boldsymbol{y}_u) \hat{\boldsymbol{x}}_u^{(k)}, \tag{2.165}$$

donde $\hat{\boldsymbol{x}}_u^{(k)}$ representa la estimación MAP acotada obtenida para la componente k-ésima del GMM y $P(k|\boldsymbol{y}_r, \boldsymbol{x}_u \leq \boldsymbol{y}_u)$ es su probabilidad a posteriori.

Aplicando la regla de Bayes, la probabilidad a posteriori de cada gaussiana se puede expresar de la siguiente forma:

$$P(k | \boldsymbol{y}_r, \boldsymbol{x}_u \leq \boldsymbol{y}_u) = \frac{p(\boldsymbol{y}_r, \boldsymbol{x}_u \leq \boldsymbol{y}_u | k) P(k)}{\sum_{k'=1}^{M} p(\boldsymbol{y}_r, \boldsymbol{x}_u \leq \boldsymbol{y}_u | k') P(k')}. \tag{2.166}$$

En la ecuación anterior, $P(k)$ $(k = 1, \ldots, M)$ son los pesos (probabilidades a priori) de las componentes del GMM. Por otra parte, la probabilidad de observación $p(\boldsymbol{y}_r, \boldsymbol{x}_u \leq$

$\boldsymbol{y}_u|k)$ se puede obtener, bajo la suposición de diagonalidad de las matrices de covarianza, usando la ecuación (2.161).

El otro término requerido por el estimador de la ecuación (2.165) son las estimas parciales $\hat{\boldsymbol{x}}_u^{(k)}$ para cada gaussiana k del GMM . Éstas se computan de acuerdo al siguiente estimador MAP acotado:

$$\hat{\boldsymbol{x}}_u^{(k)} = \underset{\boldsymbol{x}_u}{\operatorname{argmax}}\, p(\boldsymbol{x}_u|\boldsymbol{x}_u \leq \boldsymbol{y}_u, \boldsymbol{y}_r, k). \tag{2.167}$$

En el caso de utilizar matrices de covarianza diagonales, el estimador anterior coincide con el valor mínimo, elemento a elemento, entre los valores de voz observados y la media marginal de los datos perdidos,

$$\hat{\boldsymbol{x}}_u^{(k)} = \text{mín}(\boldsymbol{y}_u, \boldsymbol{\mu}_u^{(k)}). \tag{2.168}$$

Si se emplean matrices de covarianza completas, desafortunadamente, no existe una expresión cerrada para el cálculo de (2.167). En tal caso, Raj propone en [224] un algoritmo iterativo para calcular la estimación MAP anterior. No obstante, el procedimiento iterativo propuesto puede llegar a ser muy costoso computacionalmente hablando, por lo que se sugiere su simplificación como,

$$\hat{\boldsymbol{x}}_u^{(k)} = \text{mín}(\boldsymbol{\mu}_{u|r}^{(k)}, \boldsymbol{y}_u), \tag{2.169}$$

donde $\boldsymbol{\mu}_{u|r}^{(k)}$ es la media condicional definida en (2.157).

BCMI. Una técnica de reconstrucción similar a la que acabamos de presentar es BCMI (*Bounded Conditional Mean Imputation*, imputación basada en la media condicional acotada) [88]. En esta técnica, al igual que CBR, utiliza un GMM para modelar las características de la voz, pero en este caso se fuerza que las matrices de covarianza sean completas. Con ello se pretende aprovechar mejor las correlaciones entre las distintas características para obtener una reconstrucción más precisa. Otra diferencia entre ambas técnicas es el uso del estimador MAP en CBR, mientras que BCMI recurre a un estimador MMSE para obtener las estimas de las características perdidas:

$$\hat{\boldsymbol{x}}_u = \mathbb{E}[\boldsymbol{x}_u|\boldsymbol{y}_r, \boldsymbol{x}_u \leq \boldsymbol{y}_u] = \int_{-\infty}^{\boldsymbol{y}_u} \boldsymbol{x}_u p(\boldsymbol{x}_u|\boldsymbol{y}_r, \boldsymbol{y}_u) d\boldsymbol{x}_u, \tag{2.170}$$

donde se ha utilizado una notación simplificada en los límites de la integral para denotar la integral multidimensional definida sobre el hipercubo $(-\infty, y_{u,1}] \times (-\infty, y_{u,2}] \times \ldots$

Desarrollando la probabilidad condicional $p(\boldsymbol{x}_u|\boldsymbol{x}_r, \boldsymbol{y}_u)$ en (2.170) usando el GMM de voz limpia, se obtiene una expresión muy similar a la obtenida para CBR,

$$\hat{\boldsymbol{x}}_u = \sum_{k=1}^{M} P(k|\boldsymbol{y}_r, \boldsymbol{y}_u) \underbrace{\int_{-\infty}^{\boldsymbol{y}_u} \boldsymbol{x}_u p(\boldsymbol{x}_u|\boldsymbol{y}_r, \boldsymbol{y}_u, k) d\boldsymbol{x}_u}_{\hat{\boldsymbol{x}}_u^{(k)}}. \tag{2.171}$$

Constatamos que el estimador resultante es una combinación lineal de estimas parciales $\hat{\boldsymbol{x}}_u^{(k)}$ ponderadas por sus respectivas probabilidades a posteriori $P(k|\boldsymbol{y}_r, \boldsymbol{y}_u)$. Tras aplicar la regla de Bayes, se puede comprobar que las probabilidades a posteriori son proporcionales al producto de los pesos de las gaussianas $P(k)$ multiplicados por la siguiente probabilidad de observación:

$$p(\boldsymbol{y}_r, \boldsymbol{y}_u|k) = p(\boldsymbol{y}_r|k) \int_{-\infty}^{\boldsymbol{y}_u} \boldsymbol{x}_u p(\boldsymbol{x}_u|\boldsymbol{y}_r, k) d\boldsymbol{x}_u. \tag{2.172}$$

En la ecuación anterior, $p(\boldsymbol{y}_r|k) = \mathcal{N}(\boldsymbol{y}_r; \boldsymbol{\mu}_r^{(k)}, \boldsymbol{\Sigma}_r^{(k)})$ es directamente computable mediante la distribución marginal de los datos fiables obtenida para la gaussiana k. En cambio, no existe una expresión analítica cerrada para el cálculo de la integral múltiple, cuando las matrices de covarianza utilizadas no son diagonales. Éste es uno de los problemas aún no resueltos dentro del paradigma de datos perdidos y que es compartido por todas aquellas técnicas que usan mezclas de gaussianas (marginalización, SFD e imputación).

Varias son las tentativas que se han propuesto para resolver este problema. Ya veremos en la sección 4.2 que una de las soluciones más simples, pero a la vez más efectivas, consiste en forzar que la matriz de covarianza de $p(\boldsymbol{x}_u|\boldsymbol{y}_r, k)$ sea diagonal, poniendo a cero todos los elementos de la matriz que quedan fuera de la diagonal principal. Otra estrategia que también puede aplicarse para evaluar esta integral es la integración de Montecarlo [115] o el método numérico de integración basado en rejilla propuesto en [240]. Aunque en teoría estos métodos son capaces de aproximar el valor de la integral con una precisión arbitraria, el coste de los mismos se vuelve prohibitivo a medida que la dimensión de $\boldsymbol{x}_u$ crece. En [240], además de la técnica de integración numérica que se ha comentado, también se propone otra solución que sorprende por su simplicidad y su elegancia: el valor de la integral se aproxima por el valor $\boldsymbol{x}_u$ que maximiza la probabilidad $p(\boldsymbol{x}_u|\boldsymbol{y}_r, k)$ dentro del área de integración. Cuando se compara esta técnica con los métodos de integración numérica, se observa que ambas técnicas obtienen resultados similares, pero la aproximación por el máximo es mucho más eficiente.

La solución propuesta en [88] para BCMI es diagonalizar la matriz de covarianza de la PDF $p(\boldsymbol{x}_u|\boldsymbol{y}_r, k) = \mathcal{N}(\boldsymbol{x}_u; \boldsymbol{\mu}_{u|r}^{(k)}, \boldsymbol{\Sigma}_{u|r}^{(k)})$ en la ecuación (2.172). Para ello los datos son transformados usando la siguiente expresión,

$$\tilde{\boldsymbol{x}}_u^{(k)} = \boldsymbol{H}(\boldsymbol{x}_u - \boldsymbol{\mu}_{u|r}^{(k)}), \tag{2.173}$$

donde $\boldsymbol{H}$ es la matriz triangular superior obtenida tras factorizar por Cholesky la matriz de precisión de la PDF anterior, esto es,

$$\boldsymbol{\Sigma}_{u|r}^{(k)^{-1}} = \boldsymbol{H}\boldsymbol{H}^\top. \tag{2.174}$$

Tras aplicar la transformación afín de la ecuación (2.173) a la ecuación (2.172), en [88] se prueba que la probabilidad de observación se puede aproximar por[1]

$$p(\boldsymbol{y}_r, \boldsymbol{y}_u|k) \approx \mathcal{N}(\boldsymbol{y}_r; \boldsymbol{\mu}_r^{(k)}, \boldsymbol{\Sigma}_r^{(k)}) \prod_{j \in \boldsymbol{s}_u} \Phi(\tilde{y}_{u,j}^{(k)}), \tag{2.175}$$

con

$$\tilde{\boldsymbol{y}}_u^{(k)} = \boldsymbol{H}(\boldsymbol{y}_u - \boldsymbol{\mu}_{u|r}^{(k)}) \tag{2.176}$$

y $\Phi(\cdot)$ es la función de distribución de probabilidad acumulada (CDF, *Cumulative Distribution Function*) para la distribución normal, esto es,

$$\Phi(x) = \int_{-\infty}^{x} \mathcal{N}(u)du. \tag{2.177}$$

Análogamente a lo hecho durante el cómputo de las probabilidades de observación, durante el cálculo de $\hat{\boldsymbol{x}}_u^{(k)}$ se diagonaliza la matriz de covarianza de $p(\boldsymbol{x}_u|\boldsymbol{y}_r, k)$ mediante la transformación definida en la ecuación (2.173). Tras esta transformación, se obtiene la siguiente expresión para el cálculo de las estimas parciales:

$$\hat{\boldsymbol{x}}_u^{(k)} \approx \boldsymbol{\mu}_{u|r}^{(k)} - \boldsymbol{H}^{-1} \left[\frac{\mathcal{N}\left(\tilde{y}_{u,1}^{(k)}\right)}{\Phi\left(\tilde{y}_{u,1}^{(k)}\right)}, \frac{\mathcal{N}\left(\tilde{y}_{u,2}^{(k)}\right)}{\Phi\left(\tilde{y}_{u,2}^{(k)}\right)}, \cdots \right]^T. \tag{2.178}$$

Imputación en el dominio cepstral. Un denominador común de las dos técnicas de reconstrucción presentadas hasta ahora, es el uso de modelos probabilísticos de la voz con poca resolución (GMMs) que la modelan en el domino espectral. Asimismo, ambas técnicas se plantean como técnicas robustas de extracción de características, en el sentido de que proporcionan características de voz libres de distorsiones al motor de reconocimiento. Una estrategia alternativa a estas técnicas es la que Van Segbroeck propone en [265]. A diferencia de las técnicas de reconstrucción anteriores, el objetivo principal de este trabajo es la aplicación de las técnicas basadas en el paradigma de datos perdidos al dominio cepstral, o a cualquier otro dominio que suponga una transformación lineal de las características log-Mel [261, 262, 263]. Otra diferencia clave de la aproximación planteada por Van Segbroeck es el uso del propio modelo acústico del reconocedor (HMM entrenado con coeficientes cepstrales y sus derivadas) como modelo probabilístico a priori de alta resolución para la estimación de los elementos perdidos del espectro.

[1] La equivalencia entre ambas ecuaciones no es exacta, ya que tras aplicar la transformación de (2.173) a (2.172) también se modifican los límites de la integral, no quedando estos alineados respecto a los nuevos ejes. Esta pequeña discrepancia, no obstante, no es considerada en [88].

Formalmente, el problema que Van Segbroeck analiza [265] consiste en, para cada gaussiana k del modelo acústico, estimar la energía de la voz en aquellos elementos del vector de entrada etiquetados como perdidos. Para llevar a cabo esta tarea, en [265] se elige el criterio de optimización ML. Esto da lugar a la siguiente optimización con restricciones:

$$\begin{aligned} \hat{\boldsymbol{x}} &= \underset{\boldsymbol{x}}{\operatorname{argmin}} \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}^{(k)})^{\top}\boldsymbol{P}^{(k)}(\boldsymbol{x}-\boldsymbol{\mu}^{(k)}) \\ \text{sujeto a } & \ \boldsymbol{x}_r = \boldsymbol{y}_r \ \text{ y } \ \boldsymbol{x}_u \leq \boldsymbol{y}_u. \end{aligned} \tag{2.179}$$

En la ecuación anterior $\boldsymbol{P}^{(k)}$ es la matriz de precisión de la gaussiana k-ésima expresada en el dominio log-Mel y $\boldsymbol{\mu}^{(k)}$ es la media de dicha gaussiana expresada también en ese mismo dominio. Las fórmulas que nos permiten transformar estos parámetros entre el dominio cepstral y el domino log-Mel son (se omiten los índices de las gaussianas por claridad),

$$\boldsymbol{\mu} = \boldsymbol{C}^{-1}\boldsymbol{\mu}^{c}, \tag{2.180}$$

$$\boldsymbol{P} = \boldsymbol{C}^{-1}\boldsymbol{\Sigma}^{c}\boldsymbol{C} + \boldsymbol{\Psi}, \tag{2.181}$$

donde el superíndice c indica que las características se expresan en el dominio cepstral y $\boldsymbol{\Psi}$ es una matriz cuadrada que se emplea para forzar a que el rango de la matriz de precisión resultante, $\boldsymbol{P}$, no sea deficiente.

Planteado el problema de optimización a resolver en la ecuación (2.179), en [265] se demuestra que la optimización de esta función es equivalente a resolver un problema de mínimos cuadrados con restricción de no negatividad (NNLSQ, *Non-Negative Least Square*). Para resolver el problema NNLSQ, en [265] se emplea una estrategia iterativa basada en el algoritmo del descenso en gradiente. Este algoritmo se repetirá, siempre y cuando la máscara de segregación indique que hay datos perdidos, para cada gaussiana del modelo y cada instante de tiempo, lo cual puede llegar a ser inviable en ciertos sistemas. Como resultado final de todo este proceso, el algoritmo de Viterbi devolverá la secuencia de palabras más probable reconocida dados el espectrograma incompleto y la información proporcionada por la máscara de segregación. Asimismo, si se desea también puede obtenerse el espectrograma de la señal reconstruida, sin más que combinar las estimaciones parciales obtenidas para el camino de máxima probabilidad.

Imputación rala. Fruto del impacto que ha supuesto entre la comunidad científica la teoría del *Compressed Sensing* (CS) [45, 76], en los últimos ha crecido el interés por las técnicas que aplican restricciones de dispersión al problema de reconstrucción de espectrogramas incompletos. Una de las primeras técnicas que combinan el paradigma

de datos perdidos con la teoría del CS fue propuesta en [114]. Con objeto de simplificar la descripción de esta técnica, supondremos que la tarea de reconocimiento a abordar es reconocer dígitos aislados. Asimismo, también supondremos que el espectrograma de cada dígito se representa mediante un supervector de tamaño $Z = D \times T$, donde D es el número de componentes en frecuencia y T la duración temporal. Por simplicidad en la exposición, la duración temporal T se considerará fija para todos los dígitos, pudiendo recurrir a técnicas de interpolación/diezmado para conseguir tal duración fija. En la técnica propuesta en [114], al contrario que las técnicas de reconstrucción anteriores, no se utiliza un modelo paramétrico de la voz (GMM o HMM), sino se representa mediante un diccionario que contiene N espectrogramas de ejemplo $\{\boldsymbol{a}_n\}$ $(n = 1, \ldots, N)$. Generalmente N es mucho mayor que el número de dígitos a reconocer, habiendo por tanto varios ejemplos por cada dígito. Con esto se pretende reflejar la variabilidad acústica que pueda existir en la pronunciación de dicho dígito.

Dado $\boldsymbol{x}$ un espectrograma cualquiera, suponemos que éste puede representarse mediante una combinación lineal de los N ejemplos del diccionario,

$$\boldsymbol{x} = \sum_{n=1}^{N} w_n \boldsymbol{a}_n = \boldsymbol{A}\boldsymbol{w}, \tag{2.182}$$

donde $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N)$ es la matriz $Z \times N$ que contiene los ejemplos del diccionario en sus columnas y $\boldsymbol{w}$ es el vector con los pesos asociados a cada ejemplo.

De entre las posibles alternativas que hay para hallar el vector de pesos $\boldsymbol{w}$, en [114] se elige aquella que, según la teoría del CS, permita reconstruir $\boldsymbol{x}$ usando una combinación rala de los ejemplos del diccionario,

$$\begin{aligned} \boldsymbol{w} &= \underset{\tilde{\boldsymbol{w}}}{\operatorname{argmin}} \|\tilde{\boldsymbol{w}}\|_0 \\ \text{sujeto a } \ \boldsymbol{x} &= \boldsymbol{A}\tilde{\boldsymbol{w}}. \end{aligned} \tag{2.183}$$

En esta ecuación el operador $\|\cdot\|_0$ indica la norma cero, ℓ_0, del vector que tiene por argumento, esto es, el número de elementos distintos de cero. Este problema, no obstante, es sumamente complicado de resolver, ya que se trata de un problema combinatorio NP-duro. Bajo ciertas restricciones sobre el diccionario $\boldsymbol{A}$, es posible demostrar que la solución al problema de optimización anterior equivale a la minimización de la norma ℓ_1. Este último problema puede traducirse, a su vez, en el siguiente problema de mínimos cuadrados con término de regularización λ [114, 256],

$$\boldsymbol{w} = \underset{\tilde{\boldsymbol{w}}}{\operatorname{argmin}} \|\boldsymbol{A}\tilde{\boldsymbol{w}} - \boldsymbol{x}\|_2 + \lambda \|\tilde{\boldsymbol{w}}\|_1 . \tag{2.184}$$

La idea expresada en la ecuación anterior por la que $\boldsymbol{x}$ se representa como una combinación lineal rala de elementos de un diccionario, la podemos extender al caso

que nos ocupa, a saber, la reconstrucción de los elementos no fiables del espectro. Supongamos ahora que el espectrograma del dígito a reconocer se representa mediante el supervector $\boldsymbol{y}$ de dimensión $D \times T$. Dada la máscara de segregación $\boldsymbol{m}$, también de dimensión $D \times T$, podemos dividir al vector observado en sus componentes fiables $\boldsymbol{y}_r$ y las no fiables $\boldsymbol{y}_u$. De forma similar, podemos aplicar la máscara $\boldsymbol{m}$ al diccionario $\boldsymbol{A}$ para obtener las submatrices $\boldsymbol{A}_r$ y $\boldsymbol{A}_u$, en donde $\boldsymbol{A}_r$ contiene las filas de $\boldsymbol{A}$ para las que la máscara toma el valor 1 y $\boldsymbol{A}_u$ aquellas donde toma el valor 0. En base a estos términos, es posible calcular el vector de pesos $\boldsymbol{w}$ que minimiza la distorsión de reconstrucción de $\boldsymbol{y}_r$ y que, además, cumple con la condición de dispersión impuesta por la norma ℓ_1,

$$\boldsymbol{w} = \underset{\tilde{\boldsymbol{w}}}{\operatorname{argmin}} \left\| \boldsymbol{A}_r \tilde{\boldsymbol{w}} - \boldsymbol{y}_r \right\|_2 + \lambda \left\| \tilde{\boldsymbol{w}} \right\|_1 . \tag{2.185}$$

Finalmente, el espectro estimado de la voz limpia se obtiene a partir del vector de pesos calculado en (2.185) y de los elementos fiables de la señal de voz observada,

$$\hat{\boldsymbol{x}} = \begin{cases} \hat{\boldsymbol{x}}_r = \boldsymbol{y}_r \\ \hat{\boldsymbol{x}}_u = \text{mín}(\boldsymbol{A}_u \boldsymbol{w}, \boldsymbol{y}_u) \end{cases} . \tag{2.186}$$

Uno de las ventajas de la técnica de imputación que acabamos de presentar, es que ésta modela de forma explícita la evolución temporal de la voz en el diccionario de ejemplos $\boldsymbol{A}$. Esto permite un uso más eficiente de la redundancia de la voz, consiguiendo con ello una estimación más precisa de las componentes perdidas del espectro. Como contrapartida debe mencionarse que, tal y como se ha descrito, la técnica propuesta en [114] sólo es aplicable a sistemas de reconocimiento de palabras aisladas. Para abordar este problema, en [112] se presenta una extensión en la que se emplea una ventana deslizante para extraer los ejemplos del diccionario y, por tanto, poder aplicar la imputación a sistemas de habla continua. Otras extensiones a esta técnica también han considerado su aplicación a reconstrucción con máscaras de segregación continuas [111], así como enfoques híbridos que combinan el realce de las características de la voz y la propagación de la incertidumbre de la reconstrucción al reconocedor [113].

Imputación usando máscaras continuas. Las distintas estrategias de imputación presentadas anteriormente comparten una crítica común y es el uso de máscaras de segregación binarias. Como se demuestra en [31], el uso de máscaras de segregación continuas proporciona un mejor rendimiento que las binarias en situaciones reales, es decir, cuando las máscaras deben estimarse. En efecto, los errores en la estimación de la máscara binaria pueden afectar de forma decisiva a la tarea de imputación. Por

ejemplo, si un elemento no distorsionado del espectrograma se etiqueta como perdido en la máscara, las técnicas de imputación tratarán de reemplazarlo por cierta estima, agravando con ello la distorsión. Si, por el contrario, el elemento era originalmente no fiable y por error se le etiqueta como fiable, no sólo se conservará intacto, sino que además afectará a la estimación del resto de elementos no fiables de espectro.

Por las razones anteriores, algunos autores han propuesto versiones de sus técnicas de imputación que pueden operar con máscaras continuas. Por ejemplo, en [90, 225] se proponen sendas extensiones paras las técnicas CBR y BCMI estudiadas anteriormente. Brevemente, la estimación parcial obtenida por estas técnicas para cada elemento del vector observado $\boldsymbol{y} = (y_1, \ldots, y_D)$ se obtiene mediante la siguiente combinación lineal:

$$\mathbb{E}[x_i|y_i, m_i, k] = m_i y_i + (1 - m_i)\tilde{\mu}_i^{(k)}(y_i), \tag{2.187}$$

donde m_i es el valor de la máscara para el elemento i-ésimo del vector, y_i puede considerarse como la estimación de la voz limpia para valores altos de SNR y $\tilde{\mu}_i^{(k)}(y_i)$ es la estima de la energía de la voz obtenida para la gaussiana k en el caso de que el ruido enmascare completamente la voz. La estimación resultante, como el lector puede apreciar, consta ahora de una decisión suave, en lugar de la decisión ruda que el uso de máscaras binarias acarrea. Experimentalmente se ha demostrado que este tipo de decisiones son más robustas a los errores en la estimación de la segregación del espectro [90, 225].

Con esto concluye el estudio de las aproximaciones propuestas en la literatura para el reconocimiento de la voz en presencia de datos perdidos. Como puede observarse, un requisito compartido por la gran mayoría de las técnicas estudiadas es disponer de una máscara que permita identificar las regiones del espectro dominadas por la voz del locutor, de aquellas dominadas por el ruido. Dada la gran importancia de esta tarea, en la siguiente sección procedemos a revisar, de forma breve, distintas alternativas que se han propuesto para lograr tal cometido.

2.2.5.5. Estimación de la máscara de segregación

En la literatura podemos encontrar una miríada de técnicas propuestas para la estimación de las máscaras de segregación (o máscaras *missing data*) requeridas por las técnicas MD, estando la gran mayoría de estas técnicas recogidas en el artículo de revisión bibliográfica [49]. En este apartado intentaremos organizar estas técnicas en distintas categorías y daremos ejemplos de las técnicas más representativas dentro de cada categoría.

La primera distinción que podemos hacer en las máscaras de segregación es si éstas son ideales, también denominadas máscaras oráculo, o si han sido estimadas. Las

máscaras oráculo suponen un conocimiento perfecto de la segregación entre voz y ruido en el espectrograma observado de la voz. Generalmente esta suposición no es realista, debiéndose de estimar en la práctica la fiabilidad de cada elemento del espectrograma en base a otra información colateral disponible a priori, como ya veremos. No obstante, dado que contienen la segregación ideal del espectro, este tipo de máscaras suelen ser muy útiles a la hora de establecer la cota superior a la que podemos aspirar en el rendimiento de una técnica dada, ya sea de una técnica de estimación de máscaras, de imputación o de marginalización. El cálculo de este tipo de máscaras sólo puede llevarse a cabo en aquellas situaciones en las que sea posible estimar, de forma precisa, la densidad espectral de potencia del ruido. Por esta razón, las máscaras oráculo se han empleado con profusión en aquellos trabajos que trabajan con bases de datos en las que se tiene acceso a grabaciones estéreo (p.ej. Aurora2 [141] y Aurora4 [140]).

Por otro lado tenemos las máscaras estimadas que, como su propio nombre indica, se obtienen como resultado de un procedimiento que permite estimar la fiabilidad de cada uno de los elementos que constituyen el espectrograma de la señal observada. Como cabe prever, el rendimiento de estas máscaras es menor que el de las máscaras oráculo, al estar sujetas a numerosas fuentes de incertidumbre fruto del proceso de estimación al que se ven sometidas.

Otra clasificación alternativa de las máscaras de segregación es la derivada del tipo de decisión voz/ruido que se aplica. En primer lugar encontramos las máscaras binarias, en donde cada elemento del espectro se etiqueta bien como voz ($m_i = 1$) o como ruido ($m_i = 0$). Por otro lado, en las máscaras continuas el grado de fiabilidad de un elemento en cuestión se mueve ahora en el intervalo $m_i \in [0, 1]$. Estas máscaras suelen conducir a una interpretación probabilística del proceso de enmascaramiento de la voz, en donde cada elemento de la máscara mide la probabilidad de que la energía de la voz domine a la del ruido en dicha componente tiempo-frecuencia del espectrograma [21, 40, 116],

$$m_i = P(x_i > n_i). \tag{2.188}$$

Las dos clasificaciones que hemos realizado sobre las máscaras de segregación se centran en aspectos concretos de la máscara en sí, pero no en el método empleado para obtenerlas. Desde este último punto de vista, podemos clasificar las técnicas empleadas para estimar las máscaras de segregación en cuatro grupos: (i) técnicas que estiman la SNR local, (ii) técnicas de clasificación, (iii) enfoques CASA y (iv) otros enfoques. En los siguientes párrafos enumeraremos los detalles más relevantes de cada grupo de técnicas.

Máscaras basadas en la estimación de la SNR local. El método más simple y directo para el cálculo de la fiabilidad de cada uno de los elementos del espectro,

consiste en obtener la SNR de los mismos a partir de cierta estimación de la potencia del ruido (ver sección 2.3) en esa componente tiempo-frecuencia. Si la SNR local es mayor que 0 dB, la potencia de la voz será mayor que la del ruido y, por lo tanto, podremos considerar que ese elemento es fiable.

Notemos como $Y_i(t)$ a la respuesta del filtro Mel i-ésimo ($i = 1, \ldots, D$) para el instante de tiempo t y como $\hat{N}_i(t)$ a la potencia estimada del ruido, ambos términos en escala lineal. La SNR estimada para dicho elemento se puede definir como

$$SNR_i(t) = 20 \log_{10} \frac{Y_i(t) - \hat{N}_i(t)}{\hat{N}_i(t)}. \tag{2.189}$$

En base a esta SNR estimada, podemos calcular la fiabilidad de cada elemento del espectrograma. Si lo que se desea es una máscara binaria, dicha fiabilidad se puede calcular umbralizando los valores de SNR en base a cierto valor γ,

$$m_i(t) = \begin{cases} 1 & \text{si } SNR_i(t) \geq \gamma \\ 0 & \text{en otro caso} \end{cases}. \tag{2.190}$$

El umbral γ se suele determinar experimentalmente usando conjuntos de validación, de tal forma se maximice la tasa de reconocimiento en relación a la señal reconstruida. En general, este umbral suele tomar valor dentro del rango $\gamma \in [1, 7]$ dB, consiguiendo con ello un compromiso entre información descartada del espectro original, esto es, elementos fiables considerados no fiables, y la aceptación de falsos positivos, es decir, frecuencias dominadas por la energía del ruido consideradas como fiables. En general, el último caso suele ser más perjudicial para el proceso de reconocimiento, por lo que se tiende a preferir umbrales de SNR altos [249].

Por otro lado, si lo que se desea es una máscara continua, podemos transformar el rango de posibles valores de SNR al intervalo $[0, 1]$ mediante una función sigmoide de la forma [31],

$$m_i(t) = \frac{1}{1 + e^{-\alpha(SNR_i(t) - \beta)}} \tag{2.191}$$

siendo α el parámetro que controla la pendiente de la sigmoide y β el centro de la misma, esto es, el nivel de SNR para el cual la máscara toma el valor 0,5. Estos valores, de nuevo, se estiman empíricamente usando conjuntos de validación.

Una de las críticas más importantes que comúnmente se le achacan a las técnicas de estimación espectral del ruido y, por ende, a las máscaras basadas en la estimación de la SNR local, son las suposiciones de ruido estacionario que suelen realizar. No obstante, el ruido raramente es estacionario, por lo que estas suposiciones acarrean una desviación de la estima respecto a la distribución real el mismo. Debido a esto, diversos investigadores se han decantado por una estrategia basada en la clasificación de los elementos del espectro para el cómputo de las máscaras de segregación.

Segregación basada en la clasificación bayesiana del espectro. Desde el punto de vista del modelo de enmascaramiento de la ecuación (2.152), el espectro observado puede interpretarse como una mezcla exclusiva elemento a elemento de los espectros de voz limpia y ruido. Bajo esta perspectiva, la estimación de la máscara de segregación puede verse ahora como un proceso de clasificación binario de los elementos del espectro en las categorías voz y ruido. En la literatura podemos encontrar diferentes técnicas que se encuadran dentro de esta nueva estrategia de obtención de las máscaras de segregación.

En [241] se propone un clasificador bayesiano que emplea GMMs independientes para diferenciar la voz del ruido. El GMM que modela el ruido se entrena con voz distorsionada por ruido blanco gaussiano. En el caso del clasificador de la voz, éste emplea a su vez dos GMMs diferentes: uno para representar los sonidos sonoros y otros para los sordos. Cada uno de estos GMMs se entrenan con un conjunto de características extraídas de la señal de voz que muestran ser robustas a la distorsión producida por el ruido. Las características extraídas en los segmentos sonoros son la salida de un filtro peine que muestrea la señal en los armónicos de la frecuencia fundamental, el ratio entre la energía de las sub-bandas y la energía total de cada trama y la forma espectral del mismo medida en términos de su curtosis y el grado en que ésta tiende a ser plana. Para los segmentos sordos, se utilizan las mismas características de la voz sonora salvo a aquellas que involucran a la frecuencia fundamental. En comparación con máscaras tradicionales basadas en estimación del ruido, las máscaras basadas en clasificación bayesiana muestran mejores resultados de reconocimiento sobre una amplio rango de tipos de ruido y niveles de SNR [241].

En [162] se proponen varias mejoras al clasificador propuesto en [241]. Estas mejoras pueden resumirse en el empleo de un rango mayor de ruidos coloreados para entrenar el clasificador de ruido y el uso de un conjunto extendido de características de voz. De nuevo, las máscaras de segregación obtenidas producen mejoras significativas en comparación con las basadas en la estimación de la SNR local.

Enfoques basados en el CASA. Un tercer grupo de técnicas de estimación de las máscaras de segregación lo constituyen aquellas que emplean primitivas del CASA para tal fin. En [183, 186], por ejemplo, se propone un algoritmo de estimación multi-pitch especialmente pensado para su uso en el generador de fragmentos de un sistema de reconocimiento SFD. Tras la etapa de reconocimiento, el sistema SFD es capaz de devolver no sólo la secuencia de palabras más probable, sino también la hipótesis de segregación (máscara) que maximiza la tasa de reconocimiento. Esta máscara puede ser usada posteriormente por un sistema de reconocimiento convencional basado en

imputación, consiguiendo este último sistema una mejor tasa de reconocimiento que la alcanzada por el sistema SFD por sí solo [184].

Otros trabajos también han utilizado la información que proporciona la frecuencia fundamental de la voz para estimar las máscaras de segregación (ver p.ej. [28]). La idea subyacente parte de la suposición de que la voz es la única fuente armónica dominante en la señal observada. Bajo este supuesto, aquellas componentes en frecuencia que estén armónicamente relacionadas pueden agruparse y ser consideradas como parte de la señal de voz en la máscara. No obstante, este razonamiento falla para aquellos sonidos en los que las cuerdas vocales no vibran y, por consiguiente, no existe periodicidad en la voz. En estos casos lo habitual es combinar las máscaras basadas en la armonicidad de la voz con máscaras basadas en la estimación de la SNR local [28], o con máscaras que explotan otras primitivas básicas de agrupación como por ejemplo el sincronismo en los *onsets*/*offsets* [144, 145].

Otros enfoques para la estimación de las máscaras de segregación. Además de los estrategias anteriores, en la literatura también se han propuesto otros enfoques alternativos para la estimación de las máscaras de segregación. Por un lado podemos distinguir aquellas técnicas que usan *arrays* de micrófonos para estimar la posición de cada una de las fuentes presentes en la escena auditiva y, en base a información, construir la máscara de segregación [133, 271]. Por otro lado nos encontramos con aquellas técnicas que estiman la probabilidad de presencia de la voz (SPP, *Speech Presence Probability*), esto es, la probabilidad de que la energía de la voz sea mayor que la del ruido, en base a modelos estadísticos de voz y ruido disponibles a priori [21, 40, 116]. Para estimar la SPP, el primer paso es construir, implícita o explícitamente, el modelo de voz distorsionada a partir de los dos modelos anteriores. A continuación se infiere la máscara de segregación dado el modelo de voz ruidosa recurriendo a cierto criterio de optimización (p.ej. ML). Este segundo enfoque combina las ventajas de las técnicas basadas en la estimación de la SNR al emplear un modelo de ruido y de las técnicas de clasificación bayesiana por usar un modelo de voz limpia.

2.3. Estimación del modelo de ruido

La gran mayoría de técnicas de reconocimiento robusto de voz estudiadas en las secciones anteriores requieren del conocimiento de un modelo de ruido para operar (adaptar los modelos acústicos o estimar el vector de voz limpia). En esta tesis supondremos que el modelo de ruido viene dado por $\mathcal{M}_n = \{\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n, \boldsymbol{\mu}_h\}$, donde $\boldsymbol{\mu}_n$ y $\boldsymbol{\Sigma}_n$ son los parámetros que modelan el ruido aditivo (media y matriz de covarianza) y $\boldsymbol{\mu}_h$

representa el desplazamiento en los dominios log-Mel o cepstral de las característica de voz debido al ruido convolutivo, supuesto que éste es estacionario ($\mathbf{\Sigma}_h = \mathbf{0}$). Aunque en el transcurso de esta sección se asumirá por simplicidad que el ruido aditivo se representa mediante una única gaussiana, distintos autores [95, 166, 250] han considerado un GMM para tal fin con resultados ligeramente superiores.

En la literatura se han propuesto numerosas técnicas de estimación del ruido, las cuales varían su modo de operación en función del objetivo último para el que fueron diseñadas. Por un lado, encontramos con técnicas de estimación espectral del ruido que en su origen fueron ideadas como complemento a las técnicas de realce de voz. Por otro lado, existen un conjunto de técnicas específicamente pensadas para la estimación del modelo de ruido en el contexto del reconocimiento robusto de voz. Como veremos a continuación, estas técnicas suponen el ruido como una variable aleatoria oculta y tratan de estimar sus características de forma iterativa usando el algoritmo EM [66]. En los siguientes apartados se analizarán en detalle estos dos enfoques.

2.3.1. Estimación espectral del ruido

Uno de los enfoques más simples para la estimación de la estadística del ruido aditivo consiste en identificar los segmentos de no voz de la señal de entrada y, a partir de estos, obtener una estimación del espectro del ruido. En este sentido multitud de técnicas suponen que las primeras y últimas tramas de cada frase (típicamente 10-30 tramas) son silencio, por lo que pueden ser utilizadas para estimar la densidad de potencia espectral del ruido. Este esquema tan simple de estimación del ruido se ha empleado con éxito en tareas de reconocimiento donde el ruido puede ser considerado estacionario, bien porque así lo sea o bien porque las frases que se reconocen sean cortas (alrededor de 2-3 segundos). En particular, en la literatura encontraos numerosos trabajos (p.ej. [35, 125, 126, 237, 239]) que aplican este enfoque para estimar el ruido en la tarea de reconocimiento definida por Aurora2 [141], obteniendo un rendimiento más que notable.

En aquellas situaciones en las que el ruido no es estacionario, el enfoque anterior no proporciona buenas estimas del ruido. La extensión lógica al mismo consiste en utilizar un VAD (*Voice Activity Detector*, detector de actividad de voz) [14, 227, 247, 257] que permita identificar los segmentos de voz y silencio de la frase y utilizar los últimos para estimar progresivamente el espectro del ruido. Este enfoque, no obstante, no se encuentra libre de problemas. El primero de ellos, y quizás el más acuciante, reside en la dificultad de la detección de los segmentos de voz a muy bajas SNRs. En efecto, para valores de SNR bajos, el error de clasificación en la detección de la voz se dispara [258]. Por otra parte, el uso de VAD obliga a suponer que el ruido es estacionario en los

segmentos de voz. En tales casos la estimación del ruido se suele calcular interpolando las estimas obtenidas para los segmentos de silencio que circundan el segmento de voz en cuestión.

Es posible, no obstante, estimar el espectro del ruido incluso en los segmentos de voz. Así, distintos trabajos [52, 195, 196] aprovechan la estructura armónica de la voz en los segmentos sonoros para estimar el ruido en los valles del espectro. No obstante, estas técnicas requieren conocer a priori el *pitch* de la voz para operar. Al igual que ocurre con la detección de actividad de voz, la identificación robusta de la frecuencia fundamental para SNRs bajas es un problema arduo. Como último problema de las técnicas de estimación de ruido basadas en VAD, y compartido por la mayoría de técnicas de estimación espectral, podemos comentar la dificultad de la estimación del ruido convolutivo a partir de los segmentos de silencio. En el siguiente apartado se verá cómo las estrategias de encuadre estadístico permiten estimar este ruido, junto con el ruido aditivo, siguiendo un marco matemático formal.

Siguiendo con el estudio de las técnicas de estimación de ruido, nos encontramos con aquellas que estiman la densidad de potencia espectral del ruido llevando una traza de la evolución temporal de los valores mínimos de energía. En este apartado describiremos cuatro de las técnicas más conocidas: MS (*Minimum Statistics noise estimation*, estimación de ruido basada en estadística de mínimos) [190], MCRA (*Minima Controlled Recursive Averaging*, promediado recursivo controlado de los valores mínimos) [54] y su versión mejorada IMCRA (*Improved* MCRA, MCRA mejorado) [53] y, por último, ANF (*Adaptive Noise Flooring*, umbralización adaptativa del ruido) [185].

En [190] Martin propone la técnica MS de estimación de la densidad de potencia espectral del ruido (aditivo). Dada su elevada eficiencia y la buena precisión que consigue a la hora de estimar tanto ruidos estacionarios como impredecibles, esta técnica se ha empleado en combinación con multitud de técnicas de procesamiento robusto de voz como realce de voz [182, 190], técnicas de compensación de características [264], adaptación de modelos [17], etc. Las principales cualidades de esta técnica respecto a las estudiadas hasta ahora son, por un lado, el no requerir de un VAD para distinguir los segmentos de voz de los que sólo contienen ruido y, por otro, el poder estimar la potencia del ruido incluso en los segmentos de voz. Para estimar el ruido, esta técnica se fundamenta en las siguientes observaciones: (i) en la mayoría de los casos voz y ruido pueden considerarse estadísticamente independientes y (ii) la potencia de la señal ruidosa decae con frecuencia a niveles que son representativos del ruido subyacente, ocurriendo esto incluso durante los periodos de actividad de voz.

La primera observación es clara y, de hecho, es una suposición básica en la mayoría de las técnicas de procesamiento robusto de la voz. Respecto a la segunda observación,

hemos comentado anteriormente que es posible estimar la potencia del ruido fielmente en los valles del espectro de los segmentos sonoros de voz [195]. Basándose en estas dos observaciones, la técnica MS sigue los valores mínimos en cada frecuencia del espectro de una versión suavizada de la señal de voz ruidosa sobre una ventana deslizante de 1 segundo aproximadamente. Estos valores mínimos son los que posteriormente, tras una corrección aditiva que depende de la varianza de la señal ruidosa suavizada, son los que se asocian al espectro del ruido. Para el cálculo de la versión suavizada de la señal ruidosa, en [190] se deriva la expresión para el factor de suavizado óptimo que, por un lado, permita una estimación fiable de las características de los ruidos cuasi-estacionarios y, por otro, sea capaz de seguir los ruidos no estacionarios.

Como podemos comprobar, la idea básica de la técnica MS consiste en seguir la evolución temporal de los valores mínimos del espectro, ya que estos son indicativos de la energía del ruido aditivo presente. Esta misma idea se ha incorporado en el módulo de supresión de ruido del método de extracción de características PNCC [157] descrito en la sección 2.2.1. De forma similar, la técnica MCRA [54] y su versión mejorada IMCRA [53] también realizan un seguimiento sobre los valores mínimos del espectro de la señal de voz ruidosa. La mayor diferencia de IMCRA respecto a MS radica en la elección más cuidadosa del factor de suavizado. Asimismo, IMCRA lleva a cabo un suavizado tanto en tiempo como en frecuencia, implementado además un doble bucle de suavizado y seguimiento del valor mínimo, lo que le permite un mayor control sobre el valor de ruido estimado. Todas estas características adicionales hacen que IMCRA muestre un mejor rendimiento que MS a la hora de estimar distintos tipos de ruidos [53].

La técnica ANF propuesta por Ma en [185] tiene como objetivo la estimación del espectro de aquellos ruidos que varían lentamente en el tiempo. Así, la técnica es capaz de estimar con precisión el espectro de las fuentes sonoras estacionaras presentes en una grabación (p.ej. ruido de un ventilador, motor de un coche, etc.). En el caso de ruidos impredecibles, esta técnica se combina en [185] con la técnica SFD estudiada en la sección 2.2.5.3. El funcionamiento de la técnica ANF puede resumirse de la siguiente forma: las componentes del ruido se estiman a partir de una ventana deslizante (de aproximadamente 5 segundos) que contiene tramas de voz expresada en el dominio log-Mel o un dominio equivalente (en [185] se utilizan características extraídas según la escala gammatone y comprimidas logarítmicamente). Posteriormente se aplica el algoritmo EM [66] para ajustar un modelo de mezclas de gaussianas (generalmente con 2 ó 3 componentes) a los datos de la ventana. Dada la gran dimensión de los vectores de datos, antes de estimar el GMM se aplica un método de selección de características (p.ej. se escogen ciertas componentes del vector) o de reducción de la dimensión (p.ej.

PCA) a las tramas de la ventana, de forma que la estimación de los parámetros del GMM sea robusta. Finalmente, se selecciona como modelo de ruido para las tramas de voz presentes en la ventana la componente del GMM cuya media tiene menor energía. De nuevo vemos que esta técnica supone que los valores de energía mínimos en el espectro son indicativos del ruido aditivo presente en el entorno.

2.3.2. Estimación estadística del modelo de ruido

Una estrategia alternativa para la estimación del modelo de ruido consiste en suponer que $\boldsymbol{n}$ y $\boldsymbol{h}$, esto es, las variables aleatorias que definen el ruido aditivo y convolutivo, respectivamente, son variables ocultas. Bajo esta suposición es posible aplicar el algoritmo EM para encontrar los parámetros del modelo de ruido $\mathcal{M}_n = \{\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n, \boldsymbol{\mu}_h\}$ que maximizan la probabilidad de la voz ruidosa $\boldsymbol{Y}$:

$$\hat{\mathcal{M}}_n = \underset{\mathcal{M}_n}{\operatorname{argmax}}\, p(\boldsymbol{Y}|\mathcal{M}_x, \mathcal{M}_n), \tag{2.192}$$

donde $\mathcal{M}_x$ es el modelo de voz limpia.

Como se puede apreciar, varias son las ventajas de este enfoque respecto a las técnicas estudiadas en la sección anterior:

- En primer lugar, este enfoque permite una estimación conjunta de ambos tipos de ruido, aditivo y convolutivo, bajo un enfoque completamente estadístico.
- Ahora es posible estimar el ruido de forma simple incluso en los segmentos de voz.
- El algoritmo de estimación se puede extender para que también estime las características dinámicas del ruido.
- Por último, dado que los parámetros del modelo de ruido se estiman siguiendo un criterio ML, ahora no resulta crítico si estos no describen fielmente el verdadero ruido subyacente. Aunque pueda parecer lo contrario, esto supone una ventaja, ya que de esta forma los parámetros del modelo de ruido pueden compensar, en cierta medida, cualquier aproximación introducida en el modelo de distorsión de la voz empleado durante la estimación de los mismos (p.ej. aproximación VTS).

Dada la dificultad de estimar los parámetros del modelo de ruido de forma directa, en la mayoría de los trabajos propuestos en la literatura se recurre a una aproximación iterativa basada en el algoritmo EM [66], si bien otros trabajos han recurrido a un enfoque basado en el algoritmo de descenso en gradiente [154, 175, 176].

En este apartado introduciremos los pasos básicos para estimar el modelo de ruido usando el algoritmo EM y la aproximación VTS. Este algoritmo se utilizará como base en la sección 4.3.3 para derivar un algoritmo EM para la estimación del modelo de ruido, pero en este caso se utilizará el modelo de enmascaramiento de la voz (ver sección 4.1) en vez de la aproximación VTS. En el caso de la aproximación basada en VTS, la función auxiliar empleada por el algoritmo EM es

$$\begin{aligned}\mathcal{Q}(\mathcal{M}_n, \hat{\mathcal{M}}_n) &= \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_t^{(k)} \log p(\boldsymbol{y}_t|k; \mathcal{M}_x, \hat{\mathcal{M}}_n) \\ &= \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_t^{(k)} \left\{ -\frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}_y^{(k)}| - \frac{1}{2}(\boldsymbol{y}_t - \hat{\boldsymbol{\mu}}_y^{(k)})^\top \hat{\boldsymbol{\Sigma}}_y^{(k)^{-1}} (\boldsymbol{y}_t - \hat{\boldsymbol{\mu}}_y^{(k)}) \right\}, \quad (2.193)\end{aligned}$$

donde $\mathcal{M}_n$ y $\hat{\mathcal{M}}_n$ denotan la hipótesis actual del modelo de ruido y la versión actualizada que se estima en cada iteración, respectivamente. Por otra parte, $\gamma_t^{(k)}$ es la probabilidad a posteriori de la componente k-ésima del modelo de voz limpia (gaussiana k en caso de emplear un GMM) en el instante de tiempo t. Esta probabilidad se calcula utilizando $\mathcal{M}_n$ y el algoritmo *forward-backward* [220].

Para el cálculo de los parámetros del modelo de voz ruidosa, $(\hat{\boldsymbol{\mu}}_y^{(k)}, \hat{\boldsymbol{\Sigma}}_y^{(k)}); k = 1, \ldots, K$; presentes en la ecuación (2.193) a partir de los modelos de voz $\mathcal{M}_x$ y de ruido $\mathcal{M}_n$, es necesario recurrir a ciertas aproximaciones que simplifiquen la no linealidad del modelo de distorsión de la voz estudiado en la sección 2.1.1. En la mayoría de las técnicas propuestas en la literatura [70, 94, 159, 166, 171, 199], esta no linealidad se resuelve aproximando el modelo de distorsión mediante la técnica VTS estudiada en la sección 2.2.2.2. En este caso, el valor de los parámetros a estimar se obtiene derivando la función auxiliar de la ecuación (2.193) y resolviendo el sistema de ecuaciones resultante obtenido tras igualar a cero. Suponiendo constantes tanto las matrices jacobianas obtenidas en la formulación VTS como la matriz de covarianza del ruido aditivo $\boldsymbol{\Sigma}_n$, la derivada de la función auxiliar (2.193) respecto al vector media del ruido aditivo viene dada por [175],

$$\begin{aligned}&\frac{\partial}{\partial \hat{\boldsymbol{\mu}}_n} \mathcal{Q}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\mu}}_h; \boldsymbol{\mu}_n, \boldsymbol{\mu}_h) \\ &\quad = \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_t^{(k)} \frac{\partial}{\partial \hat{\boldsymbol{\mu}}_n} \left\{ -\frac{1}{2}\log|\boldsymbol{\Sigma}_y^{(k)}| - \frac{1}{2}(\boldsymbol{y}_t - \hat{\boldsymbol{\mu}}_y^{(k)})^\top \boldsymbol{\Sigma}_y^{(k)^{-1}} (\boldsymbol{y}_t - \hat{\boldsymbol{\mu}}_y^{(k)}) \right\} \\ &\quad = \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_t^{(k)} \left\{ 0 - \frac{1}{2}\frac{\partial}{\partial \hat{\boldsymbol{\mu}}_n}(\boldsymbol{y}_t - \hat{\boldsymbol{\mu}}_y^{(k)})^\top \boldsymbol{\Sigma}_y^{(k)^{-1}} (\boldsymbol{y}_t - \hat{\boldsymbol{\mu}}_y^{(k)}) \right\}\end{aligned}$$

$$= \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_t^{(k)} \boldsymbol{J}_n^{(k)^\top} \boldsymbol{\Sigma}_y^{(k)^{-1}} \left(\boldsymbol{y}_t - \boldsymbol{\mu}_y^{(k)} + \boldsymbol{J}_n^{(k)} \boldsymbol{\mu}_n + \boldsymbol{J}_h^{(k)} \boldsymbol{\mu}_h - \boldsymbol{J}_n^{(k)} \hat{\boldsymbol{\mu}}_n - \boldsymbol{J}_h^{(k)} \hat{\boldsymbol{\mu}}_h \right)$$
$$= \boldsymbol{d} - \boldsymbol{E}\hat{\boldsymbol{\mu}}_n - \boldsymbol{F}\hat{\boldsymbol{\mu}}_h, \tag{2.194}$$

siendo

$$\boldsymbol{d} = \sum_{k=1}^{K} \boldsymbol{J}_n^{(k)^\top} \boldsymbol{\Sigma}_y^{(k)^{-1}} \sum_{t=1}^{T} \gamma_t^{(k)} \left(\boldsymbol{y}_t - \boldsymbol{\mu}_y^{(k)} + \boldsymbol{J}_n^{(k)} \boldsymbol{\mu}_n + \boldsymbol{J}_h^{(k)} \boldsymbol{\mu}_h \right), \tag{2.195}$$

$$\boldsymbol{E} = \sum_{k=1}^{K} \gamma^{(k)} \boldsymbol{J}_n^{(k)^\top} \boldsymbol{\Sigma}_y^{(k)^{-1}} \boldsymbol{J}_n^{(k)}, \tag{2.196}$$

$$\boldsymbol{F} = \sum_{k=1}^{K} \gamma^{(k)} \boldsymbol{J}_n^{(k)^\top} \boldsymbol{\Sigma}_y^{(k)^{-1}} \boldsymbol{J}_h^{(k)} \tag{2.197}$$

y $\gamma^{(k)} = \sum_{t=1}^{T} \gamma_t^{(k)}$.

Análogamente, la derivada de la función auxiliar (2.193) respecto al vector de ruido convolutivo se calcula como

$$\frac{\partial}{\partial \hat{\boldsymbol{\mu}}_h} \mathcal{Q}(\hat{\boldsymbol{\mu}}_n, \hat{\boldsymbol{\mu}}_h; \boldsymbol{\mu}_n, \boldsymbol{\mu}_h)$$
$$= \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_t^{(k)} \frac{\partial}{\partial \hat{\boldsymbol{\mu}}_h} \left\{ -\frac{1}{2} \log |\boldsymbol{\Sigma}_y^{(k)}| - \frac{1}{2} (\boldsymbol{y}_t - \hat{\boldsymbol{\mu}}_y^{(k)})^\top \boldsymbol{\Sigma}_y^{(k)^{-1}} (\boldsymbol{y}_t - \hat{\boldsymbol{\mu}}_y^{(k)}) \right\}$$
$$= \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_t^{(k)} \left\{ 0 - \frac{1}{2} \frac{\partial}{\partial \hat{\boldsymbol{\mu}}_h} (\boldsymbol{y}_t - \hat{\boldsymbol{\mu}}_y^{(k)})^\top \boldsymbol{\Sigma}_y^{(k)^{-1}} (\boldsymbol{y}_t - \hat{\boldsymbol{\mu}}_y^{(k)}) \right\}$$
$$= \sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_t^{(k)} \boldsymbol{J}_h^{(k)^\top} \boldsymbol{\Sigma}_y^{(k)^{-1}} \left(\boldsymbol{y}_t - \boldsymbol{\mu}_y^{(k)} + \boldsymbol{J}_n^{(k)} \boldsymbol{\mu}_n + \boldsymbol{J}_h^{(k)} \boldsymbol{\mu}_h - \boldsymbol{J}_n^{(k)} \hat{\boldsymbol{\mu}}_n - \boldsymbol{J}_h^{(k)} \hat{\boldsymbol{\mu}}_h \right)$$
$$= \boldsymbol{u} - \boldsymbol{V}\hat{\boldsymbol{\mu}}_n - \boldsymbol{W}\hat{\boldsymbol{\mu}}_h \tag{2.198}$$

y

$$\boldsymbol{u} = \sum_{k=1}^{K} \boldsymbol{J}_h^{(k)^\top} \boldsymbol{\Sigma}_y^{(k)^{-1}} \sum_{t=1}^{T} \gamma_t^{(k)} \left(\boldsymbol{y}_t - \boldsymbol{\mu}_y^{(k)} + \boldsymbol{J}_n^{(k)} \boldsymbol{\mu}_n + \boldsymbol{J}_h^{(k)} \boldsymbol{\mu}_h \right), \tag{2.199}$$

$$\boldsymbol{V} = \sum_{k=1}^{K} \gamma^{(k)} \boldsymbol{J}_h^{(k)^\top} \boldsymbol{\Sigma}_y^{(k)^{-1}} \boldsymbol{J}_n^{(k)}, \tag{2.200}$$

$$\boldsymbol{W} = \sum_{k=1}^{K} \gamma^{(k)} \boldsymbol{J}_h^{(k)^\top} \boldsymbol{\Sigma}_y^{(k)^{-1}} \boldsymbol{J}_h^{(k)}. \tag{2.201}$$

Igualando a cero las ecuaciones (2.194) y (2.198) obtenemos el siguiente sistema de ecuaciones,

$$\begin{bmatrix} \boldsymbol{E} & \boldsymbol{F} \\ \boldsymbol{V} & \boldsymbol{W} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\mu}}_n \\ \hat{\boldsymbol{\mu}}_h \end{bmatrix} = \begin{bmatrix} \boldsymbol{d} \\ \boldsymbol{u} \end{bmatrix}. \tag{2.202}$$

Resolviendo el sistema de ecuaciones anterior, obtenemos las siguientes soluciones para las estimas del modelo de ruido,

$$\hat{\boldsymbol{\mu}}_n = (\boldsymbol{E} - \boldsymbol{F}\boldsymbol{W}^{-1}\boldsymbol{V})^{-1}(\boldsymbol{d} - \boldsymbol{F}\boldsymbol{W}^{-1}\boldsymbol{u}), \tag{2.203}$$

$$\hat{\boldsymbol{\mu}}_h = (\boldsymbol{W} - \boldsymbol{V}\boldsymbol{E}^{-1}\boldsymbol{F})^{-1}(\boldsymbol{u} - \boldsymbol{V}\boldsymbol{E}^{-1}\boldsymbol{d}). \tag{2.204}$$

Las dos ecuaciones anteriores definen el método iterativo mediante el cual se pueden estimar las medias del modelo de ruido. Por contra, no existe una expresión analítica que permita al cálculo de la matriz de covarianza del ruido aditivo $\boldsymbol{\Sigma}_n$. Es por ello que en la literatura se haya recurrido a métodos basados en el algoritmo de descenso en gradiente para estimar esta matriz [176].

Tal y como se ha visto, para derivar las expresiones del algoritmo EM e ha empleado la aproximación VTS para linealizar el modelo de distorsión de la voz. Aunque ésta ha sido la aproximación predominante, en la literatura también podemos encontrar otros enfoques alternativos. De entre estos enfoques, destacaremos aquí aquellos que aproximan numéricamente la distribución a posteriori del ruido. Por ejemplo, en [86, 91, 96] se propone el uso de los filtros de partículas para la estimación de los ruidos no estacionarios. En lineas generales, estos algoritmos generan un conjunto de muestras de ruido (partículas) que se ponderan según un criterio probabilístico en función de la observación ruidosa y un modelo de voz limpia (GMM) disponible a priori. Estas partículas se combinan entonces para obtener la estimación final de ruido. En [89], por su parte, se usa la denominada transformada *unscented* (UT, *unscented transform*) [151] en lugar de VTS para obtener una aproximación de la PDF conjunta de la voz ruidosa, el ruido aditivo y el ruido convolutivo. Esta PDF se emplea dentro del algoritmo EM para obtener las estimas oportunas del modelo de ruido.

2.4. Resumen

Esta capítulo ha tratado sobre la problemática que se genera en los sistemas de RAH cuando trabajan sobre voz distorsionada, así como las posibles soluciones que se han propuesto a este problema. En primer lugar, en la sección 2.1 se presentó un expresión analítica que modela el efecto de las dos principales fuentes de ruido, aditivo y convolutivo, en la señal de voz emitida por el locutor, haciendo especial hincapié en la distorsión producida en las características log-Mel y MFCC. En base a este modelo se dedujo que el principal efecto del ruido sobre la voz es, desde el punto de vista estadístico, una modificación de las distribuciones de probabilidad con respecto a las de la voz limpia, produciendo ello una discrepancia entre las condiciones de entrenamiento y las de evaluación.

A fin de reducir la discrepancia producida por el ruido, en la sección 2.2 se revisaron las principales propuestas para robustecer los sistemas de RAH frente al ruido. A grandes rasgos, se vio que estas propuestas se agrupan en tres grandes grupos: (i) extracción robusta de características (sección 2.2.1), (ii) adaptación de los modelos del reconocedor (sección 2.2.2) y (iii) modificación de las características con las que se reconocen (sección 2.2.3).

Las extracción robusta de características se encarga de computar una serie de características de voz que sean más insensibles a la acción del ruido. Generalmente, tal y como se estudió en la sección 2.2.1, esto implica trabajar en dominios más robustos al ruido, como es el caso de la autocorrelación. No obstante, su capacidad para robustecer los sistemas de RAH es limitada ya que, entre otras cosas, hacen pocas suposiciones sobre las características del ruido.

Las técnicas de adaptación de modelos modifican los parámetros del modelo acústico del reconocedor para que representen fielmente las condiciones en las que el sistema se explota. Dentro de adaptación de modelos, en la sección 2.2.2 distinguíamos entre dos grupos de técnicas: las de adaptación estadística (sección 2.2.2.1) y la adaptación basada en modelos de distorsión (2.2.2.2). En la adaptación estadística, los modelos se modifican en base a estadísticas recopiladas usando conjuntos de adaptación con elocuciones grabadas en el entorno de explotación del sistema (p.ej. voz de un locutor determinado o elocuciones contaminadas con un ruido dado). Generalmente las técnicas que se encuadran en este grupo son técnicas genéricas que no fueron diseñadas específicamente para tratar el problema del ruido. Las técnicas más conocidas de este grupo son MLLR y MAP. En segundo lugar están las técnicas de adaptación basadas en modelos de distorsión descritas en la sección 2.2.2.2. Al contrario que las técnicas del grupo anterior, la adaptación basada en modelos de distorsión sí que aborda específicamente el problema de la adaptación del reconocedor al ruido. Para ello se derivan un conjunto de transformaciones que se aplican a los parámetros del modelo acústico a fin de reducir la discrepancia al reconocer voz ruidosa. Estas transformaciones se derivan mediante distintas aproximaciones del modelo analítico de distorsión (p.ej. aproximaciones PMC, VTS y Algonquin).

La tercera alternativa para robustecer el RAH frente al ruido consiste en procesar las características extraídas de la voz ruidosa antes de usarlas para reconocer. Como se vio en la sección 2.2.3, las técnicas de este grupo modifican las características de las elocuciones que se reconocen para que sean más parecidas a las usadas durante la fase de entrenamiento. Esta tarea puede realizarse de dos formas alternativas: normalizando las características o compensándolas para reducir la distorsión producida por el ruido. Las técnicas de normalización de características (sección 2.2.3.1) son técnicas que se aplican

tanto en la fase de entrenamiento como en la de evaluación. Persiguen la transformación de las características de voz a un dominio donde les afecte menos la distorsión del ruido. Técnicas dentro de este grupo muy conocidas son CMN, MVN y HEQ. Por otro lado, las técnicas de compensación o realce (secciones 2.2.3.2, 2.2.3.3 y 2.2.3.4) persiguen mitigar el ruido en las características extraídas, de forma que al final se obtenga una señal pseudolimpia más parecida a las empleadas durante el entrenamiento del sistema.

Además de las tres aproximaciones anteriores para el reconocimiento robusto de voz, en este capítulo se han analizado otras dos aproximaciones híbridas: el reconocimiento con incertidumbre (sección 2.2.4) y las técnicas MD derivadas del paradigma de datos perdidos (sección 2.2.5).

El reconocimiento con incertidumbre se trata de una estrategia que combina las bondades de las técnicas de compensación de características y la adaptación de modelos. En primer lugar se procede a estimar las características relativas a la voz limpia y, además, se computan medidas de fiabilidad de la estimación realizada. Posteriormente, las estimas de voz y las medidas de fiabilidad son usadas por un algoritmo de reconocimiento modificado que otorga más peso a las estimaciones más fiables, controlando de esta forma el efecto en el reconocimiento de estimaciones muy poco fiables.

Por otra parte, las técnicas MD se basan en un modelo de distorsión simplificado que considera el efecto del ruido aditivo sobre la voz como una pérdida de datos en el espectro de ésta. A fin de salvar esta pérdida de información, dos son las estrategias de reconocimiento robusto propuestas. La primera de ellas (sección 2.2.5.3) supone modificar el proceso de descodificación para que en el cómputo de las probabilidades de observación se diferencie entre elementos fiables del espectro (correspondientes a la voz) y elementos perdidos (correspondientes al ruido). La segunda estrategia MD considerada en la sección 2.2.5.4 es la estimación de los valores perdidos del espectro. Para llevar a cabo esta tarea, las técnicas de estimación (imputación) emplean máscaras que segregan el espectro observado en elementos fiables y perdidos, y modelos de voz que embeben las correlaciones entre las distintas características.

Concluyendo este capítulo, en la sección 2.3 se presentaron distintas estrategias para estimar los modelos de ruido requeridos por la mayoría de técnicas de reconocimiento robusto estudiadas.

CAPÍTULO 3

Compensación basada en datos estéreo

En el capítulo 2 hacíamos una revisión de las distintas aproximaciones al reconocimiento robusto que se han propuesto en la literatura. Veíamos que un grupo amplio de estas aproximaciones lo constituyen las técnicas de compensación de los parámetros extraídos de la señal de voz. Estas técnicas tienen como objeto mitigar, en la medida de lo posible, la degradación producida por el ruido en dichos parámetros. De entre las técnicas de compensación estudiadas, también veíamos que existe un subconjunto de las mismas que emplean grabaciones estéreo para estimar las transformaciones que se aplican a los parámetros de la voz para compensar la distorsión producida por el ruido.

En este capítulo seguimos esta misma filosofía y proponemos un conjunto de técnicas de compensación basadas en datos estéreo. Al estimar las transformaciones de compensación usando grabaciones estéreo, no hay necesidad de emplear un modelo de distorsión analítico como lo hacen las técnicas estudiadas en la sección 2.2.3.3. La ventaja, por tanto, de nuestra propuesta respecto a estas últimas técnicas es clara: las técnicas que proponemos pueden abordar, en teoría, cualquier tipo de discrepancia no deseada en los parámetros de la voz, sea producida por el ruido (aditivo, de canal o reverberación) o fruto de las diferencias entre locutores, al no emplear ningún modelo que restrinja el tipo de distorsiones que se pueden combatir. Es más, el no usar un modelo de distorsión conlleva no hacer suposiciones sobre las características del ruido a combatir ni la forma en la que éste afecta a los parámetros de voz. Estos dos aspectos, particularmente el primero de ellos, pueden considerarse como los puntos más frágiles de las técnicas de compensación basadas en modelos de distorsión, en el sentido que,

por ejemplo, una mala estimación de la potencia del ruido puede resultar desastrosa para el proceso de compensación de las características de la voz.

Además del uso de datos estéreo, las técnicas que proponemos emplean diccionarios de cuantificación vectorial (VQ, *Vector Quantization*) durante el modelado de los espacios de características de la voz limpio y distorsionado. Fruto de ello, es posible obtener técnicas de compensación muy eficientes. Los resultados experimentales que presentaremos en el capítulo 6 mostrarán el buen rendimiento alcanzado por nuestra propuesta, siendo éste comparable, y superando en ciertas situaciones, al rendimiento mostrado por otras técnicas de compensación más complejas desde el punto de vista computacional.

Viendo con perspectiva las técnicas de compensación que introduciremos a lo largo de este capítulo y el siguiente, observamos que uno de los nexos de unión que comparten dichas técnicas se materializa en la forma de estimar los parámetros de voz limpia. Así, nos percatamos de que todas estas técnicas adoptan el criterio de mínimo error cuadrático medio (MMSE, *Minimum Mean Square Error*) para el cómputo de esas estimas. Asimismo, durante el cálculo de las probabilidades requeridas por el estimador, se asume que el espacio de características de la voz puede modelarse de forma precisa mediante un diccionario que contiene un conjunto de celdas o regiones que dividen dicho espacio. Como ya veremos, la partición podrá ser disjunta, conduciendo a diccionarios VQ, o bien podrá permitir cierto solapamiento entre las distintas regiones, conduciendo a modelos probabilísticos de mezcla de PDFs (p.ej. GMM). Con objeto de unificar todas estas técnicas bajo un marco matemático común que permita una comparación fácil entre las mismas, en la sección siguiente procederemos a introducir las ecuaciones básicas requeridas por los estimadores MMSE basados en diccionario.

3.1. Estimación MMSE basada en diccionario

La compensación de características de voz puede formularse como el siguiente problema: dada la secuencia observada de vectores de características $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$, queremos obtener la secuencia $\hat{\boldsymbol{X}} = (\hat{\boldsymbol{x}}_1, \ldots, \hat{\boldsymbol{x}}_T)$ tal que $\hat{\boldsymbol{X}}$ maximice, en comparación con $\boldsymbol{Y}$, algún criterio de optimización respecto a la secuencia no observada de vectores de voz limpia $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$. Criterios comunes que se suelen elegir son el de menor error cuadrático $\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_F \leq \|\boldsymbol{Y} - \boldsymbol{X}\|_F$ [1] o mayor tasa de reconocimiento de palabras obtenida con $\hat{\boldsymbol{X}}$ frente a $\boldsymbol{Y}$. Para simplificar la tarea de compensación, supondremos inicialmente que los vectores de características son estadísticamente independientes entre sí, por lo que el problema se traduce ahora en encontrar el vector $\hat{\boldsymbol{x}}$ que satisface

[1] $\|\cdot\|_F$ denota la norma de Frobenius.

el criterio de optimización adoptado respecto a $\boldsymbol{x}$ en comparación con la observación $\boldsymbol{y}$ (se ha omitido el índice temporal por claridad). De aquí en adelante al vector $\hat{\boldsymbol{x}}$ lo denominaremos como la estimación del vector de voz limpia no observado $\boldsymbol{x}$.

El primer paso para abordar el diseño del estimador es conocer la información de la que disponemos y que puede ser relevante para obtener el valor de $\hat{\boldsymbol{x}}$. Por supuesto conocemos el valor de la observación $\boldsymbol{y}$, la cual, en función de la SNR de la señal, puede ser más o menos próxima a $\boldsymbol{x}$. También podemos disponer a priori de cierta información lateral Λ que nos ayude en el proceso de estimación de las características de la voz. Esta información puede ser, por ejemplo, una estimación de la potencia del ruido, de la SNR instantánea de la señal o de la dirección de la que proviene la voz del locutor. La información que hemos descrito hasta ahora, $\boldsymbol{y}$ y Λ, es la que se ha venido utilizando tradicionalmente en las técnicas de realce de la voz [37]. Además de esta información, y debiéndose en gran parte al trabajo pionero desarrollado por Ephraim y Malah en el campo de las técnicas de realce de encuadre estadístico [81, 82, 83], en repetidas ocasiones se ha demostrado que la inclusión de información a priori sobre la voz en el estimador produce estimaciones que son consistentemente mejores que las obtenidas cuando esta información no se contempla. Como veremos a continuación, esta información se suele materializar en modelos estadísticos de la voz.

En esta trabajo supondremos que la señal de voz, en cualquiera de sus dominios (p.ej. log-Mel o MFCC), puede representarse mediante un modelo de mezcla de PDFs $\mathcal{M}_x$:

$$\mathcal{M}_x = \{\mathcal{C}_x^{(1)}, \mathcal{C}_x^{(2)}, \ldots, \mathcal{C}_x^{(M)}\}, \tag{3.1}$$

donde $\mathcal{C}_x^{(k)} (k = 1, \ldots, M)$ caracterizan las distintas regiones o celdas[1] en las que se divide el espacio de características. Por ejemplo si el modelo $\mathcal{M}_x$ se trata de un diccionario VQ, cada una de sus celdas estará representada por un par de la forma $\mathcal{C}_x^{(k)} = \left\langle \boldsymbol{\mu}_x^{(k)}, \boldsymbol{\Sigma}_x^{(k)} \right\rangle$, siendo $\boldsymbol{\mu}_x^{(k)}$ el centroide de la celda y $\boldsymbol{\Sigma}_x^{(k)}$ su matriz de covarianza. Para GMMs tendremos que cada celda se representará mediante una tripleta $\mathcal{C}_x^{(k)} = \left\langle \pi_x^{(k)}, \boldsymbol{\mu}_x^{(k)}, \boldsymbol{\Sigma}_x^{(k)} \right\rangle$ donde $\boldsymbol{\mu}_x^{(k)}$ y $\boldsymbol{\Sigma}_x^{(k)}$ son los parámetros de la gaussiana en cuestión y $\pi_x^{(k)} \equiv P(k|\mathcal{M}_x)$ denota su probabilidad a priori (en el caso de los diccionarios VQ consideramos a las celdas equiprobables). Por generalidad de la exposición, y salvo que se indique lo contrario, en la explicación que sigue no particularizaremos $\mathcal{M}_x$ a ningún modelo en concreto, refiriéndonos a éste como el diccionario de la voz o, simplemente, el modelo de la voz.

El segundo paso en el diseño del estimador consiste en elegir el criterio de optimización que se va seguir para calcular $\hat{\boldsymbol{x}}$. Antes hemos nombrado un par de criterios

[1] En esta tesis utilizaremos indistintamente el término región o celda para referirnos a cada una de las partes en las que se divide el espacio de características.

que son susceptibles de ser empleados, aunque dado que vamos a emplear información a priori sobre la voz durante la estimación, según la teoría bayesiana existen ciertos criterios (estimadores) más apropiados para estos casos. En particular, dos son los estimadores bayesianos más utilizados en la práctica [156]: el de máxima probabilidad a posteriori, MAP, y el de mínimo error cuadrático medio, MMSE. Aunque en general la derivación del estimador MMSE suele ser más compleja que la del estimador MAP, en esta tesis preferiremos el primero de ellos por las siguientes razones:

- En la literatura podemos encontrar numerosos trabajos en los que estos estimadores son evaluados en el contexto del realce de la voz y/o la compensación de las características de la misma. En estos trabajos podemos constatar que el estimador MMSE exhibe un rendimiento que es consistentemente mayor que el alcanzado por el estimador MAP (ver [71] para más detalles).

- En segundo lugar podemos ver que, aunque ambos sean estimadores puntuales, el valor devuelto por el estimador MMSE, esto es, la media de la distribución a posteriori, es más representativo de esta distribución que el devuelto por la estimación MAP (la moda de la distribución). En este sentido podemos decir que la media de esta distribución es un estadístico más informativo de la misma que la moda.

- Como última razón tenemos que, aunque definido sobre el criterio MSE (error cuadrático medio), el estimador MMSE puede extenderse fácilmente para que abarque de forma óptima otras medidas de distorsión. Éste es un atributo muy atractivo que no se da en el estimador MAP y que es deseable puesto que permite incluir medidas de distorsión con consideraciones perceptivas en el estimador MMSE [71].

Después de la introducción anterior, ya tenemos todas las herramientas necesarias para presentar el estimador MMSE que emplearemos. Dado el vector observado $\boldsymbol{y}$, la estimación MMSE $\hat{\boldsymbol{x}}$ del vector aleatorio $\boldsymbol{x}$ es aquella que minimiza el error cuadrático medio:

$$\hat{\boldsymbol{x}} = \underset{\tilde{\boldsymbol{x}}}{\operatorname{argmin}}\, \mathbb{E}[\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|]. \tag{3.2}$$

De acuerdo a uno de los teoremas fundamentales de la teoría de la estimación, el valor $\hat{\boldsymbol{x}}$ que minimiza el error de la ecuación anterior se corresponde con el siguiente valor esperado [156]

$$\hat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}] = \int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x}, \tag{3.3}$$

esto es, la estimación MMSE $\hat{\boldsymbol{x}}$ coincide con el valor medio de la distribución a posteriori $p(\boldsymbol{x}|\boldsymbol{y})$.

El primer paso, y quizás el más complejo, a la hora de derivar el estimador de la ecuación (3.3) consiste en modelar $p(\boldsymbol{x}|\boldsymbol{y})$. Dicho modelo puede plasmarse en una expresión analítica de esta probabilidad o bien puede tratarse de un modelo numérico obtenido experimentalmente. En nuestro caso nos apoyaremos en el diccionario $\mathcal{M}_x$ de la ecuación (3.1) para obtener estas probabilidades. Dado que $\mathcal{M}_x$ es una mezcla de PDFs, la probabilidad del vector de voz limpia puede obtenerse de la siguiente forma:

$$p(\boldsymbol{x}|\mathcal{M}_x) = \sum_{k=1}^{M} p(\boldsymbol{x}|k, \mathcal{M}_x) P(k|\mathcal{M}_x), \tag{3.4}$$

donde se ha notado por $p(\boldsymbol{x}|k, \mathcal{M}_x)$ a la probabilidad de $\boldsymbol{x}$ en la celda k-ésima del modelo, esto es, $p(\boldsymbol{x}|\mathcal{C}_x^{(k)})$. Para simplificar aún más la notación, siempre que sea posible y no induzca a confusión omitiremos la referencia al modelo $\mathcal{M}_x$ en las probabilidades anteriores, luego $p(\boldsymbol{x}|k) \equiv p(\boldsymbol{x}|k, \mathcal{M}_x)$ y $P(k) \equiv P(k|\mathcal{M}_x)$.

Para obtener la probabilidad a posteriori que aparece en el estimador de la ecuación (3.3), la PDF conjunta $p(\boldsymbol{x}, k|\boldsymbol{y})$ se marginaliza sobre la variable oculta k. Asimismo, introducimos el termino Λ para referirnos a cualquier información a priori que pueda resultar de utilidad de cara a la estimación de $\hat{\boldsymbol{x}}$. En este caso tenemos que

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y}, \Lambda) &= \sum_{k=1}^{M} p(\boldsymbol{x}, k|\boldsymbol{y}, \Lambda) \\ &= \sum_{k=1}^{M} p(\boldsymbol{x}|\boldsymbol{y}, k, \Lambda) P(k|\boldsymbol{y}, \Lambda). \end{aligned} \tag{3.5}$$

Usando la ecuación (3.5) en (3.3) se obtiene la siguiente expresión final para la estimación MMSE del vector de voz limpia,

$$\begin{aligned} \hat{\boldsymbol{x}} &= \int \boldsymbol{x} \sum_{k=1}^{M} p(\boldsymbol{x}|\boldsymbol{y}, k, \Lambda) P(k|\boldsymbol{y}, \Lambda) \\ &= \sum_{k=1}^{M} P(k|\boldsymbol{y}, \Lambda) \underbrace{\int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}, k, \Lambda) d\boldsymbol{x}}_{\hat{\boldsymbol{x}}^{(k)}}. \end{aligned} \tag{3.6}$$

Como puede observarse, el estimador consiste en una combinación lineal de ciertas estimaciones parciales $\hat{\boldsymbol{x}}^{(k)}$, una para cada celda $k = 1, \ldots, M$ del diccionario $\mathcal{M}_x$, ponderadas por sus correspondientes probabilidades a posteriori $P(k|\boldsymbol{y}, \Lambda)$. En términos

generales estas probabilidades se pueden obtener recurriendo a la regla de Bayes:

$$\begin{aligned} P(k|\boldsymbol{y},\Lambda) &= \frac{p(\boldsymbol{y}|k,\Lambda)P(k|\Lambda)}{p(\boldsymbol{y}|\Lambda)} \\ &= \frac{p(\boldsymbol{y}|k,\Lambda)P(k)}{\sum_{k'=1}^{M} p(\boldsymbol{y}|k',\Lambda)P(k')} \end{aligned} \tag{3.7}$$

donde hemos supuesto que las probabilidades a priori de cada celda, $P(k)$, no se ven alteradas por la adición de ruido, lo cual es razonable salvo en aquellos casos en los que se den además ciertos efectos colaterales como por ejemplo el efecto Lombard [152, 153].

El término $p(\boldsymbol{y}|k,\Lambda)$ que aparece en la ecuación (3.7) se corresponde con la probabilidad de observación del vector $\boldsymbol{y}$ dada la celda k-ésima del modelo de voz limpia. Sin ninguna información adicional que nos indique el grado de distorsión que ha sufrido la voz, no tenemos forma de calcular esta probabilidad. Es por ello que en dicha probabilidad se recoge explícitamente el término Λ como información a priori. Esta información puede ser una estima de la potencia del ruido o de la SNR de la señal de voz. En su forma más general podemos considerar que Λ nos proporciona información estadística sobre el ruido presente en la señal observada, esto es, $\Lambda \equiv \mathcal{M}_n$, y, por tanto, el modelo $\mathcal{M}_n$ nos permite evaluar la probabilidad $p(\boldsymbol{n}|\mathcal{M}_n)$. Bajo este supuesto, la probabilidad de observación del vector de voz distorsionada viene dada por

$$\begin{aligned} p(\boldsymbol{y}|k,\mathcal{M}_n,\mathcal{M}_x) &= \int_{\boldsymbol{x}}\int_{\boldsymbol{n}} p(\boldsymbol{x},\boldsymbol{n},\boldsymbol{y}|k,\mathcal{M}_n,\mathcal{M}_x)d\boldsymbol{x}d\boldsymbol{n} \\ &= \int_{\boldsymbol{x}}\int_{\boldsymbol{n}} p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{n})p(\boldsymbol{x}|k,\mathcal{M}_x)p(\boldsymbol{n}|\mathcal{M}_n)d\boldsymbol{x}d\boldsymbol{n}, \end{aligned} \tag{3.8}$$

donde la segunda igualdad de la ecuación anterior se obtiene tras suponer, por un lado, independencia estadística entre voz y ruido y, por otro, que $\boldsymbol{y}$ es independiente de los modelos de voz y ruido dados $\boldsymbol{x}$ y $\boldsymbol{n}$.

De los términos que aparecen en la ecuación (3.8) el único desconocido es $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{n})$. Esta probabilidad nos define la relación entre los vectores $\boldsymbol{y}$, $\boldsymbol{n}$ y $\boldsymbol{x}$. Si disponemos de un modelo de distorsión analítico $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x},\boldsymbol{n})$ como el presentado en la sección 2.1, podemos considerar que $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{n}) = \delta_{\boldsymbol{f}(\boldsymbol{x},\boldsymbol{n})}(\boldsymbol{y})$. No obstante, rara vez se suele trabajar con el modelo de distorsión real, sino que en la mayoría de las ocasiones se recurren a aproximaciones lineales del mismo, $\boldsymbol{y} \approx \hat{\boldsymbol{f}}(\boldsymbol{x},\boldsymbol{n})$, tal y como la proporcionada por la técnica VTS estudiada en la sección 2.2.2.2. En estos casos es posible incorporar el error residual de la aproximación en el cómputo de la probabilidad $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{n})$ considerando, por ejemplo, una gaussiana con una matriz de covarianza $\boldsymbol{\Psi}$ función de dicho error, esto es, $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{n}) \approx \mathcal{N}(\boldsymbol{y};\hat{\boldsymbol{f}}(\boldsymbol{x},\boldsymbol{n}),\boldsymbol{\Psi})$.

Siguiendo un procedimiento similar al anterior, podemos derivar una expresión general para el cálculo de los valores $\hat{\boldsymbol{x}}^{(k)}$ presentes en la ecuación (3.6):

$$\begin{aligned}\hat{\boldsymbol{x}}^{(k)} &= \int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}, k, \Lambda) d\boldsymbol{x} \\ &\equiv \int_{\boldsymbol{x}} \int_{\boldsymbol{n}} \boldsymbol{x} p(\boldsymbol{x}, \boldsymbol{n}|\boldsymbol{y}, k, \mathcal{M}_n, \mathcal{M}_x) d\boldsymbol{x} d\boldsymbol{n} \\ &= \frac{1}{p(\boldsymbol{y}|k, \mathcal{M}_n, \mathcal{M}_x)} \int_{\boldsymbol{x}} \int_{\boldsymbol{n}} \boldsymbol{x} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n}) p(\boldsymbol{x}|k, \mathcal{M}_x) p(\boldsymbol{n}|\mathcal{M}_n) d\boldsymbol{x} d\boldsymbol{n} \end{aligned} \tag{3.9}$$

En la figura 3.1 se presenta un esquema gráfico de los distintos elementos involucrados en el estimador. Para simplificar el esquema gráfico, en esta figura se supone que la identidad de la celda k^* que con mayor probabilidad generó el vector limpio $\boldsymbol{x}$, $k^* = \text{argmax}_{\boldsymbol{x}} P(k|\boldsymbol{x})$, es conocida. Aunque esta suposición no es realista puesto que $\boldsymbol{x}$ es desconocido, nos ayudará a simplificar la descripción del estimador MMSE. Como veíamos en el capítulo 2, el ruido modifica la forma de la distribución $p(\boldsymbol{x}|k^*)$ dando lugar a la PDF $p(\boldsymbol{y}|k^*)$. Esta última PDF, en principio, es desconocida, así que para estimarla se introduce la información a priori Λ. Como se aprecia en la figura, la distorsión introducida por el ruido suele ser no lineal, lo que conlleva que, aunque la PDF limpia $p(\boldsymbol{x}|k^*)$ sea gaussiana, $p(\boldsymbol{y}|k^*, \Lambda)$ no tiene por qué serlo. En la práctica, por tanto, $p(\boldsymbol{y}|k^*, \Lambda)$ se suele aproximar por una forma más tratable (en VTS, por ejemplo, se aproxima por una gaussiana [199]).

Además de la probabilidad de observación anterior, el estimador MMSE debe calcular una estimación parcial $\hat{\boldsymbol{x}}^{(k)}$ para cada una de las celdas del espacio limpio. Esta estimación parcial se interpreta como sigue. Supongamos que el vector observado lo notamos como $\boldsymbol{y}_o$. Debido a la aleatoriedad inherente al ruido, existe un conjunto de vectores de voz limpia pertenecientes a la celda k^*-ésima que se transforman en $\boldsymbol{y}_o$ tras la adición del ruido, es decir, $f(\boldsymbol{x}, \boldsymbol{n})$ es una función sobreyectiva. Dado que estamos en un marco probabilístico, a este conjunto de vectores le podemos asociar una densidad de probabilidad que es justamente $p(\boldsymbol{x}|\boldsymbol{y}_o, k^*, \Lambda)$. Intuitivamente podemos pensar que cuanto menor sea la SNR de la señal de voz observada, mayor será la varianza de esta distribución, tendiendo a $p(\boldsymbol{x}|k^*)$ para valores de SNR muy bajos. Por contra, para valores de SNR altos, $p(\boldsymbol{x}|\boldsymbol{y}_o, k^*, \Lambda)$ se aproximará a una delta de Dirac centrada en la observación, esto es, $\delta_{\boldsymbol{y}_o}(\boldsymbol{x})$. El modelado de este comportamiento, como veremos, será un punto clave en el diseño del estimador MMSE. Finalmente, tal y como aparece reflejado en la figura 3.1, la estima parcial $\hat{\boldsymbol{x}}^{(k^*)}$ obtenida por el estimador MMSE no es más que la media de esta distribución.

La formulación del estimador MMSE que acabamos de derivar será la que manejaremos a lo largo de esta tesis. En la siguiente sección dicho estimador será particularizado a la estrategia de compensación de la voz que emplea grabaciones estéreo. Por otro la-

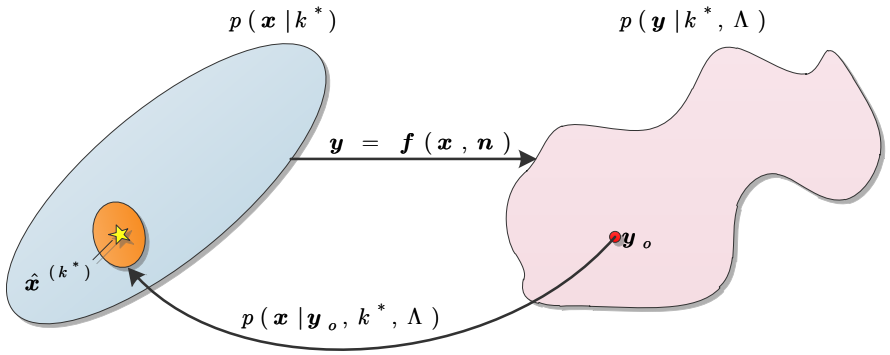

Figura 3.1: Esquema de la estimación MMSE basada en diccionario.

do, en el capítulo 4 veremos cómo puede adaptarse este estimador al paradigma de datos perdidos.

3.2. Compensación eficiente basada en diccionarios VQ

A lo largo de esta sección se derivan un conjunto de técnicas de compensación que permitan robustecer el comportamiento de los sistemas de RAH frente a degradaciones producidas en la voz. Dichas técnicas se sustentarán sobre los siguientes pilares: (i) todas se derivarán del estimador MMSE basado en diccionario introducido en la sección anterior, (ii) las transformaciones empleadas para mitigar la distorsión introducida por el ruido serán estimadas usando grabaciones estéreo, (iii) el dominio en el que trabajan es el de los MFCCs y (iv) las técnicas emplearán un modelado VQ de los espacios de parámetros limpio y distorsionado que, como veremos, traerá asociado un coste computacional muy reducido en comparación con otras técnica similares, sin que ello suponga una merma en los resultados obtenidos. En base a estos tres pilares, a continuación desarrollaremos un marco matemático común para la estimación MMSE basada en diccionarios VQ (VQMMSE) que, posteriormente, se particularizará para cada una de las técnica propuestas. En particular, la principal diferencia entre las técnicas propuestas radicará en la forma de la transformación aplicada a los coeficientes de la voz.

Por conveniencia, volvemos a reproducir aquí la fórmula general del estimador MM-

SE basado en diccionario:

$$\hat{\boldsymbol{x}} = \sum_{k_x=1}^{M_x} P(k_x|\boldsymbol{y}, \Lambda) \underbrace{\int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}, k_x, \Lambda) d\boldsymbol{x}}_{\hat{\boldsymbol{x}}^{(k_x)}}, \tag{3.10}$$

donde se ha modificado ligeramente la notación añadiendo el subíndice x al índice empleado para referenciar a las regiones del espacio de características sin distorsionar, k, así como al número total de regiones M. Las razones de este cambio se justificarán más adelante.

Como se puede apreciar en la ecuación anterior, para la estimación del vector voz limpia se requiere el cómputo de las probabilidades $P(k_x|\boldsymbol{y}, \Lambda)$ y las estimas parciales $\hat{\boldsymbol{x}}^{(k_x)}$ para cada una de las regiones. En primer lugar, comenzaremos derivando las expresiones oportunas para el cálculo de las probabilidades a posteriori. Para obtener estas probabilidades, en esta sección nos apoyaremos en un modelado dual de los espacios limpio (sin distorsionar) y ruidoso (distorsionado) de los parámetros MFCC de la voz. Así, consideraremos que la información a priori disponible sobre el entorno acústico, Λ, se materializa en un modelo estadístico de mezcla $p(\boldsymbol{y}|\mathcal{M}_y)$ que representa a los parámetros de la voz distorsionados por dicho entorno. Asimismo se dispone de un segundo modelo, $p(\boldsymbol{x}|\mathcal{M}_x)$, que representa los parámetros de voz sin distorsionar. En resumidas cuentas, se cuenta con los siguientes modelos de mezcla de PDFs

$$p(\boldsymbol{x}|\mathcal{M}_x) = \sum_{k_x=1}^{M_x} P(k_x|\mathcal{M}_x) p(\boldsymbol{x}|k_x, \mathcal{M}_x), \tag{3.11}$$

$$p(\boldsymbol{y}|\mathcal{M}_y) = \sum_{k_y=1}^{M_y} P(k_y|\mathcal{M}_y) p(\boldsymbol{y}|k_y, \mathcal{M}_y). \tag{3.12}$$

Usando este modelado dual de la voz, la probabilidad a posteriori $P(k_x|\boldsymbol{y}) \equiv P(k_x|\boldsymbol{y}, \mathcal{M}_x, \mathcal{M}_y)$ requerida por el estimador MMSE de la ecuación (3.10) se puede expresar como la distribución marginal de $P(k_x, k_y|\boldsymbol{y})$:

$$\begin{aligned} P(k_x|\boldsymbol{y}) &= \sum_{k_y}^{M_y} P(k_x, k_y|\boldsymbol{y}) \\ &= \sum_{k_y}^{M_y} P(k_x|k_y, \boldsymbol{y}) P(k_y|\boldsymbol{y}) \\ &\approx \sum_{k_y}^{M_y} P(k_x|k_y) P(k_y|\boldsymbol{y}), \end{aligned} \tag{3.13}$$

donde $P(k_y|\boldsymbol{y})$ es la probabilidad a posteriori de la región k_y-ésima del espacio ruidoso dada la observación $\boldsymbol{y}$, mientras que $P(k_x|k_y)$ modela la estadística de cómo las regiones

de los espacios limpios y distorsionado se transforman entre ellas como consecuencia del ruido.

De las dos probabilidades que aparecen en la ecuación (3.13) la única que, según lo expuesto hasta ahora, es directamente computable es $P(k_y|\boldsymbol{y})$. En la mayoría de técnicas de compensación que emplean datos estéreo como por ejemplo SPLICE [68, 79] o MEMLIN [44], el cómputo de esta probabilidad involucra la evaluación de una gaussiana $p(\boldsymbol{y}|k_y) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_y^{(k_y)}, \boldsymbol{\Sigma}_y^{(k_y)})$. Con objeto de conseguir un diseño más eficiente del estimador MMSE, aquí emplearemos diccionarios VQ para el modelado de los espacios de características, es decir, los modelos de mezcla de PDFs $p(\boldsymbol{x}|\mathcal{M}_x)$ y $p(\boldsymbol{y}|\mathcal{M}_y)$ en nuestro caso serán diccionarios VQ con M_x y M_y celdas, respectivamente, obtenidas tras aplicar, de forma independiente, el algoritmo de Linde-Buzo-Gray [178] sobre sendos conjuntos de entrenamiento que contienen parámetros MFCC limpios y distorsionados. Cada una de las celdas $k = 1, \ldots, M$ en las que se dividen los espacios de parámetros vendrá caracterizada por dos parámetros: el vector de media (o centroide) de la celda $\boldsymbol{\mu}^{(k)}$ y la matriz de covarianza de la misma $\boldsymbol{\Sigma}^{(k)}$. Asimismo, en el modelado VQ consideraremos que todas las celdas son equiprobables y que éstas son disjuntas, esto es, no se produce solapamiento entre celdas.

Aplicando este modelado VQ, la probabilidad $P(k_y|\boldsymbol{y})$ que aparece en la ecuación (3.13) se puede calcular como

$$P(k_y|\boldsymbol{y}) = \delta(k_y - k_y^*(\boldsymbol{y})), \tag{3.14}$$

donde $\delta(x)$ es la función delta de Kronecker,

$$\delta(x) = \begin{cases} 1, & \text{si } \ x = 0 \\ 0, & \text{en otro caso} \end{cases}, \tag{3.15}$$

y $k_y^*(\boldsymbol{y})$ el valor de $\boldsymbol{y}$ cuantificado o, de forma equivalente, la celda del espacio de características distorsionado a la que el vector observado pertenece. Para calcular esta celda, aquí emplearemos la siguiente distancia de Mahalanobis:

$$\mathrm{d}(\boldsymbol{y}, k_y) = (\boldsymbol{\mu}_y^{(k_y)} - \boldsymbol{y})^\top \boldsymbol{\Sigma}_y^{(k_y)^{-1}} (\boldsymbol{\mu}_y^{(k_y)} - \boldsymbol{y}). \tag{3.16}$$

Luego $k_y^* \equiv k_y^*(\boldsymbol{y})$ es aquella celda que minimiza la distancia anterior,

$$k_y^* = \underset{1 \leq k_y \leq M_y}{\operatorname{argmin}} \, \mathrm{d}(\boldsymbol{y}, k_y). \tag{3.17}$$

El otro término requerido por la ecuación (3.13) para el cálculo de las probabilidades $P(k_x|\boldsymbol{y})$ es el modelo de probabilidades cruzadas $P(k_x|k_y)$. Este modelo nos da una idea de cómo las regiones del espacio limpio se modifican por el ruido hasta transformarse en las del espacio distorsionado. Para calcular estas probabilidades en esta

sección emplearemos grabaciones estéreo. Así, asumimos la existencia de un conjunto de datos de entrenamiento sin distorsionar $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T)$ y su correspondiente conjunto asociado $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$. Este último conjunto puede obtenerse de distintas formas. La manera más común consiste en contaminar artificialmente $\boldsymbol{X}$ con el tipo de ruido deseado, siendo ésta, por ejemplo, la estrategia empleada en las bases de datos Aurora2 [141] y Aurora4 [140]. Otra estrategia alternativa requiere la existencia de varios micrófonos situados a distintas distancias del locutor. En este caso, $\boldsymbol{X}$ contiene las grabaciones realizadas con el micrófono más cercano al locutor y, por tanto, el que adquiere la señal de voz con una SNR más alta. El conjunto $\boldsymbol{Y}$, por contra, contiene las grabaciones realizadas con un micrófono lejano, el cual recoge más ruido ambiental y posiblemente tendrá una respuesta en frecuencia distinta del micrófono empleado para grabar $\boldsymbol{X}$. Un ejemplo representativo de esta segunda estrategia lo tenemos en la base de datos Aurora3 [187].

Sea como fuere la forma de obtener $\boldsymbol{Y}$, la probabilidad $P(k_x|k_y)$ puede estimarse mediante la siguiente ecuación

$$P(k_x|k_y) = \frac{\sum_{t=1}^{T} \delta(k_x - k_x^*(\boldsymbol{x}_t))\delta(k_y - k_y^*(\boldsymbol{y}_t))}{\sum_{t=1}^{T} \delta(k_y - k_y^*(\boldsymbol{y}_t))}. \tag{3.18}$$

Así vemos que $P(k_x|k_y)$ se corresponden con el cociente entre el número de pares de vectores de entrenamiento $\langle \boldsymbol{x}_t, \boldsymbol{y}_t \rangle$ en donde $\boldsymbol{x}_t$ se cuantifica por k_x y $\boldsymbol{y}_t$ por k_y, dividido por el número total de vectores distorsionados que pertenecen a la celda k_y.

Finalmente, usando las ecuaciones (3.13) y (3.14) en la ecuación (3.10), el estimador MMSE basado en diccionarios VQ (VQMMSE) resultante viene dado por

$$\begin{aligned} \hat{\boldsymbol{x}} &= \sum_{k_x=1}^{M_x} \sum_{k_y=1}^{M_y} P(k_y|\boldsymbol{y}) P(k_x|\boldsymbol{y}, k_y) \int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}, k_x, k_y) d\boldsymbol{x} \\ &\approx \sum_{k_x=1}^{M_x} P(k_x|k_y^*) \underbrace{\int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}, k_x, k_y^*) d\boldsymbol{x}}_{\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, k_x, k_y^*]}. \end{aligned} \tag{3.19}$$

Una forma alternativa de expresar el estimador VQMMSE se deriva del modelo de distorsión de los parámetros de la voz de la ecuación (2.12). De acuerdo a este modelo, el vector observado es suma del vector de voz original más una cierta corrección aditiva,

$$\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n}). \tag{3.20}$$

Para un entorno acústico determinado podemos suponer que $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{h})$ es función únicamente de $\boldsymbol{y}$, esto es, $\boldsymbol{r}(\boldsymbol{y}) \approx \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n})$. Luego dada la estimación del vector de

corrección $\hat{\boldsymbol{r}}(\boldsymbol{y})$, la estima del vector de voz limpia viene dada por

$$\hat{\boldsymbol{x}} = \boldsymbol{y} - \hat{\boldsymbol{r}}(\boldsymbol{y}). \tag{3.21}$$

Suponiendo que

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, k_x, k_y] = \boldsymbol{y} - \mathbb{E}[\boldsymbol{r}|\boldsymbol{y}, k_x, k_y], \tag{3.22}$$

entonces

$$\hat{\boldsymbol{r}}(\boldsymbol{y}) = \sum_{k_x=1}^{M_x} P(k_x|k_y^*) \underbrace{\int \boldsymbol{r} p(\boldsymbol{r}|\boldsymbol{y}, k_x, k_y^*) d\boldsymbol{r}}_{\mathbb{E}[\boldsymbol{r}|\boldsymbol{y}, k_x, k_y^*]}. \tag{3.23}$$

A raíz de las fórmulas de estimación obtenidas en (3.19) y (3.23), varios son los puntos que debemos aclarar. En primer lugar nos encontramos con la cuestión de cómo se estima el modelo $p(\boldsymbol{y}|\mathcal{M}_y)$ requerido por el estimador VQMMSE. En esta sección hemos supuesto que este modelo es conocido a priori. Esta suposición, no obstante, raramente suele ser cierta. En la práctica $p(\boldsymbol{y}|\mathcal{M}_y)$ debe estimarse utilizando, por ejemplo, la técnica VTS descrita en la sección 2.2.2.2 o la aproximación basada en el modelo de enmascaramiento de la voz que presentaremos en la sección 4.3.3. Para evitar complicar la derivación de la técnica VQMMSE con la estimación de este modelo, a lo largo de esta sección consideraremos que en la fase de entrenamiento es posible conocer el conjunto de entornos acústicos bajo los cuales el sistema de reconocimiento va a trabajar. Aunque en general esta suposición no es cierta, pueden existir algunas situaciones en las que sí sea posible conocer con gran precisión las características de los entornos a los que el sistema de reconocimiento se expondrá: por ejemplo en sistemas de reconocimiento integrados en vehículos o en entornos de oficina. En estos entornos, por tanto, será posible entrenar un diccionario VQ $p(\boldsymbol{y}|\mathcal{M}_y, e)$ para cada uno de los diferentes entornos acústicos $e = 1, \ldots, E$. El problema de cómo averiguar el entorno acústico presente en la señal de voz observada de entre los E posibles, así como resolver los casos en los cuales la distorsión que contamina la señal de entrada no haya sido considerada en la fase de entrenamiento, serán tratados con más detalle en la sección 3.2.4.

Otro punto muy importante a considerar es el proceso de cuantificación al que se ve sometida la señal de entrada. Tal y como se aprecia en las ecuaciones (3.19) y (3.23), este proceso sólo afecta al cálculo de las probabilidades de mezcla del estimador, pero no al cálculo de los valores esperados parciales $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, k_x, k_y^*]$ y $\mathbb{E}[\boldsymbol{r}|\boldsymbol{y}, k_x, k_y^*]$. De hecho se debe notar cómo el vector observado $\boldsymbol{y}$ aparece reflejado explícitamente en estos términos. Esta forma de proceder tiene como objetivo el lograr un diseño eficiente del estimador MMSE, pero sin que ello conlleve una degradación del rendimiento por el modelado VQ llevado a cabo.

La última cuestión remanente en relación al estimador VQMMSE es el cómputo de los valores esperados $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, k_x, k_y^*]$ y $\mathbb{E}[\boldsymbol{r}|\boldsymbol{y}, k_x, k_y^*]$ que aparecen en la formulación del mismo. Estos valores, grosso modo, pueden considerarse como funciones de transformación $\hat{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{y}; k_x, k_y^*)$ que devuelven una estima más o menos precisa del vector de voz limpia dado el vector observado. En los siguientes apartados propondremos distintas transformaciones afines para este propósito con una complejidad y una precisión en la compensación del ruido crecientes. Como resultado se tendrá distintas versiones del estimador VQMMSE con diferentes requerimientos computaciones y distinto poder de compensación.

3.2.1. Estimación VQMMSE cuantificada

El uso de cuantificadores VQ, tal y como se ha planteado para el caso del estimador VQMMSE, tiene una gran tradición en el campo del procesado digital de la voz. De hecho en los estándares de reconocimiento distribuido de la voz (DSR, *Distributed Speech Recognition*) [1, 2, 3, 4, 215] los vectores de características de la voz extraídos por los terminales cliente son cuantificados y transmitidos posteriormente al servidor remoto de reconocimiento. Con ello se pretende un uso eficiente de las capacidades de la red de datos, sin que ello conlleve una merma en la tasa de reconocimiento obtenida. En este apartado nosotros adoptaremos esta misma filosofía para el cálculo de los valores esperados $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, k_x, k_y^*]$ que aparecen en la ecuación (3.19).

En primer lugar tenemos que el proceso de cuantificación al que se ve sometida la señal observada conlleva que los vectores $\boldsymbol{y}$ se aproximan por los centroides de las celdas a las que pertenecen, esto es, $\boldsymbol{y} \approx \boldsymbol{\mu}_y^{(k_y^*)}$. Asimismo supondremos que el espacio de características limpias también está cuantificado, de forma que todos los vectores pertenecientes a la celda k_x se representan por su centroide $\boldsymbol{\mu}_x^{(k_x)}$ correspondiente. Suponiendo, adicionalmente, que ambos espacios (limpio y distorsionado) son independientes, podemos aproximar el valor esperado $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, k_x, k_y^*]$ por

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, k_x, k_y^*] = \mathbb{E}[\boldsymbol{x}|k_x] = \boldsymbol{\mu}_x^{(k_x)}, \tag{3.24}$$

esto es, el valor esperado coincide con el centroide de cada celda independientemente de la observación $\boldsymbol{y}$ y de la celda k_y^* a la que ésta pertenece.

Usando este valor esperado en la ecuación general (3.19), obtenemos el siguiente estimador al cual denominaremos estimador VQMMSE cuantificado (Q-VQMMSE, *Quantized* VQMMSE),

$$\hat{\boldsymbol{x}}_q = \sum_{k_x=1}^{M_x} P(k_x|k_y^*)\boldsymbol{\mu}_x^{(k_x)}. \tag{3.25}$$

El estimador Q-VQMMSE que acabamos de presentar se puede expresar, de forma alternativa, en términos de vectores de corrección aplicados a la señal de entrada, tal y como se recoge en las ecuaciones (3.21) y (3.23). Aunque el estimador será el mismo, esta notación alternativa nos permitirá, como ya veremos más adelante, derivar transformaciones de compensación más precisas. Bajo los supuestos anteriormente mencionados (ambos espacios cuantificados e independientes), el valor esperado $\mathbb{E}[\boldsymbol{r}|\boldsymbol{y}, k_x, k_y]$ se puede calcular como sigue:

$$\begin{aligned}\mathbb{E}[\boldsymbol{r}|\boldsymbol{y}, k_x, k_y] &= \mathbb{E}[\boldsymbol{r}|k_x, k_y] = \mathbb{E}[\boldsymbol{y} - \boldsymbol{x}|k_x, k_y] \\ &= \mathbb{E}[\boldsymbol{y}|k_x, k_y] - \mathbb{E}[\boldsymbol{x}|k_x, k_y] \\ &= \mathbb{E}[\boldsymbol{y}|k_y] - \mathbb{E}[\boldsymbol{x}|k_x] \\ &= \boldsymbol{\mu}_y^{(k_y)} - \boldsymbol{\mu}_x^{(k_x)}. \end{aligned} \tag{3.26}$$

Como puede apreciarse, el vector de corrección aplicado por Q-VQMMSE para cada par de celdas (k_x, k_y) es simplemente la diferencia entre los centroides de las celdas. Suponiendo que $\boldsymbol{y}$ está cuantificado, la corrección $\hat{\boldsymbol{r}}(\boldsymbol{y})$ de la ecuación (3.23) equivale a $\boldsymbol{r}^{(k_y^*)} \equiv \hat{\boldsymbol{r}}(\boldsymbol{\mu}_y^{(k_y^*)})$, esto es, la corrección promedio para la celda VQ k_y^*-ésima a la que $\boldsymbol{y}$ pertenece. Usando los valores esperados de la ecuación (3.26) en la ecuación (3.23), el valor de $\boldsymbol{r}^{(k_y^*)}$ se calcula como

$$\begin{aligned}\boldsymbol{r}^{(k_y^*)} &= \sum_{k_x=1}^{M_x} P(k_x|k_y^*)\mathbb{E}[\boldsymbol{r}|k_x, k_y^*] \\ &= \boldsymbol{\mu}_y^{(k_y^*)} - \underbrace{\sum_{k_x=1}^{M_x} P(k_x|k_y^*)\boldsymbol{\mu}_x^{(k_x)}}_{\hat{\boldsymbol{x}}_q}. \end{aligned} \tag{3.27}$$

El valor $\hat{\boldsymbol{x}}_q$ que aparece en la ecuación anterior es el devuelto por el estimador Q-VQMMSE de la ecuación (3.25). Esto implica que la corrección aplicada a los vectores observados es constante dentro de cada celda k_y del espacio distorsionado y que depende únicamente de la estadística $P(k_x|k_y)$. Como esta estadística es conocida a priori para cada entorno acústico, estos valores se podrán precalcular en la fase de entrenamiento del sistema. En los siguientes apartados ahondaremos más en esta idea para permitir que estas correcciones dependan del vector observado y, de esta forma, conseguir una compensación más dinámica y fidedigna.

Finalmente, el estimador Q-VQMMSE expresado en términos de vectores de corrección es

$$\hat{\boldsymbol{x}}_q = \boldsymbol{\mu}_y^{(k_y^*)} - \boldsymbol{r}^{(k_y^*)}, \tag{3.28}$$

ecuación que coincide con la ya presentada en (3.25).

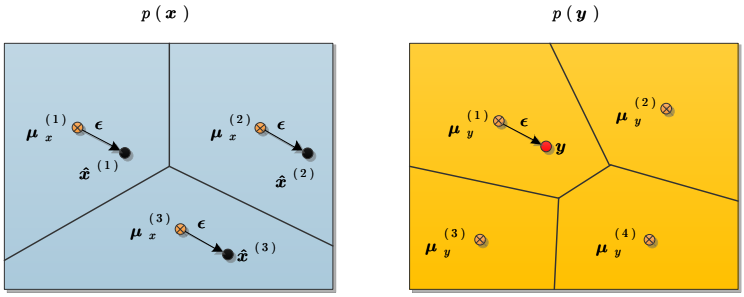

Figura 3.2: Esquema del estimador S-VQMMSE.

3.2.2. Estimación VQMMSE con correcciones simples

Examinemos de nuevo el las expresiones derivadas para la técnica Q-VQMMSE en las ecuaciones (3.28) y (3.26). En este estimador el vector observado $\boldsymbol{y}$ se sustituye por su versión cuantificada, $\boldsymbol{\mu}_y^{(k_y^*)}$, a fin de simplificar el desarrollo matemático del mismo. No obstante, este proceso de cuantificación conlleva asociada una distorsión o error de cuantificación que podría limitar el rendimiento alcanzado por el estimador. Asimismo, debemos recordar que, en líneas generales, los vectores de características utilizados por un sistema de reconocimiento no están cuantificados. Por tanto, en este apartado propondremos una nueva variante de compensación basada en el estimador VQMMSE que intenta solventar este problema.

Consideremos de nuevo la formulación basada en vectores de corrección del estimador VQMMSE que aparece en la ecuación (3.21). Al contrario que en el caso de la técnica Q-VQMMSE, en dicha ecuación aparece explícitamente el vector observado y no su versión cuantificada. No obstante, el cálculo de una corrección $\hat{\boldsymbol{r}}(\boldsymbol{y})$ para cada vector $\boldsymbol{y}$, tal y como se recoge en (3.21), podría ser inviable por la complejidad que esto conllevaría. Por tanto, la solución que aquí adoptamos se basa en suponer que las correcciones son homogéneas y constantes dentro de cada celda del espacio distorsionado. Bajo esta premisa podemos aproximar $\hat{\boldsymbol{r}}(\boldsymbol{y})$ por el valor $\boldsymbol{r}^{(k_y^*)}$ calculado para Q-VQMMSE en la ecuación (3.27). Usando este valor en la ecuación (3.21), obtenemos el siguiente estimador al que denominaremos VQMMSE con correcciones simples (S-VQMMSE, *Simple bias* VQMMSE),

$$\hat{\boldsymbol{x}} = \boldsymbol{y} - \boldsymbol{r}^{(k_y^*)} = \boldsymbol{y} - (\boldsymbol{\mu}_y^{(k_y^*)} - \hat{\boldsymbol{x}}_q) = \hat{\boldsymbol{x}}_q + \underbrace{(\boldsymbol{y} - \boldsymbol{\mu}_y^{(k_y^*)})}_{\boldsymbol{\epsilon}}, \tag{3.29}$$

donde $\hat{\boldsymbol{x}}_q$ es la estimación Q-VQMMSE de la ecuación (3.25).

En la figura 3.2 se muestra un esquema gráfico del estimador S-VQMMSE. En este ejemplo los espacios de características limpias y distorsionadas se modelan me-

diante sendos diccionarios VQ con 3 y 4 celdas, respectivamente, cuyos centroides son $(\boldsymbol{\mu}_x^{(1)}, \boldsymbol{\mu}_x^{(2)}, \boldsymbol{\mu}_x^{(3)})$ para las celdas del espacio limpio y $(\boldsymbol{\mu}_y^{(1)}, \boldsymbol{\mu}_y^{(2)}, \boldsymbol{\mu}_y^{(3)}, \boldsymbol{\mu}_y^{(4)})$ para la celdas del espacio sucio. Supuesto que el vector observado $\boldsymbol{y}$ pertenece a la primera celda del diccionario ruidoso, el estimador S-VQMMSE calcula el error de cuantificación $\boldsymbol{\epsilon} = \boldsymbol{y} - \boldsymbol{\mu}_y^{(k_y^*)}$ usado para corregir la posición de los centroides del diccionario limpio y así obtener las estimas parciales $\hat{\boldsymbol{x}}^{(k_x)}$ $(k_x = 1, 2, 3)$:

$$\begin{aligned}\hat{\boldsymbol{x}}^{(k_x)} &\equiv \mathbb{E}[\boldsymbol{x}|k_x, k_y^*, \boldsymbol{y}] = \boldsymbol{y} - \mathbb{E}[\boldsymbol{r}|k_x, k_y^*] \\ &= \boldsymbol{y} - (\boldsymbol{\mu}_y^{(k_y^*)} - \boldsymbol{\mu}_x^{(k_x)}) \\ &= \boldsymbol{\mu}_x^{(k_x)} + \boldsymbol{\epsilon}. \end{aligned} \tag{3.30}$$

En base a estas estimas parciales, la técnica S-VQMMSE calcula la estima final $\hat{\boldsymbol{x}}$ como una combinación lineal de las mismas, siendo los pesos de la combinación las probabilidades a posteriori $P(k_x|k_y^*)$ de las distintas celdas limpias k_x dada la celda k_y^* a la que pertenece el vector observado. En el valor de estas probabilidades, que modelan las transformaciones entre las distintas celdas debidas al ruido, influirán factores como la SNR y el tipo de distorsión que afecta al ruido.

El uso de vectores de corrección calculados para cada una de las regiones del espacio ruidoso no es exclusivo de la técnica S-VQMMSE que acabamos de presentar, sino que estos también han sido empleados por otras técnicas de compensación basadas en datos estéreo. En concreto, varias son las similitudes entre nuestro estimador y la técnica SPLICE estudiada en la sección 2.2.3.4. Así podemos observar que ambas técnicas calculan una estimación parcial para cada región k_y del espacio ruidoso, y que esta estima se obtiene sustrayendo la corrección $\boldsymbol{r}^{(k_y)}$ a $\boldsymbol{y}$ (en nuestro caso solamente se usa el valor estimado para la celda más probable). No obstante, mientras que en SPLICE únicamente se modela el espacio características distorsionadas, $p(\boldsymbol{y})$, en nuestra propuesta se modelan ambos espacios $p(\boldsymbol{x})$ y $p(\boldsymbol{y})$. Como consecuencia de ello la varianza en el cálculo de las correcciones $\boldsymbol{r}^{(k_y)}$ será menor para la técnica S-VQMMSE y las correcciones obtenidas serán más precisas [44].

Aunque la transformación aplicada por S-VQMMSE sea más precisa que la obtenida para la técnica Q-VQMMSE, aún existe margen de mejora que puede aprovecharse para obtener estimaciones más fidedignas. En este sentido podemos considerar que una transformación afín del tipo $\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}$ modelará mejor la distorsión no lineal introducida por el ruido que las correcciones simples usadas por la técnica S-VQMMSE. De igual modo, un modelado más detallado de la distorsión producida por el ruido en las regiones del espacio limpio y cómo éstas se transforman en sus correspondientes del espacio distorsionado, generará también grandes mejoras. Ambos aspectos, serán estudiados en detalle en la próxima sección.

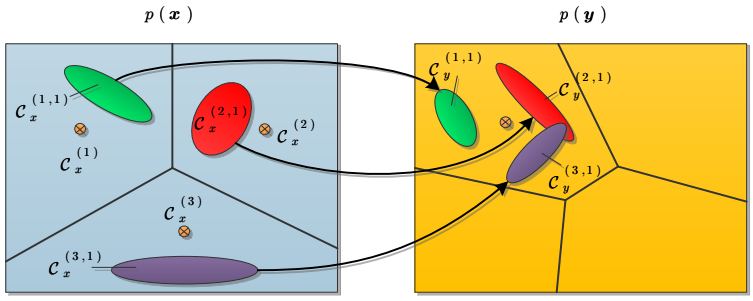

Figura 3.3: Concepto de subregión de una celda VQ y transformación entre ellas a causa al ruido.

3.2.3. Mejora del modelo de distorsión

Consideremos el esquema de la figura 3.3 en donde se muestran las transformaciones producidas a causa del ruido entre celdas de los espacios limpio y distorsionado. Debido a que la técnica VQMMSE estima de forma independiente los diccionarios VQ que modelan ambos espacios de características, no podemos asegurar que exista una correspondencia biunívoca (uno a uno) entre cada par de celdas (k_x, k_y) de ambos espacios, o dicho de otra forma, el conjunto de los vectores de características pertenecientes a una cierta celda k_x del espacio limpio, una vez distorsionado dicho conjunto por el ruido, no tiene por qué coincidir al completo con los vectores pertenecientes a una celda k_y del espacio distorsionado.

En este sentido lo máximo que podemos afirmar sobre la relación entre las celdas de ambos espacios es que, para cada celda k_x del espacio limpio, existirá un subconjunto de vectores $\mathcal{C}_x^{(k_x,k_y)} \in \mathcal{C}_x^{(k_x)}$ (subconjunto posiblemente vacío) cuya imagen en el espacio de características distorsionadas será el subconjunto $\mathcal{C}_y^{(k_x,k_y)} \in \mathcal{C}_y^{(k_y)}$ de la celda k_y-ésima:

$$\boldsymbol{x} \in \mathcal{C}_x^{(k_x,k_y)} \Leftrightarrow \boldsymbol{y} \in \mathcal{C}_y^{(k_x,k_y)}, \boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{h}, \boldsymbol{n}). \tag{3.31}$$

De aquí en adelante a estos subconjuntos los denominaremos subregiones de una celda VQ dada. Al contrario que la división de los espacios de características en celdas disjuntas, las distintas subregiones de cada celda VQ estarán altamente solapadas entre ellas debido a la aleatoriedad del ruido. Como se puede apreciar en la figura 3.3, cada celda del espacio ruidoso contará con un número de subregiones equivalente al número de celdas del espacio limpio. De forma análoga, las celdas del espacio limpio contarán con tantas subregiones, algunas posiblemente vacías, como celdas haya en el espacio

distorsionado,

$$\mathcal{C}_x^{(k_x)} = \bigcup_{k_y=1}^{M_y} \mathcal{C}_x^{(k_x,k_y)}, \tag{3.32}$$

$$\mathcal{C}_y^{(k_y)} = \bigcup_{k_x=1}^{M_x} \mathcal{C}_y^{(k_x,k_y)}. \tag{3.33}$$

A partir de la introducción anterior se deduce que, a través del conocimiento de las subregiones, es posible obtener transformaciones más precisas que permitan combatir de forma más eficaz la distorsión introducida por el ruido. Por lo tanto, el objetivo de este apartado es la extensión del estimador VQMMSE para que contemple el uso de subregiones. Aparte de esto, también consideraremos un modelado más general de la transformación que se produce entre cada par de subregiones. Así, como se muestra en la figura 3.3, la distorsión producida por el ruido en las subregiones del espacio limpio no sólo se manifiesta en una traslación de las mismas, sino también en rotaciones y cambios de escala. A fin de conseguir una mejor representación de las deformaciones que sufren las subregiones, en este apartado proponemos además su modelado mediante un conjunto de transformaciones afines cuyos parámetros se derivarán de la estadística de las subregiones.

El primer paso para lograr los dos objetivos planteados consiste en obtener una expresión paramétrica que modele la estadística de las distintas subregiones. Por simplicidad aquí supondremos que las subregiones se modelan mediante densidades de probabilidad gaussianas con los siguientes parámetros:

$$\mathcal{C}_x^{(k_x,k_y)} \sim \mathcal{N}(\boldsymbol{\mu}_x^{(k_x,k_y)}, \boldsymbol{\Sigma}_x^{(k_x,k_y)}), \tag{3.34}$$

$$\mathcal{C}_y^{(k_x,k_y)} \sim \mathcal{N}(\boldsymbol{\mu}_y^{(k_x,k_y)}, \boldsymbol{\Sigma}_y^{(k_x,k_y)}). \tag{3.35}$$

Los vectores de media y matrices de covarianza de las PDFs anteriores pueden calcularse fácilmente a partir de los vectores de características de entrenamiento asignados a cada subregión. Notando mediante N al número de pares de vectores entrenamiento $\langle \boldsymbol{x}_t, \boldsymbol{y}_t \rangle$ en los que $\boldsymbol{x}_t$ pertenece a la celda k_x-ésima y $\boldsymbol{y}_t$ a la celda k_y,

$$N = \sum_{t=1}^{T} \delta(k_x - k_x^*(\boldsymbol{x}_t))\delta(k_y - k_y^*(\boldsymbol{y}_t)), \tag{3.36}$$

entonces las expresiones para el cálculo de las medias de las subregiones son

$$\boldsymbol{\mu}_x^{(k_x,k_y)} = \frac{1}{N} \sum_{t=1}^{T} \delta(k_x - k_x^*(\boldsymbol{x}_t))\delta(k_y - k_y^*(\boldsymbol{y}_t))\boldsymbol{x}_t, \tag{3.37}$$

$$\boldsymbol{\mu}_y^{(k_x,k_y)} = \frac{1}{N} \sum_{t=1}^{T} \delta(k_x - k_x^*(\boldsymbol{x}_t))\delta(k_y - k_y^*(\boldsymbol{y}_t))\boldsymbol{y}_t, \tag{3.38}$$

y las matrices de covarianza se calculan como

$$\boldsymbol{\Sigma}_x^{(k_x,k_y)} = \frac{1}{N-1}\sum_{t=1}^{T}\delta(k_x - k_x^*(\boldsymbol{x}_t))\delta(k_y - k_y^*(\boldsymbol{y}_t))\left(\boldsymbol{x}_t - \boldsymbol{\mu}_x^{(k_x,k_y)}\right)\left(\boldsymbol{x}_t - \boldsymbol{\mu}_x^{(k_x,k_y)}\right)^\top, \tag{3.39}$$

$$\boldsymbol{\Sigma}_y^{(k_x,k_y)} = \frac{1}{N-1}\sum_{t=1}^{T}\delta(k_x - k_x^*(\boldsymbol{x}_t))\delta(k_y - k_y^*(\boldsymbol{y}_t))\left(\boldsymbol{y}_t - \boldsymbol{\mu}_y^{(k_x,k_y)}\right)\left(\boldsymbol{y}_t - \boldsymbol{\mu}_y^{(k_x,k_y)}\right)^\top. \tag{3.40}$$

De las ecuaciones anteriores se deduce que el modelado VQ puede considerarse como una simplificación del modelado basado subregiones propuesto en este apartado. Así podemos establecer equivalencias entre ambos modelados sin más que considerar que en las técnicas Q-VQMMSE y S-VQMMSE no hay distinción entre celdas VQ y subregiones y, por tanto, todas las subregiones de una celda dada coinciden con la propia celda. Desde este punto de vista el modelado basado en subregiones puede considerarse como más preciso. Los resultados de reconocimiento que presentaremos en el capítulo 6 avalarán esta afirmación.

Después de haber definido el concepto de subregión y haber presentado un modelo probabilístico para cada una de ellas, a continuación abordamos el estudio de las transformaciones que se producen entre las subregiones y su modelado matemático. Como primer intento consideraremos el uso de transformaciones del tipo de las usadas por las técnicas SSM y FE-Joint estudiadas en la sección 2.2.3.4. Recordemos que en estas técnicas ambos espacios de características se modelan de forma conjunta mediante un GMM $p(\boldsymbol{x},\boldsymbol{y})$. Este modelo se emplea posteriormente en el marco de la estimación MMSE para calcular las estimas oportunas de la voz, las cuales se calculan como una combinación lineal de las medias de las distribuciones condicionales $p(\boldsymbol{x}|\boldsymbol{y},k)$ para cada gaussiana k del GMM (ver ecuación (2.143)). Aplicando este tipo de transformaciones a nuestro problema, las estimaciones parciales $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y},k_x,k_y]$ empleadas por la técnica VQMMSE pueden calcularse como

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y},k_x,k_y] = \boldsymbol{\mu}_x^{(k_x,k_y)} + \boldsymbol{\Sigma}_{xy}^{(k_x,k_y)}{\boldsymbol{\Sigma}_y^{(k_x,k_y)}}^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}_y^{(k_x,k_y)}\right), \tag{3.41}$$

donde $\boldsymbol{\Sigma}_{xy}^{(k_x,k_y)}$ es la matriz de correlación cruzada entre los vectores de características pertenecientes a ambas subregiones. De nuevo esta matriz puede calcularse fácilmente usando grabaciones estéreo.

Además de la transformación anterior, en este apartado también proponemos un método alternativo para el cálculo $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y},k_x,k_y]$ basado en la transformación de blanqueo (*whitening transformation*, en inglés) [216]. Esta transformación normaliza, en primer lugar, los vectores distorsionados en media y varianza usando los parámetros

de la subregión del espacio distorsionado y, posteriormente, transforma estos vectores normalizados al espacio limpio usando los parámetros (media y matriz de covarianza) de la subregión limpia. Formalmente, la estimación que proponemos para el cómputo de la estimas parciales es

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, k_x, k_y] = \boldsymbol{\mu}_x^{(k_x,k_y)} + \left(\boldsymbol{\Sigma}_x^{(k_x,k_y)}\right)^{1/2} \left(\boldsymbol{\Sigma}_y^{(k_x,k_y)}\right)^{-1/2} \left(\boldsymbol{y} - \boldsymbol{\mu}_y^{(k_x,k_y)}\right). \tag{3.42}$$

El exponente 1/2 al que están elevadas ambas matrices de covarianza denota la raíz cuadrada de la matriz en cuestión. Dada una matriz $\boldsymbol{\Sigma}$ su matriz raíz cuadrada $\boldsymbol{\Sigma}^{1/2}$ satisface $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$. Si $\boldsymbol{\Sigma}$ es simétrica y definida positiva, tal es el caso de las matrices de covarianza, su raíz cuadrada es

$$\boldsymbol{\Sigma}^{1/2} = \boldsymbol{V}\,\mathrm{sqrt}(\boldsymbol{D})\boldsymbol{V}^\top, \tag{3.43}$$

siendo $\boldsymbol{V}$ la matriz con los autovectores de $\boldsymbol{\Sigma}$, $\boldsymbol{D}$ la matriz diagonal de autovalores y $\mathrm{sqrt}(\cdot)$ denota la operación raíz cuadrada aplicada a todos los elementos de su matriz argumento.

Para concluir este apartado, hemos de notar que en función del método elegido para el cálculo de los valores esperados $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, k_x, k_y]$ se obtendrá un estimador u otro. Así, a la técnica resultante de combinar los valores esperados de la ecuación (3.41) y el estimador VQMMSE de la ecuación (3.19) la denominaremos estimación VQMMSE con transformaciones basadas en la estadística conjunta (J-VQMMSE, *Joint* VQMMSE). Si los valores esperados empleados por el estimador son aquellos que devuelve la ecuación de blanqueo (3.42), entonces la técnica será conocida como estimación VQMMSE basada en transformaciones de blanqueo (W-VQMMSE, *Whitening-transformation based* VQMMSE). Cualesquiera que sea el estimador resultante, éste ya no podrá expresado en términos de una corrección aplicada al vector observado, sino que la forma final será una transformación afín del tipo $\boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}$.

3.2.4. Compensación basada en un esquema de múltiples modelos

Una suposición básica de las técnicas de compensación descritas en los apartados anteriores es el conocimiento a priori de las características de la fuente o fuentes que distorsionan la calidad de la voz. No obstante, como ya hemos comentado en varias ocasiones, raramente éste es el caso. En la práctica los sistemas de reconocimiento automático de voz deben enfrentarse a distorsiones cuyas características no son conocidas en principio y que son variables con el tiempo. Por ello, en este apartado estudiaremos

la forma de extender las técnicas de compensación propuestas para que éstas funcionen bajo distorsiones no contempladas en el entrenamiento.

La solución que aquí aportamos a este problema se basa en un esquema de reconocimiento con modelos múltiples [160, 277]. Según este esquema, en la fase de entrenamiento se puede disponer de cierto conocimiento sobre el conjunto de entornos acústicos a los que va a estar expuesto el sistema de reconocimiento. Notemos como $e = 1, \ldots, E$ el índice del entorno acústico en cuestión. Básicamente cada entorno modela la mezcla de un conjunto de fuentes de distorsión que degradan la calidad de la voz. En este sentido podremos modelar cada entorno e por la densidad de potencia espectral del ruido aditivo, la SNR de la señal de voz observada, la respuesta en frecuencia del canal y, opcionalmente, el tiempo de reverberación de la sala donde se desarrolla la interacción con el sistema. Conociendo todos estos parámetros, es posible recolectar datos de entrenamiento estéreo en ese ambiente determinado o generarlos artificialmente.

Las grabaciones estéreo adquiridas para cada entorno serán empleadas para entrenar el conjunto de diccionarios VQ requeridos por la técnica VQMMSE. Posteriormente, cada técnica de compensación derivará las transformaciones oportunas para cada entorno acústico (vectores de corrección o subregiones) que, como se estudió anteriormente, poseen la capacidad de mitigar la distorsión producida por el entorno en las características de la voz. Sin embargo, ahora dado el vector observado $\boldsymbol{y}$, el esquema de compensación basado en modelos múltiples calcula una estimación del vector de voz limpia $\hat{\boldsymbol{x}}^{(e)}$ para cada uno de los posibles entornos $e = 1, \ldots, E$. Estos valores se combinan de forma ponderada para obtener la una estima global como,

$$\hat{\boldsymbol{x}} = \sum_{e=1}^{E} P(e|\boldsymbol{y})\hat{\boldsymbol{x}}^{(e)}, \tag{3.44}$$

donde

$$P(e|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|e)P(e)}{\sum_{e'=1}^{E} p(\boldsymbol{y}|e')P(e')}. \tag{3.45}$$

Por simplicidad aquí supondremos que los entornos son equiprobables, luego, $P(e) = 1/E$. Al igual que otros trabajos (p.ej. [44, 160]), para modelar las probabilidades de observación $p(\boldsymbol{y}|e)$ aquí emplearemos GMMs entrenados de forma independiente para cada entorno,

$$p(\boldsymbol{y}|k) = \sum_{k_y^e}^{M_y^e} P(k_y^e)\mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_y^{(k_y^e)}, \boldsymbol{\Sigma}_y^{(k_y^e)}). \tag{3.46}$$

Con esto finalizamos la exposición del esquema de modelos múltiples. En comparación con otras técnicas de reconocimiento robusto, este esquema posee como principal ventaja la sencillez, tal y como puede deducirse de las de las ecuaciones anteriores.

Además constatamos la alta eficiencia del mismo: dado que los entornos acústicos son conocidos a priori, gran parte del cálculo invertido en las transformaciones de compensación puede precalcularse [160]. Por último, resaltaremos también el buen rendimiento mostrado por este esquema para los entornos acústicos considerados durante el entrenamiento, así como para ambientes (ruidos) similares [277].

3.2.5. Análisis de la complejidad computacional

En el título de esta sección hemos utilizado el atributo *eficiente* para calificar a las técnicas de compensación derivadas del estimador VQMMSE. En este apartado ratificaremos este calificativo basándonos en medidas objetivas de la complejidad computacional en tiempo de las distintas técnicas. En particular, emplearemos la notación $O(\boldsymbol{f}(\boldsymbol{y}))$ para especificar la cota superior de la complejidad asintótica de cada estimador $\boldsymbol{f}(\boldsymbol{y})$ en términos del número de operaciones requeridas por el mismo [18]. Además de los estimadores presentados en esta sección, el estudio de la eficiencia computacional también abarcará otras técnicas de compensación similares a las propuestas y que servirán de referencia. En concreto estas técnicas serán SPLICE [79] y MEMLIN [44] que, como vimos en la sección 2.2.3.4, difieren principalmente de las aquí propuestas en el uso de GMMs en el modelado de los espacios de características en lugar de emplear diccionarios VQ.

El primer paso que hemos de dar para obtener el orden de eficiencia de cada método es simplificar la expresión del estimador asociado. En este sentido podemos constatar que algunos de los términos que aparecen en las expresiones de los estimadores pueden agruparse y ser precalculados durante la fase de entrenamiento, con el consiguiente ahorro de operaciones que ello conlleva. En el apéndice A pueden consultarse las expresiones simplificadas para las técnicas derivadas del estimador VQMMSE, junto con las expresiones de las técnicas SPLICE y MEMLIN.

La tabla 3.1 muestra un resumen de la complejidad de cada técnica, así como la expresión simplificada asociada a cada estimador. En el caso de las técnicas descritas en el apartado 3.2.3 que contemplan el uso de subregiones, esto es, J-VQMMSE y W-VQMMSE, se presenta además la complejidad en función del tipo de matrices de covarianza que aparece en las ecuaciones (3.41) y (3.42): matrices de covarianza completas, diagonales o identidades. Para el caso de las matrices de covarianza diagonales se ha utilizado la notación '$\circ$' para denotar el producto vectorial elemento a elemento. Para simplificar el cálculo de la eficiencia asintótica, hemos ignorado el tiempo asociado al cómputo de las probabilidades a posteriori de cada región del diccionario, es decir, la evaluación de la gaussianas $p(\boldsymbol{y}|k_y)$ en SPLICE y MEMLIN, y la evaluación de la distancia de Mahalanobis $\mathrm{d}(\boldsymbol{y}, k_y)$ definida en (3.16) para el caso de las técnicas derivadas

Técnica	Expresión simplificada	Complejidad asintótica
SPLICE	$\hat{\boldsymbol{x}} = \boldsymbol{y} - \sum_{k_y=1}^{M_y} \boldsymbol{r}_{k_y} P(k_y\|\boldsymbol{y})$	$O(M_y D)$
MEMLIN	$\hat{\boldsymbol{x}} = \boldsymbol{y} - \sum_{k_y=1}^{M_y} \boldsymbol{r}_{k_y} P(k_y\|\boldsymbol{y})$	$O(M_y D)$
Q-VQMMSE	$\hat{\boldsymbol{x}} = \boldsymbol{c}$	$O(1)$
S-VQMMSE	$\hat{\boldsymbol{x}} = \boldsymbol{y} - \boldsymbol{r}$	$O(D)$
J-VQMMSE		
$\boldsymbol{\Sigma}$ identidad	$\hat{\boldsymbol{x}} = \boldsymbol{y} - \boldsymbol{r}$	$O(D)$
$\boldsymbol{\Sigma}$ diagonal	$\hat{\boldsymbol{x}} = \boldsymbol{a} \circ \boldsymbol{y} + \boldsymbol{b}$	$O(D)$
$\boldsymbol{\Sigma}$ completa	$\hat{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}$	$O(D^2)$
W-VQMMSE		
$\boldsymbol{\Sigma}$ identidad	$\hat{\boldsymbol{x}} = \boldsymbol{y} - \boldsymbol{r}$	$O(D)$
$\boldsymbol{\Sigma}$ diagonal	$\hat{\boldsymbol{x}} = \boldsymbol{a} \circ \boldsymbol{y} + \boldsymbol{b}$	$O(D)$
$\boldsymbol{\Sigma}$ completa	$\hat{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}$	$O(D^2)$

Tabla 3.1: Comparación de la eficiencia computacional entre diferentes técnicas de compensación basadas en datos estéreo.

del estimador VQMMSE. En este sentido debemos decir que el coste computacional de evaluar $p(\boldsymbol{y}|k_y)$ y $\mathrm{d}(\boldsymbol{y}, k_y)$ es muy similar, ya que en ambas expresiones aparece una forma cuadrática en la que interviene la matriz de precisión $\boldsymbol{\Sigma}^{(k_y)^{-1}}$.

Como se aprecia en la tabla 3.1, la complejidad de las técnicas estudiadas depende principalmente de dos factores: la dimensión del vector de características D y, para el caso de SPLICE y MEMLIN, del número de componentes M_y del GMM que modela el espacio de características distorsionadas. A simple vista se observa claramente que las técnicas propuestas basadas en diccionarios VQ son más eficientes que aquellas que emplean GMMs (supuesto que $D \ll M_y$ como suele ser habitual), debido a que en las primeras la celda más probable es la única que interviene en el cálculo del valor estimado, pudiendo, por tanto, precomputarse gran parte de las operaciones requeridas por los estimadores.

En la tabla también se aprecia cómo el modelado en subregiones no supone un coste computacional extra respecto al modelado VQ clásico. Así, por ejemplo, vemos que la técnica S-VQMMSE es igual de eficiente que la técnica W-VQMMSE con matrices de covarianza diagonales, pero como se verá en la sección 6.2.3.1, la última proporciona mejores resultados de reconocimiento que la primera. Se podría argumentar, no obstante, que un comportamiento similar al conseguido mediante el modelado en subregiones se puede lograr usando diccionarios VQ con M^2 celdas: puesto que cada celda VQ contiene a lo sumo M subregiones (M_y las celdas del espacio limpio y M_x las del espacio distorsionado), cada espacio de características contendrá M^2 subregiones en

total. No obstante, existen algunas diferencias sutiles entre ambos enfoques. En primer lugar tenemos que las celdas VQ son, por definición, disjuntas entre sí, mientras que las subregiones pueden solaparse entre ellas. Por otro lado tenemos que tras estimar el diccionario VQ con M^2 celdas usando el algoritmo de Linde-Buzo-Gray [178], éstas no tienen por qué coincidir con las M^2 subregiones de diccionario inicial. Finalmente podemos constatar que para el cálculo de la celda más probable k_y^* son necesarias M^2 comparaciones en el diccionario de M^2 celdas, mientras que en el diccionario con subregiones sólo se efectúan M comparaciones. En este sentido podemos concluir que el modelado en subregiones constituye una aproximación eficiente al diccionario con M^2 celdas.

Comparando los estimadores basados en GMMs (SPLICE y MEMLIN) con aquellos que emplean diccionarios VQ (Q-VQMMSE, S-VQMMSE, J-VQMMSE y W-VQMMSE), también se podría argumentar que los últimos pueden considerarse como una aproximación de los primeros en donde únicamente interviene la componente más probable del GMM $k_y^* = \text{argmax}_{1 \leq k_y \leq M_Y} P(k_y|y)$ en cada instante de tiempo. Aunque, en efecto, se pueden derivar dichas versiones de SPLICE y MEMLIN, existen diferencias entre éstas y las técnicas derivadas del estimador VQMMSE. En primer lugar nos encontramos que los algoritmos empleados para entrenar GMMs y diccionarios VQ difieren, luego los modelos obtenidos serán diferentes. En particular, el algoritmo de las k-medias [178] usado para entrenar cuantificadores vectoriales asume que la probabilidad a priori de las celdas es constante, no siendo así en el algoritmo EM [66]. Asimismo, las regiones que componen el diccionario VQ son disjuntas entre sí, mientras que las gaussianas de un GMM se solapan entre ellas. Como consecuencia de todas estas diferencias tenemos que la representación de los espacios de características obtenida en ambos casos diferirá y, por tanto, también diferirán los parámetros de las transformaciones de compensación.

3.3. Resumen

En este capítulo hemos sentado las bases para las técnicas de compensación de características de voz que propondremos a lo largo de esta tesis. Así, de entre los múltiples enfoques posibles que existen para mitigar la distorsión producida por el ruido, en este trabajo adoptaremos un marco estadístico bayesiano que usa información a priori sobre la distribución de las características de voz. Fruto de este marco se ha propuesto un estimador MMSE genérico para los vectores de características de voz, en donde se asume que estos se representan mediante modelos de mezclas de PDFs. Éste será el estimador que desarrollaremos a lo largo de esta memoria y que iremos particularizando en los sucesivos capítulos para cada caso en concreto.

Como primera aproximación, en la sección 3.2 el estimador MMSE se ha adaptado convenientemente para trabajar con grabaciones estéreo de señales de voz a distinta calidad. Usando estas grabaciones, durante la fase de entrenamiento se estiman un conjunto de diccionarios VQ para modelar los espacios de características limpias y distorsionadas. Asimismo, durante esta fase también se calculan una serie de transformaciones orientadas a compensar la degradación producida en la voz por cada ambiente acústico en particular. Como no hay ningún modelo de distorsión subyacente que restrinja la forma de dichas transformaciones, éstas pueden compensar, en teoría, cualquier tipo de distorsión que afecten a las grabaciones de voz ruidosa. Durante la fase de *test* el estimador VQMMSE aplica dichas transformaciones para realzar la calidad de la voz degradada, supuesto que las características de la distorsión que le afecta son conocidas.

Finalmente, en la sección 3.2.4 se ha presentado un esquema de compensación en donde no se requiere conocer a priori las características de las degradaciones. Este esquema, denominado de modelos múltiples, compensa la voz usando los parámetros aprendidos para una serie de entornos acústicos similares. En comparación con otras estrategias, este esquema destaca por su simplicidad.

CAPÍTULO 4

Compensación basada en un modelo de enmascaramiento

Aunque flexibles, las técnicas de compensación que usan grabaciones estéreo se fundamentan en suposiciones que restringen su uso a una variedad de situaciones limitada. En primer lugar porque asumen que dichas grabaciones existen o pueden adquirirse fácilmente, cosa que no tiene por qué ser cierta en ciertos entornos. En segundo lugar porque suponen que la variabilidad acústica presente durante el uso del sistema de reconocimiento va a ser limitada y se restringirá a un conjunto de ambientes acústicos conocidos. Sin embargo, su aplicación a situaciones de ruido no contemplado en el entrenamiento resulta en una merma considerable de la tasa de reconocimiento de palabras [43, 44, 122, 123]. Como tercera limitación de las técnicas basadas en grabaciones estéreo debemos comentar el gran volumen de datos que éstas necesitan. En efecto, al no usar modelos paramétricos de distorsión que restrinjan la forma de las transformaciones de compensación, la estimación robusta de estas transformaciones requiere de un gran volumen de datos.

A la vista de todas estas restricciones, en este capítulo adoptaremos una aproximación diferente para abordar el problema de reconocimiento robusto de voz en presencia de ruido. En concreto, aquí proponemos un conjunto de técnicas de compensación derivadas de modelos paramétricos de distorsión que definen relaciones analíticas entre la señal de voz observada con las fuentes que la componen (voz del locutor, ruido acústico, reverberación, etc.). De todas las fuentes de distorsión que pueden afectar a la señal de voz, en este capítulo nos centraremos en combatir únicamente el ruido aditivo, por ser éste el que afecta de forma más grave al rendimiento de los reconocedores [120]. Para combatir el resto de distorsiones, por ejemplo el ruido de canal, las técnicas aquí

descritas también pueden ser usadas en combinación con otras técnicas de normalización como CMN [20] o HEQ [238, 259] que reducen la discrepancia con los datos de entrenamiento.

Como decimos, los modelos de distorsión son funciones matemáticas que relacionan la señal de voz ruidosa con la señal de voz original y el ruido aditivo. En las siguientes secciones estudiaremos cómo, en el dominio logarítmico del banco de filtros Mel, el efecto del ruido aditivo sobre la voz se traduce en el enmascaramiento de ciertas regiones del espectro de la señal de voz original, mientras que otras apenas se ven alteradas. El espectro de la señal resultante será, por consiguiente, una mezcla de ambos espectros, componiéndose de regiones (parches) de los espectros de la voz y del ruido. La expresión analítica que modela este comportamiento y que denominaremos como modelo de enmascaramiento de la voz, nos permitirá desarrollar estimadores para las regiones enmascaradas del espectro de voz. En particular, en este capítulo estudiaremos dos técnicas diferentes: TGI y MMSR.

4.1. Modelo de enmascaramiento de la voz

Considérese el modelo de distorsión de los parámetros de la voz en el dominio log-Mel que aparece reflejado en la ecuación (2.9),

$$\boldsymbol{y} \approx \boldsymbol{x} + \boldsymbol{h} + \log(\mathbf{1} + e^{\boldsymbol{n}-\boldsymbol{x}-\boldsymbol{h}}). \tag{4.1}$$

Suponiendo nulo el ruido de canal, tenemos que

$$\boldsymbol{y} \approx \log(e^{\boldsymbol{x}} + e^{\boldsymbol{n}}), \tag{4.2}$$

que es el modelo que aparece recogido en la ecuación (2.11).

Definamos $\text{máx}(\boldsymbol{x}, \boldsymbol{n})$ y $\text{mín}(\boldsymbol{x}, \boldsymbol{n})$ como las operaciones de máximo y mínimo aplicadas sobre cada una de las componentes de sus dos vectores argumentos. En base a estas dos funciones podemos reescribir el modelo de distorsión anterior de la siguiente forma

$$\begin{aligned}\boldsymbol{y} &\approx \log(e^{\text{máx}(\boldsymbol{x},\boldsymbol{n})} + e^{\text{mín}(\boldsymbol{x},\boldsymbol{n})}) \\ &= \log\left(e^{\text{máx}(\boldsymbol{x},\boldsymbol{n})} \circ \left(\mathbf{1} + e^{\text{mín}(\boldsymbol{x},\boldsymbol{n})-\text{máx}(\boldsymbol{x},\boldsymbol{n})}\right)\right) \\ &= \text{máx}(\boldsymbol{x}, \boldsymbol{n}) + \underbrace{\log\left(\mathbf{1} + e^{\text{mín}(\boldsymbol{x},\boldsymbol{n})-\text{máx}(\boldsymbol{x},\boldsymbol{n})}\right)}_{\boldsymbol{\varepsilon}(\boldsymbol{x},\boldsymbol{n})}.\end{aligned} \tag{4.3}$$

El término $\boldsymbol{\varepsilon}(\boldsymbol{x}, \boldsymbol{n})$ de la ecuación anterior puede considerarse como un error residual de aproximación. En este sentido podemos ver que el error es máximo cuando la energía

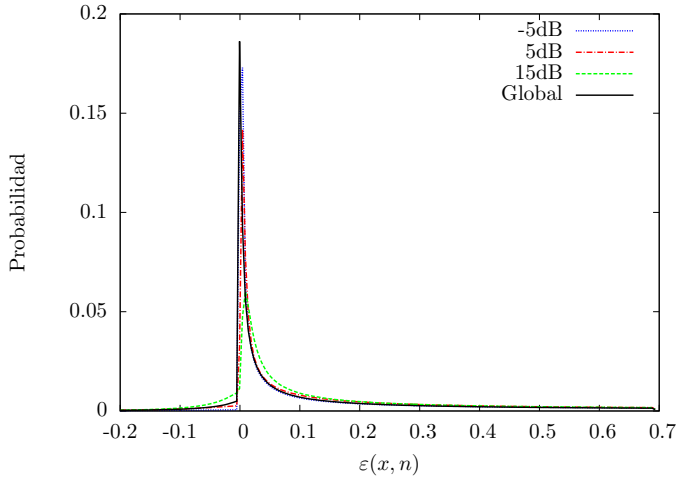

Figura 4.1: Histograma de $\varepsilon(x, n)$ para el conjunto de test A de la base de datos Aurora2.

del ruido iguala a la de la voz y, por tanto, $\varepsilon(x_i, n_i) = \log(2) \approx 0,69$ $(i = 1, \ldots, D)$. Un análisis más detallado de la estadística de este término se muestra en la figura 4.1, en la cual se presenta el histograma de $\varepsilon(x_i, n_i)$ calculado sobre todos los canales log-Mel para varios niveles de SNR del conjunto de test A de Aurora2 [141]. En esta figura también aparece reflejada la estadística global de este error calculada para todos los niveles de SNR presentes en esta base de datos. La magnitud de este error se ha calculado comparando el valor de la voz ruidosa en cada canal con el valor máximo entre voz y ruido (la energía del ruido en este caso se calcula a partir de las frases ruidosas y sus frases limpias correspondientes). Vemos que la mayor densidad de error se encuentra en torno a cero, decreciendo su probabilidad exponencialmente hasta llegar al valor máximo posible en $\log(2)$. En la gráfica también se aprecia la existencia de valores negativos de error debidos a la no consideración de la correlación entre las fases de voz y ruido en el modelo de distorsión de la ecuación (4.1). No obstante, se observa que la probabilidad de estos valores negativos es despreciable.

Del hecho anterior, y dado que además la magnitud relativa de $\boldsymbol{\varepsilon}(\boldsymbol{x}, \boldsymbol{n})$ en comparación con los valores de $\boldsymbol{x}$ y $\boldsymbol{n}$ es insignificante, podemos suponer nulo $\boldsymbol{\varepsilon}(\boldsymbol{x}, \boldsymbol{n})$ incurriendo en un error despreciable y, por tanto, simplificar la ecuación (4.3) como,

$$\boldsymbol{y} \approx \text{máx}(\boldsymbol{x}, \boldsymbol{n}). \tag{4.4}$$

Este modelo de distorsión, que fue propuesto inicialmente en [206, 267] en el contexto de descomposición de modelos (ver sección 2.2.2.2), se ha denominado en la literatura de múltiples formas: aproximación *log-max* [207, 233, 267], modelo MIXMAX (*Mixture*

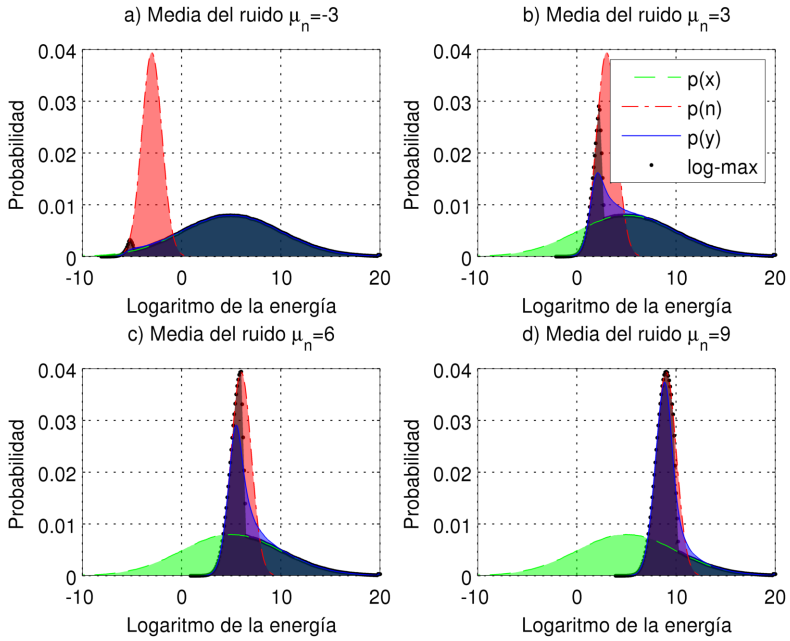

Figura 4.2: Aproximación *log-max* de la distribución $p(y)$ para distintos niveles de ruido.

of Maxima, mezcla de valores máximos) [206, 222, 229] o *modelo de enmascaramiento de la voz*[1] [90, 121, 125]. En esta tesis emplearemos indistintamente cualquiera de estos nombres, aunque preferiremos el último de ellos. La ventaja de este modelo frente a otras aproximaciones (p.ej. VTS [199]) es, como puede apreciarse, su sencillez. Asimismo en [222, 229] se demuestra que el modelo de enmascaramiento se corresponde con el valor esperado de la función de interacción exacta entre dos fuentes sonoras[2], bajo el supuesto de uniformidad estadística en las diferencias entre las fases de ambas fuentes. En el caso de una escena acústica compuesta por más de dos fuentes sonoras, el resultado anterior sigue siendo válido siempre y cuando, para una componente en frecuencia dada, una de las fuentes domine al resto.

[1] Se ha de notar que el término *enmascaramiento* empleado en este capítulo para describir la pérdida de cierta información de la señal de voz por la distorsión producida por el ruido, es una simplificación del proceso de enmascaramiento perceptivo que se produce en el oído humano entre dos o más sonidos simultáneos (en tiempo o frecuencia). En este sentido debemos decir que el enmascaramiento aquí considerado es independiente de la frecuencia y de las señales (voz y ruido) que entran en juego, dependiendo únicamente de la energía relativa entre las señales.

[2] La función de interacción o modelo de interacción es la expresión analítica que relaciona las fuentes sonoras latentes, $\boldsymbol{x}$ y $\boldsymbol{n}$ en nuestro caso, con el valor observado $\boldsymbol{y}$. Para el caso que nos ocupa, la función de interacción exacta coincide con el modelo de distorsión de la voz recogido en la ecuación (2.3), es decir, el modelo que tiene en cuenta la relación entre las fases de ambas fuentes, pero expresado en el dominio log-Mel.

La figura 4.2 muestra distintos ejemplos unidimensionales de las aproximaciones obtenidas para la distribución real de la voz distorsionada $p(y)$ considerando el modelo de enmascaramiento y varios niveles de ruido. En estos ejemplos se ha asumido que el ruido de canal es nulo y que las variables aleatorias que intervienen en el modelo (x, n, y) vienen expresadas en el dominio del banco de filtros logarítmico. Al igual que en la figura 2.9, consideramos que las distribuciones de probabilidad $p(x)$ y $p(n)$ son gaussianas: $p(x) = \mathcal{N}_x(\mu_x = 5, \sigma_x = 5)$, mientras que $p(n)$ cuenta con una desviación típica fija de $\sigma_n = 1$ y cuya media oscila entre -3 y 9 con incrementos de 3 unidades. La distribución teórica de $p(y)$ se ha calculado entonces usando el método de Montecarlo y el modelo de distorsión de la voz de la ecuación (2.11). Como se aprecia en la figura, la aproximación *log-max* proporciona estimaciones muy precisas tanto para niveles de SNR altos (gráfica a), como para SNRs bajas (gráfica d). Para SNRs intermedias (gráficas b y c), la aproximación *log-max* se aleja de la distribución teórica $p(y)$ debido a la no gaussianidad de ésta. Es interesante comparar las aproximaciones obtenidas por la técnica VTS en la figura 2.9 con las mostradas en la figura 4.2 para el modelo de enmascaramiento. Podemos comprobar que ambas aproximaciones al modelo de distorsión teórico son bastante precisas para niveles de SNR extremos (altos y bajos). Es en los casos intermedios, esto es, cuando la energía de la voz coincide aproximadamente con la del ruido, cuando ambas aproximaciones fallan: mientras que VTS no es capaz de modelar propiamente la no linealidad resultante a estas SNRs al aproximar $p(\boldsymbol{y})$ por una gaussiana, la aproximación *log-max* sí que resulta en una PDF no lineal, pero con un error mayor para valores de x y n próximos.

Volviendo a la expresión del modelo de enmascaramiento recogida en la ecuación (4.4), vemos que la distorsión producida por el ruido aditivo en la voz puede considerarse como un problema de enmascaramiento, ya que algunas de las características de la voz se ven enmascaradas por el ruido. Reordenando la observación como $\boldsymbol{y} = (\boldsymbol{y}_r, \boldsymbol{y}_u)$, llamamos *elementos fiables*, $\boldsymbol{y}_r$, a aquellos elementos de $\boldsymbol{y}$ donde domina la energía de la voz y, por tanto, ésta apenas ha sufrido alteración alguna por el ruido ($\boldsymbol{y}_r \approx \boldsymbol{x}_r$). Por otro lado, los *elementos perdidos o no fiables*, $\boldsymbol{y}_u$, del vector observado son aquellos en los que la energía del ruido es dominante y la voz se encuentra enmascarada. En este último caso, la única información que se dispone sobre los valores de voz enmascarados, $\boldsymbol{x}_u$, es el rango en el que estos se encuentra, el cual está acotado inferiormente por $-\infty$ (o cualquier otro umbral estimado a partir de los datos de entrenamiento) y superiormente por los propios valores observados $\boldsymbol{y}_u$.

El objetivo de este capítulo, como hemos mencionado anteriormente, es el planteamiento de distintas técnicas de reconstrucción (estimación) de los valores perdidos del espectro, dados los valores fiables observados. En el diseño de estas técnicas se han de

abordar los siguientes problemas:

- *Estimación de la fiabilidad de los elementos del vector observado*: consiste en identificar las zonas fiables y no fiables del espectro observado para, a partir de esta información, poder reconstruir posteriormente los elementos perdidos. En este capítulo comenzaremos proponiendo una técnica de reconstrucción que usa máscaras de segregación binarias (sección 4.2) e iremos avanzado hacia técnicas más robustas que emplean máscaras continuas. Finalmente en la sección 4.3 estudiaremos el caso de un estimador que nos permite usar estimas de ruido en el contexto del paradigma de datos perdidos y calcular la fiabilidad de los datos observados a partir de ellas.

- *Estimación de los elementos perdidos del espectro*: se trata del proceso de inferencia estadística mediante el cual se estima la energía de la voz en aquellos elementos del espectro identificados como perdidos por la máscara de segregación. Para esta tarea, aquí adaptaremos el estimador MMSE propuesto en la sección 3.1. No obstante, al contrario que las técnicas de compensación propuestas en la sección 3.2, en este capítulo se emplearán GMMs para el modelado de los espacios de características de la voz.

Aunque los hemos considerado como separados, los dos problemas anteriores están, de hecho, íntimamente ligados: como prueba tenemos la técnica SFD estudiada en la sección 2.2.5.3. Inspirados por esta filosofía, en la sección 4.3 estudiaremos un método iterativo basado en el algoritmo EM [66] para la estimación conjunta de los elementos perdidos del espectro y del modelo de ruido.

4.2. Reconstrucción espectral usando máscaras de segregación

4.2.1. Introducción

En la sección anterior hemos estudiado que la distorsión producida por el ruido acústico en los parámetros de la señal de voz puede interpretarse, de forma alternativa, como un problema de enmascaramiento. Tal y como aparece reflejado en la figura 2.12, esto conlleva que ciertas regiones del espectro de la voz van a estar enmascaradas por el ruido y viceversa. Bajo el supuesto de que los modelos acústicos del reconocedor se entrenan con voz limpia, el reconocimiento con señales de voz parcialmente enmascaradas producirá discrepancias que acarrearán una pérdida importante en el rendimiento

del sistema de reconocimiento. A fin de robustecer estos sistemas frente a tales efectos, en este apartado nos planteamos la estimación de la energía de la voz en las regiones donde ésta se encuentra enmascarada, de manera que al final del proceso de estimación se obtenga un espectrograma completo que pueda ser empleado por un reconocedor convencional (p.ej. pueden calcularse parámetros MFCC a partir del espectro estimado). En particular, de los dos problemas mencionados en la sección anterior, aquí nos centramos en estimar las regiones perdidas supuesto que la máscara de segregación que las identifica se encuentra disponible a priori.

A modo de ejemplo, la figura 4.3 muestra un diagrama del proceso de reconstrucción de regiones enmascaradas de una imagen. En lugar de espectros de voz, hemos optado por mostrar imágenes con objeto de simplificar la exposición del algoritmo de reconstrucción planteado. La extrapolación de la idea recogida en esta figura al caso de espectrogramas log-Mel es directa: ambos casos, imágenes y espectros de voz, se representan mediante matrices en las que ciertos elementos son observables y otros se ven enmascarados. De hecho, por extrapolación con las imágenes, en ciertas ocasiones nos referiremos a los elementos de un espectrograma como *píxeles*.

Volviendo a la figura, nos encontramos con una observación ruidosa que cuenta con varias zonas ocultas por ruido (las figuras geométricas que aparecen en la imagen). Para reconstruir el contenido de las zonas enmascaradas de la imagen, el primer paso consiste en estimar una máscara de segregación binaria que distinga las zonas pertenecientes al objeto de interés (identificadas por el color blanco) de las regiones pertenecientes al ruido (negro en la imagen). En este apartado, como se ha comentado antes, supondremos que esta máscara es conocida a priori. En la sección 2.2.5.5 se presentó un resumen de las distintas técnicas propuestas en la literatura para la estimación de las máscaras de segregación. Asimismo, en la sección 4.3.2 propondremos un método robusto para el cálculo de las máscaras de segregación de la voz.

Continuando con la figura 4.3, una vez conocida la máscara de segregación, tratamos de estimar las regiones de la imagen para las que la máscara toma el color negro. Existen múltiples formas de llevar a cabo este proceso. Por ejemplo, la reconstrucción puede tratarse de una simple interpolación de los *píxeles* presentes en la vecindad de la región a estimar [223]. En este apartado, no obstante, optaremos por la filosofía bayesiana en donde el estimador usa información a priori en forma de modelos de fuente sobre el objeto que se estima. En la figura esta información se representa mediante otras imágenes de los objetos que aparecen en la imagen (el oso y el árbol). En nuestro caso emplearemos modelos estadísticos que representen la distribución de los parámetros log-Mel en tiempo y frecuencia. A este respecto hay que señalar que, como no podría ser de otra manera, mejores modelos a priori redundarán en estimaciones más precisas

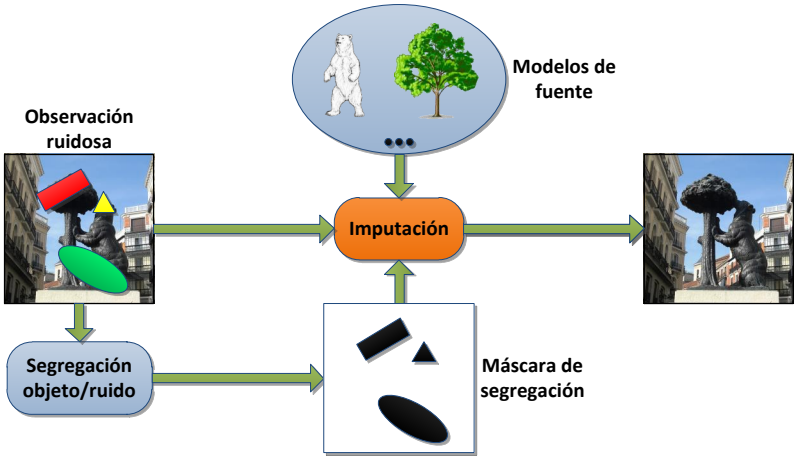

Figura 4.3: Analogía del proceso de estimación de las características perdidas del espectro.

de las características del espectro [184].

Usando estos modelos, las regiones enmascaradas de la imagen son imputadas[1] en base a cierto criterio de mínimo error. Como indicamos en el capítulo 3, en este trabajo adoptaremos el criterio MMSE, el cual nos permitirá explotar de forma eficiente la información recogida en los modelos a priori para calcular las estimas oportunas de las regiones enmascaradas. No obstante, hay que notar que, al contrario que para el caso de las imágenes, en el enmascaramiento de dos fuentes sonoras podemos explotar la información que nos proporcionan las observaciones ruidosas como cota superior para la estimación de la energía de la fuente enmascarada. Dicho de forma más simple, la aproximación *log-max* de la ecuación (4.4) nos indica que, en el enmascaramiento de dos fuentes sonoras, el valor de energía observado y constituye una cota superior para la energía de la voz enmascarada $x \in (-\infty, y]$, por lo que podemos explotar esta información durante la reconstrucción espectral para lograr estimas más precisas [57].

Finalmente el espectro reconstruido puede ser directamente usado por un reconocedor de voz estándar para descodificar el mensaje contenido en la señal de voz, siendo éste el objetivo último de la presente tesis. En tal caso, además del espectro estimado, el reconocedor puede explotar medidas calculadas durante el proceso de estimación que indiquen la bondad de la estimación efectuada. El cómputo de estas medidas y su explotación por el reconocedor se estudiarán en el capítulo 5. Frente al uso como técnica de compensación de características para reconocimiento robusto, una aplicación alternativa del algoritmo de reconstrucción que se ha esquematizado en la figura

[1] Por imputación entendemos sustituir un valor erróneo o perdido por una estimación más precisa del mismo.

4.3 es como técnica de realce de voz. Así, se podría emplear el algoritmo *overlap-add* (solapar-sumar) [130] para sintetizar una señal de voz en el dominio del tiempo a partir del espectro reconstruido y de la fase de la señal ruidosa original.

4.2.2. Imputación de los elementos perdidos del espectro

En este apartado procedemos a derivar la expresión correspondiente al estimador de las características perdidas del espectro en las señales de voz distorsionadas por ruido acústico. La técnica resultante jugará el mismo papel que el bloque de imputación que aparece en la figura 4.3, pero su aplicación estará orientada a la reconstrucción de espectros de voz, no a imágenes. Partiendo del criterio de estimación MMSE, la estima del vector se voz limpia se puede expresar como

$$\hat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, \Lambda] = \int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}, \Lambda) d\boldsymbol{x}. \tag{4.5}$$

El término Λ que aparece en la expresión del estimador se refiere a cualquier información a priori que pueda ser de utilidad para estimar $\hat{\boldsymbol{x}}$. En este sentido hemos constatado anteriormente que el conocimiento sobre la fiabilidad de cada elemento de $\boldsymbol{y}$ es una información indispensable de cara a la tarea de reconstrucción espectral. Es por ello que en este apartado supondremos que la fiabilidad de cada elemento de $\boldsymbol{y}$ es conocida a priori, y que ésta se representa mediante una máscara binaria $\boldsymbol{m}$ tal que,

$$m_i = \begin{cases} 1, & (x_i = y_i) \wedge (n_i \leq y_i) \\ 0, & (n_i = y_i) \wedge (x_i < y_i) \end{cases}, \tag{4.6}$$

siendo x_i y n_i las variables ocultas correspondientes a las energías log-Mel de la voz y el ruido, respectivamente, para el elemento $i = 1, \ldots, D$ del vector $\boldsymbol{y}$.

Para simplificar la notación, representaremos mediante $\boldsymbol{s}_r$ y $\boldsymbol{s}_u$ a los conjuntos que contienen los índices de los elementos fiables y no fiables, respectivamente, de la observación $\boldsymbol{y}$:

$$\boldsymbol{s}_r = \{i : m_i = 1\}, \tag{4.7}$$
$$\boldsymbol{s}_u = \{j : m_j = 0\}, \tag{4.8}$$

con $i, j \in \{1, \ldots, D\}$.

En base a estos conjuntos podemos segregar fácilmente la observación $\boldsymbol{y}$ en sus componentes fiables $\boldsymbol{y}_r$ y no fiables $\boldsymbol{y}_u$:

$$\boldsymbol{y}_r = \{y_i : i \in \boldsymbol{s}_r\}, \tag{4.9}$$
$$\boldsymbol{y}_u = \{y_j : j \in \boldsymbol{s}_u\}, \tag{4.10}$$

donde notaremos por R a la dimensión del vector $\boldsymbol{y}_r$ y por U a la del vector $\boldsymbol{y}_u$ por U, de tal forma que $D = R + U$.

Análogamente a lo realizado con la observación $\boldsymbol{y}$, el vector de voz limpia puede segregarse en elementos fiables y perdidos, $\boldsymbol{x} = (\boldsymbol{x}_r, \boldsymbol{x}_u)$. Recordemos que el modelo de enmascaramiento de la voz nos impone las siguientes restricciones: $\boldsymbol{x}_r = \boldsymbol{y}_r$ y $\boldsymbol{x}_u \leq \boldsymbol{y}_u$[1]. Usando estas definiciones, la probabilidad a posteriori $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{m})$ de la ecuación (4.5) se puede factorizar como el producto de los siguientes dos términos

$$p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{m}) \equiv p(\boldsymbol{x}_r, \boldsymbol{x}_u|\boldsymbol{y}_r, \boldsymbol{y}_u) = p(\boldsymbol{x}_r|\boldsymbol{y}_r)p(\boldsymbol{x}_u|\boldsymbol{y}_r, \boldsymbol{y}_u), \tag{4.11}$$

donde se asume que $\boldsymbol{x}_r$ es independiente de $\boldsymbol{y}_u$ dado $\boldsymbol{y}_r$.

Dado que los factores de la ecuación anterior actúan sobre elementos diferentes del vector $\boldsymbol{x}$, podemos fragmentar la ecuación (4.5) en dos estimadores diferentes: uno para $\boldsymbol{x}_r$ y otro para $\boldsymbol{x}_u$. Como no podía ser de otro modo, el estimador para $\boldsymbol{x}_r$ coincide con la propia observación $\boldsymbol{y}_r$:

$$\hat{\boldsymbol{x}}_r = \int \boldsymbol{x}_r p(\boldsymbol{x}_r|\boldsymbol{y}_r) d\boldsymbol{x}_r = \boldsymbol{y}_r, \tag{4.12}$$

ya que $\boldsymbol{x}_r = \boldsymbol{y}_r$ y, por tanto, $p(\boldsymbol{x}_r|\boldsymbol{y}_r) = \delta_{\boldsymbol{y}_r}(\boldsymbol{x}_r)$.

El cálculo del valor esperado $\boldsymbol{x}_u$, por otra parte, se torna más complejo. En primer lugar supondremos que las características de voz se modelan mediante un modelo de mezcla de gaussianas con M componentes,

$$p(\boldsymbol{x}) = \sum_{k=1}^{M} P(k)\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)}). \tag{4.13}$$

A partir de este modelo, la probabilidad $p(\boldsymbol{x}_u|\boldsymbol{y}_r, \boldsymbol{y}_u)$ se obtiene tras marginalizar $p(\boldsymbol{x}_u, k|\boldsymbol{y}_r, \boldsymbol{y}_u)$ respecto a la variable oculta k:

$$p(\boldsymbol{x}_u|\boldsymbol{y}_r, \boldsymbol{y}_u) = \sum_{k=1}^{M} p(\boldsymbol{x}_u|\boldsymbol{y}_r, \boldsymbol{y}_u, k)P(k|\boldsymbol{y}_r, \boldsymbol{y}_u). \tag{4.14}$$

Desarrollando el estimador MMSE de la ecuación (4.5) en base a la probabilidad condicional de la ecuación (4.14), nos queda el siguiente método de imputación de las características perdidas del espectro

$$\hat{\boldsymbol{x}}_u = \sum_{k=1}^{M} P(k|\boldsymbol{y}_r, \boldsymbol{y}_u) \underbrace{\int \boldsymbol{x}_u p(\boldsymbol{x}_u|\boldsymbol{y}_r, \boldsymbol{y}_u, k) d\boldsymbol{x}_u}_{\hat{\boldsymbol{x}}_u^{(k)}}. \tag{4.15}$$

[1] Las desigualdades se aplican por componentes.

Observamos que la expresión resultante puede considerarse como un caso particular del estimador MMSE basado en diccionario estudiado en la sección 3.1. En particular, la técnica de reconstrucción obtenida se deriva de este estimador suponiendo que: (i) los parámetros de voz limpia se modelan mediante un GMM y (ii) la información a priori que emplea el estimador es una máscara binaria que segrega la observación en componentes fiables y perdidas. Al igual que el estimador MMSE basado en diccionario, la ecuación (4.15) incluye una sumatoria de diferentes estimas parciales $\hat{\boldsymbol{x}}_u^{(k)}$, una por cada gaussiana $k = 1, \ldots, M$ del GMM, ponderadas por sus respectivas probabilidades a posteriori $P(k|\boldsymbol{y}_r, \boldsymbol{y}_u)$. Debemos resaltar cómo, en ambos términos, las estimas parciales y las probabilidades, aparecen explícitamente reflejados los subvectores de elementos fiables $\boldsymbol{y}_r$ y no fiables $\boldsymbol{y}_u$. Mientras que $\boldsymbol{y}_u$ impone una cota superior para $\boldsymbol{x}_u$, $\boldsymbol{y}_r$ restringe los valores que $\boldsymbol{x}_u$ puede tomar al condicionar la probabilidad a posteriori de las características perdidas y, de esta forma, poder explotar de forma precisa la correlación existente entre los distintos elementos.

Aplicando la regla de Bayes, la probabilidad $P(k|\boldsymbol{y}_r, \boldsymbol{y}_u)$ de la ecuación (4.15) se expresa como sigue

$$P(k|\boldsymbol{y}_r, \boldsymbol{y}_u) = \frac{p(\boldsymbol{y}_r, \boldsymbol{y}_u|k)P(k)}{\sum_{k'=1}^{M} p(\boldsymbol{y}_r, \boldsymbol{y}_u|k')P(k')}. \tag{4.16}$$

En la expresión anterior $P(k)$ $(k = 1, \ldots, M)$ son los pesos de las gaussianas del GMM (4.13). Por otro lado, la probabilidad de observación $p(\boldsymbol{y}_r, \boldsymbol{y}_u|k)$ se puede factorizar como el producto de los siguientes términos

$$p(\boldsymbol{y}_r, \boldsymbol{y}_u|k) = p(\boldsymbol{y}_r|k)p(\boldsymbol{y}_u|\boldsymbol{y}_r, k). \tag{4.17}$$

De los dos factores anteriores, $p(\boldsymbol{y}_r|k) = \mathcal{N}(\boldsymbol{y}_r; \boldsymbol{\mu}_r^{(k)}, \boldsymbol{\Sigma}_{rr}^{(k)})$ es directamente computable y equivale a la distribución marginal de los elementos fiables del vector. Los parámetros de esta distribución se calculan fácilmente a partir de la distribución original $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})$ y de la segmentación recogida en la máscara binaria,

$$\boldsymbol{\mu}^{(k)} = \begin{pmatrix} \boldsymbol{\mu}_r^k \\ \boldsymbol{\mu}_u^k \end{pmatrix}, \tag{4.18}$$

$$\boldsymbol{\Sigma}^{(k)} = \begin{pmatrix} \boldsymbol{\Sigma}_{rr}^k & \boldsymbol{\Sigma}_{ru}^k \\ \boldsymbol{\Sigma}_{ur}^k & \boldsymbol{\Sigma}_{uu}^k \end{pmatrix}. \tag{4.19}$$

Por otro lado, $p(\boldsymbol{y}_u|\boldsymbol{y}_r, k)$ se deriva de la PDF conjunta $p(\boldsymbol{x}_u, \boldsymbol{n}_u, \boldsymbol{y}_u|\boldsymbol{y}_r, k)$ ($\boldsymbol{n}_u$ representa el vector de energías log-Mel correspondientes al ruido para los elementos del vector $\boldsymbol{y}_u$) tras marginalizar esta última PDF respecto a las variables $\boldsymbol{x}_u$ y $\boldsymbol{n}_u$:

$$p(\boldsymbol{y}_u|\boldsymbol{y}_r, k) = \iint p(\boldsymbol{x}_u, \boldsymbol{n}_u, \boldsymbol{y}_u|\boldsymbol{y}_r, k) d\boldsymbol{x}_u d\boldsymbol{n}_u, \tag{4.20}$$

donde la integral múltiple se evalúa sobre todo el espacio $\mathbb{R}^{2U}$.

De nuevo podemos expresar la densidad de probabilidad $p(\boldsymbol{x}_u, \boldsymbol{n}_u, \boldsymbol{y}_u|\boldsymbol{y}_r, k)$ como producto de factores más simples

$$\begin{aligned} p(\boldsymbol{x}_u, \boldsymbol{n}_u, \boldsymbol{y}_u|\boldsymbol{y}_r, k) &= p(\boldsymbol{y}_u|\boldsymbol{x}_u, \boldsymbol{n}_u)p(\boldsymbol{x}_u, \boldsymbol{n}_u|\boldsymbol{y}_r, k) \\ &= p(\boldsymbol{y}_u|\boldsymbol{x}_u, \boldsymbol{n}_u)p(\boldsymbol{n}_u)p(\boldsymbol{x}_u|\boldsymbol{y}_r, k), \end{aligned} \tag{4.21}$$

obteniéndose la primera igualdad tras suponer que $\boldsymbol{y}_u$ es independiente del resto de términos dados $\boldsymbol{x}_u$ y $\boldsymbol{n}_u$.

De los tres términos en los que se factoriza la probabilidad de la ecuación (4.21), la probabilidad del ruido, $p(\boldsymbol{n}_u)$, es la única que desconocemos. No obstante, más adelante comprobaremos cómo este término no tiene repercusión alguna en el cálculo de la probabilidad a posteriori de la ecuación (4.16) por actuar como una constante. La probabilidad condicional $p(\boldsymbol{y}_u|\boldsymbol{x}_u, \boldsymbol{n}_u)$, por otra parte, viene dada por

$$p(\boldsymbol{y}_u|\boldsymbol{x}_u, \boldsymbol{n}_u) = \frac{p(\boldsymbol{x}_u, \boldsymbol{n}_u|\boldsymbol{y}_u)p(\boldsymbol{y}_u)}{p(\boldsymbol{x}_u)p(\boldsymbol{n}_u)}. \tag{4.22}$$

Según el modelo de enmascaramiento de la ecuación (4.4) tenemos las siguientes desigualdades $\boldsymbol{n}_u = \boldsymbol{y}_u$ y $\boldsymbol{x}_u < \boldsymbol{y}_u$, luego

$$\begin{aligned} p(\boldsymbol{x}_u, \boldsymbol{n}_u|\boldsymbol{y}_u) &= \overbrace{\delta_{\boldsymbol{y}_u}(\boldsymbol{n}_u)}^{p(\boldsymbol{n}_u|\boldsymbol{y}_u)} \overbrace{\kappa\, p(\boldsymbol{x}_u)\mathbb{1}_{\boldsymbol{x}_u<\boldsymbol{y}_u}}^{p(\boldsymbol{x}_u|\boldsymbol{y}_u)} \\ &\propto \delta_{\boldsymbol{y}_u}(\boldsymbol{n}_u)p(\boldsymbol{x}_u)\mathbb{1}_{\boldsymbol{x}_u<\boldsymbol{y}_u}, \end{aligned} \tag{4.23}$$

donde $\mathbb{1}_{\boldsymbol{x}_u<\boldsymbol{y}_u}$ denota la función indicatriz, esto es,

$$\mathbb{1}_{\boldsymbol{x}_u<\boldsymbol{y}_u} = \begin{cases} 1 & \text{si } x_{u,i} < y_{u,i},\ i \in [1, U] \\ 0 & \text{en otro caso} \end{cases} \tag{4.24}$$

y κ es una constante de normalización para forzar a que la integral de $p(\boldsymbol{x}_u)$ sea unitaria sobre el intervalo definido por la función indicatriz.

El único término de la ecuación (4.21) que aún no hemos considerado es $p(\boldsymbol{x}_u|\boldsymbol{y}_r, k)$, esto es, la probabilidad condicional para cada gaussiana del GMM de los datos perdidos dados los observados. Puesto que la distribución original (sin segmentar) es gaussiana, se puede demostrar que $p(\boldsymbol{x}_u|\boldsymbol{y}_r, k) = \mathcal{N}(\boldsymbol{x}_u; \boldsymbol{\mu}_{u|r}^{(k)}, \boldsymbol{\Sigma}_{u|r}^{(k)})$ es también una distribución normal y que sus parámetros vienen dados por [211]:

$$\boldsymbol{\mu}_{u|r}^{(k)} = \boldsymbol{\mu}_u^{(k)} + \boldsymbol{\Sigma}_{ur}^{(k)}\boldsymbol{\Sigma}_{rr}^{(k)^{-1}}\left(\boldsymbol{y}_r - \boldsymbol{\mu}_r^{(k)}\right), \tag{4.25}$$

$$\boldsymbol{\Sigma}_{u|r}^{(k)} = \boldsymbol{\Sigma}_{uu}^{(k)} - \boldsymbol{\Sigma}_{ur}^{(k)}\boldsymbol{\Sigma}_{rr}^{(k)^{-1}}\boldsymbol{\Sigma}_{ru}^{(k)}. \tag{4.26}$$

Retomando de nuevo el cálculo de la probabilidad $p(\boldsymbol{y}_u|\boldsymbol{y}_r, k)$ definida en (4.20), la expresión final de ésta se deriva de las ecuaciones (4.21), (4.22) y (4.23),

$$\begin{aligned} p(\boldsymbol{y}_u|\boldsymbol{y}_r, k) &= \iint p(\boldsymbol{x}_u, \boldsymbol{n}_u, \boldsymbol{y}_u|\boldsymbol{y}_r, k) d\boldsymbol{x}_u d\boldsymbol{n}_u \\ &= \iint p(\boldsymbol{y}_u|\boldsymbol{x}_u, \boldsymbol{n}_u) p(\boldsymbol{n}_u) p(\boldsymbol{x}_u|\boldsymbol{y}_r, k) d\boldsymbol{x}_u d\boldsymbol{n}_u \\ &= \iint \frac{p(\boldsymbol{x}_u, \boldsymbol{n}_u|\boldsymbol{y}_u) p(\boldsymbol{y}_u)}{p(\boldsymbol{x}_u) p(\boldsymbol{n}_u)} p(\boldsymbol{n}_u) p(\boldsymbol{x}_u|\boldsymbol{y}_r, k) d\boldsymbol{x}_u d\boldsymbol{n}_u \\ &= p(\boldsymbol{y}_u) \iint \frac{\delta_{\boldsymbol{y}_u}(\boldsymbol{n}_u) p(\boldsymbol{x}_u) \mathbb{1}_{\boldsymbol{x}_u < \boldsymbol{y}_u}}{p(\boldsymbol{x}_u)} p(\boldsymbol{x}_u|\boldsymbol{y}_r, k) d\boldsymbol{x}_u d\boldsymbol{n}_u \\ &= p(\boldsymbol{y}_u) \iint \delta_{\boldsymbol{y}_u}(\boldsymbol{n}_u) \mathbb{1}_{\boldsymbol{x}_u < \boldsymbol{y}_u} p(\boldsymbol{x}_u|\boldsymbol{y}_r, k) d\boldsymbol{x}_u d\boldsymbol{n}_u. \end{aligned} \tag{4.27}$$

Puesto que $\boldsymbol{x}_u$ y $\boldsymbol{n}_u$ son independientes, podemos expresar la integral múltiple anterior como el producto de dos integrales independientes, una para cada variable. Asimismo observamos que $p(\boldsymbol{y}_u)$ es constante, luego puede obviarse del cálculo. Con estos dos observaciones, la expresión final para el cómputo de $p(\boldsymbol{y}_u|\boldsymbol{y}_r, k)$ es

$$\begin{aligned} p(\boldsymbol{y}_u|\boldsymbol{y}_r, k) &\propto \iint \delta_{\boldsymbol{y}_u}(\boldsymbol{n}_u) p(\boldsymbol{x}_u|\boldsymbol{y}_r, k) \mathbb{1}_{\boldsymbol{x}_u < \boldsymbol{y}_u} d\boldsymbol{x}_u d\boldsymbol{n}_u \\ &= \underbrace{\int_{-\infty}^{\infty} \delta_{\boldsymbol{y}_u}(\boldsymbol{n}_u) d\boldsymbol{n}_u}_{=1} \times \int_{-\infty}^{\infty} p(\boldsymbol{x}_u|\boldsymbol{y}_r, k) \mathbb{1}_{\boldsymbol{x}_u < \boldsymbol{y}_u} d\boldsymbol{x}_u \\ &= \int_{-\infty}^{\boldsymbol{y}_u} p(\boldsymbol{x}_u|\boldsymbol{y}_r, k) d\boldsymbol{x}_u = \int_{-\infty}^{\boldsymbol{y}_u} \mathcal{N}(\boldsymbol{x}_u; \boldsymbol{\mu}_{u|r}^{(k)}, \boldsymbol{\Sigma}_{u|r}^{(k)}) d\boldsymbol{x}_u \end{aligned} \tag{4.28}$$

donde la integral se evalúa sobre el hipervolumen definido por $(-\infty, y_{u,1}] \times (-\infty, y_{u,2}] \times \ldots \times (-\infty, y_{u,U}]$.

A pesar de haber obtenido una expresión analítica para el cálculo de la probabilidad $p(\boldsymbol{y}_u|\boldsymbol{y}_r, k)$, nos encontramos que la integral múltiple de la ecuación (4.28) no tiene solución analítica para matrices de covarianza no diagonales. Para salvar este obstáculo, aquí asumiremos que $\boldsymbol{\Sigma}_{u|r}^{(k)}$ es diagonal, es decir, sólo retenemos los elementos de la diagonal principal de esta matriz. Bajo esta aproximación, $p(\boldsymbol{y}_u|\boldsymbol{y}_r, k)$ se define como el siguiente producto

$$\begin{aligned} p(\boldsymbol{y}_u|\boldsymbol{y}_r, k) &\approx \prod_{i=1}^{U} \int_{-\infty}^{y_{u,i}} \mathcal{N}\left(x_u; \mu_{u|r,i}^{(k)}, \sigma_{u|r,i}^{(k)}\right) dx_u \\ &= \prod_{i=1}^{U} \Phi\left(\overline{y}_{u,i}^{(k)}\right), \end{aligned} \tag{4.29}$$

donde Φ denota la función de distribución normal acumulada,

$$\Phi(y) = \int_{-\infty}^{y} \mathcal{N}(x) dx \tag{4.30}$$

e $\overline{y}_{u,i}^{(k)}$ es la variable $y_{u,i}$ tipificada,

$$\overline{y}_{u,i}^{(k)} = \frac{y_{u,i} - \mu_{u|r,i}^{(k)}}{\sigma_{u|r,i}^{(k)}}. \tag{4.31}$$

Recapitulando nuestro objetivo inicial, las probabilidades $P(k|\boldsymbol{y}_r, \boldsymbol{y}_u)$ requeridas por el estimador MMSE de la ecuación (4.15) dependen de las probabilidades de observación $p(\boldsymbol{y}_r, \boldsymbol{y}_u|k)$ recogidas en (4.17). Todo el desarrollo matemático previo nos conduce a que estas últimas probabilidades se aproximan mediante el siguiente producto de probabilidades de elementos fiables y no fiables

$$p(\boldsymbol{y}_r, \boldsymbol{y}_u|k) = p(\boldsymbol{y}_r|k)p(\boldsymbol{y}_u|\boldsymbol{y}_r, k) \approx \mathcal{N}(\boldsymbol{y}_r; \boldsymbol{\mu}_r^{(k)}, \boldsymbol{\Sigma}_{rr}^{(k)}) \prod_{i=1}^{U} \Phi(\overline{y}_{u,i}^{(k)}), \tag{4.32}$$

donde $\mathcal{N}(\boldsymbol{y}_r; \boldsymbol{\mu}_r^{(k)}, \boldsymbol{\Sigma}_{rr}^{(k)})$ será una distribución gaussiana no necesariamente diagonal.

Una vez que hemos derivado las expresiones oportunas para el cálculo de las probabilidades a posteriori $P(k|\boldsymbol{y}_r, \boldsymbol{y}_u)$, nuestros pasos se encaminan al cómputo de las estimas parciales $\hat{\boldsymbol{x}}_u^{(k)}$ presentes en el estimador MMSE. Tal y como se refleja en la ecuación (4.15), estas estimas equivalen al siguiente valor esperado

$$\hat{\boldsymbol{x}}_u^{(k)} = \int \boldsymbol{x}_u p(\boldsymbol{x}_u|\boldsymbol{y}_r, \boldsymbol{y}_u, k) d\boldsymbol{x}_u. \tag{4.33}$$

Para el cómputo de $\hat{\boldsymbol{x}}_u^{(k)}$ en primer lugar hemos de expresar $p(\boldsymbol{x}_u|\boldsymbol{y}_r, \boldsymbol{y}_u, k)$ en una forma más conveniente. Así constatamos que esta probabilidad puede verse como la distribución marginal de $p(\boldsymbol{x}_u, \boldsymbol{n}_u|\boldsymbol{y}_r, \boldsymbol{y}_u, k)$ respecto a $\boldsymbol{n}_u$,

$$p(\boldsymbol{x}_u|\boldsymbol{y}_r, \boldsymbol{y}_u, k) = \int p(\boldsymbol{x}_u, \boldsymbol{n}_u|\boldsymbol{y}_r, \boldsymbol{y}_u, k) d\boldsymbol{n}_u. \tag{4.34}$$

Entonces $p(\boldsymbol{x}_u, \boldsymbol{n}_u|\boldsymbol{y}_r, \boldsymbol{y}_u, k)$ se factoriza como el siguiente producto

$$\begin{aligned} p(\boldsymbol{x}_u, \boldsymbol{n}_u|\boldsymbol{y}_r, \boldsymbol{y}_u, k) &= p(\boldsymbol{n}_u|\boldsymbol{y}_u)p(\boldsymbol{x}_u|\boldsymbol{y}_r, \boldsymbol{y}_u, k) \\ &= \delta_{\boldsymbol{y}_u}(\boldsymbol{n}_u)\frac{p(\boldsymbol{x}_u|\boldsymbol{y}_r, k)\mathbb{1}_{\boldsymbol{x}_u < \boldsymbol{y}_u}}{P(\boldsymbol{x}_u < \boldsymbol{y}_u|\boldsymbol{y}_r, k)}, \end{aligned} \tag{4.35}$$

siendo $P(\boldsymbol{x}_u < \boldsymbol{y}_u|\boldsymbol{y}_r, k)$ la probabilidad acumulada que resulta de integrar $p(\boldsymbol{x}_u|\boldsymbol{y}_r, k)$ en el hipervolumen definido por $(-\infty, y_{u,1}] \times (-\infty, y_{u,2}] \times \ldots \times (-\infty, y_{u,U}]$.

Usando las ecuaciones (4.34) y (4.35) en (4.33), las estimaciones parciales $\hat{\boldsymbol{x}}_u^{(k)}$ para cada gaussiana $k = 1, \ldots, M$ del GMM se calculan de la siguiente forma

$$\begin{aligned}
\hat{\boldsymbol{x}}_u^{(k)} &= \iint \boldsymbol{x}_u p(\boldsymbol{x}_u, \boldsymbol{n}_u | \boldsymbol{y}_r, \boldsymbol{y}_u, k) d\boldsymbol{x}_u d\boldsymbol{x}_u \\
&= \iint \boldsymbol{x}_u \left(\delta_{\boldsymbol{y}_u}(\boldsymbol{n}_u) \frac{p(\boldsymbol{x}_u | \boldsymbol{y}_r, k) \mathbb{1}_{\boldsymbol{x}_u < \boldsymbol{y}_u}}{P(\boldsymbol{x}_u < \boldsymbol{y}_u | \boldsymbol{y}_r, k)} \right) d\boldsymbol{x}_u d\boldsymbol{x}_u \\
&= \frac{1}{P(\boldsymbol{x}_u < \boldsymbol{y}_u | \boldsymbol{y}_r, k)} \int_{-\infty}^{\infty} \boldsymbol{x}_u p(\boldsymbol{x}_u | \boldsymbol{y}_r, k) \mathbb{1}_{\boldsymbol{x}_u < \boldsymbol{y}_u} d\boldsymbol{x}_u \\
&= \frac{1}{\Phi(\boldsymbol{y}_u; \boldsymbol{\mu}_{u|r}^{(k)}, \boldsymbol{\Sigma}_{u|r}^{(k)})} \int_{-\infty}^{\boldsymbol{y}_u} \boldsymbol{x}_u \mathcal{N}(\boldsymbol{x}_u; \boldsymbol{\mu}_{u|r}^{(k)}, \boldsymbol{\Sigma}_{u|r}^{(k)}) d\boldsymbol{x}_u.
\end{aligned} \tag{4.36}$$

De nuevo nos encontramos que la integral anterior no presenta una solución analítica para distribuciones cuya matriz de covarianza no sea diagonal, así que volvemos a suponer independencia estadística entre los distintos elementos ($i = 1, \ldots, U$) del vector $\hat{\boldsymbol{x}}_u$ a estimar al igual que se ha hecho en la ecuación (4.29):

$$\hat{x}_{u,i}^{(k)} \approx \frac{1}{\Phi(y_{u,i}; \mu_{u|r,i}^{(k)}, \sigma_{u|r,i}^{(k)})} \int_{-\infty}^{y_{u,i}} x_u \mathcal{N}(x_u; \mu_{u|r,i}^{(k)}, \sigma_{u|r,i}^{(k)}) dx_u. \tag{4.37}$$

La ecuación anterior se corresponde con la media de una distribución normal truncada a la derecha[1] por la observación $y_{u,i}$, por lo que, según la ecuación (B.12), el valor final de $\hat{x}_{u,i}^{(k)}$ viene dado por

$$\hat{x}_{u,i}^{(k)} \approx \mu_{u|r,i}^{(k)} - \sigma_{u|r,i}^{(k)} \frac{\mathcal{N}(\overline{y}_{u,i}^{(k)})}{\Phi(\overline{y}_{u,i}^{(k)})}. \tag{4.38}$$

Con esto finaliza la derivación de la técnica de imputación basada en el modelo de enmascaramiento de la voz. A modo de resumen, a continuación presentamos el algoritmo de reconstrucción íntegro que hemos desarrollado a lo largo de este apartado, el cual denominaremos de aquí en adelante como método de imputación basado en gaussianas truncadas (TGI, *Truncated-Gaussian based Imputation*):

- **Entrada:** espectro log-Mel de voz distorsionada $(\boldsymbol{y}(1), \ldots, \boldsymbol{y}(T))$, máscara de segregación binaria $(\boldsymbol{m}(1), \ldots, \boldsymbol{m}(T))$ y GMM de voz limpia $\mathcal{M}_x$.
- **Salida:** espectro log-Mel de voz estimada $(\hat{\boldsymbol{x}}(1), \ldots, \hat{\boldsymbol{x}}(T))$.

1. Para cada instante de tiempo $t = 1, \ldots, T$

   a. $\hat{\boldsymbol{x}}_r(t) = \boldsymbol{y}_r(t)$.

[1] En el apéndice B se hace una breve revisión de esta distribución de probabilidad.

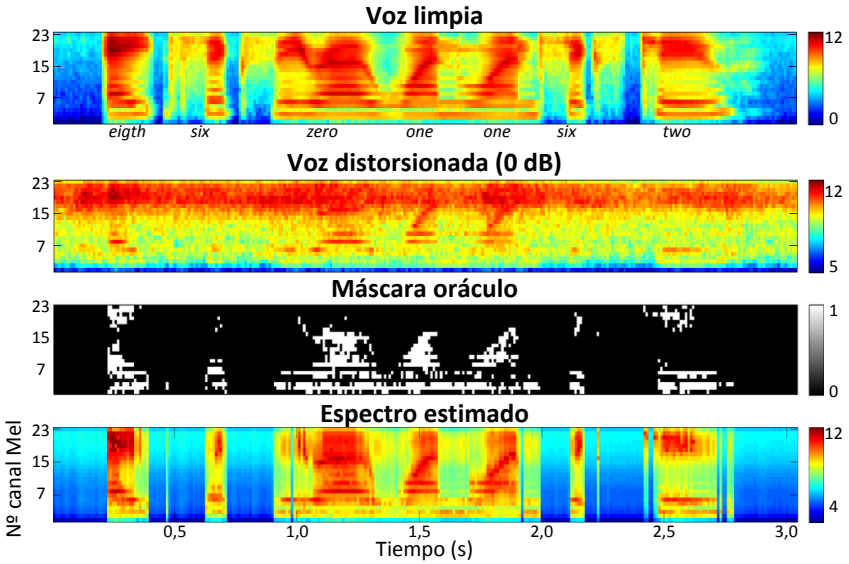

Figura 4.4: Ejemplo de reconstrucción espectral efectuada por la técnica de imputación TGI.

b. Calcular la probabilidad a posteriori $P(k|\boldsymbol{y}_r, \boldsymbol{y}_u)$ para cada gaussiana $k = 1, \ldots, M$ del GMM usando las ecuaciones (4.16) y (4.32).

c. Calcular las estimaciones parciales $\hat{\boldsymbol{x}}_u^{(k)}$ para $k = 1, \ldots, M$ usando la ecuación (4.37) para $k = 1, \ldots, M$.

d. Combinar las estimas parciales usando (4.15) para obtener $\hat{\boldsymbol{x}}_u(t)$.

Finalmente, y concluyendo esta sección, en la figura (4.4) presentamos varios espectrogramas log-Mel para ejemplificar el proceso de reconstrucción espectral llevado a cabo por la técnica TGI que acabamos de presentar. Los espectros se corresponden a distintas versiones de la elocución en inglés *eigth six zero one one six two* (ocho seis cero uno uno seis dos) extraída de la base de datos Aurora2 [141]. El primer espectro es el correspondiente a la señal de voz sin distorsionar. El segundo espectro se obtiene tras sumar ruido de tipo *subway* a la señal de voz anterior. El nivel de SNR elegido, tal y como aparece en la figura, es 0 dB. A partir de las dos señales anteriores es posible calcular la SNR local para cada elemento del espectro distorsionado. Esta SNR ideal, sin sesgo, es la empleada para calcular la máscara de segregación oráculo que se muestra en la figura. En este caso el umbral usado para binarizar las SNR anteriores ha sido de 7 dB. Como se puede apreciar, la máscara captura fielmente la fiabilidad de cada elemento del espectro, distinguiendo aquellos en los que domina la energía de la voz (identificados mediante el color blanco en la imagen), de aquellos completamen-

te afectados por el ruido (color negro). A partir del espectro de voz distorsionada, la máscara de segregación y un modelo estadístico de voz (GMM con 256 gaussianas), la técnica TGI estima la energía de las características perdidas de la voz, dando como resultado el espectro estimado que aparece en la figura. En comparación con el espectro original distorsionado, el espectro reconstruido es más similar a la señal de voz original sin distorsionar, lo que permite prever una mejora significativa en la precisión de reconocimiento.

4.3. Reconstrucción espectral usando modelos de ruido

Desde el punto de vista del modelo de enmascaramiento de la ecuación (4.4), los elementos del espectrograma de una señal de voz distorsionada pueden identificarse como fiables (dominados por la energía de la voz) y no fiables (domina la energía del ruido). Acabamos de ver en la figura 4.4 que el resultado del proceso de reconstrucción espectral llevado a cabo por las técnicas de imputación cuando éstas trabajan en condiciones ideales, es decir, cuando emplean máscaras de segregación oráculo, es comparable al espectro original de la voz sin distorsionar. En el capítulo 6 comprobaremos que, en efecto, el rendimiento alcanzado por el reconocedor de voz con ambos espectros es bastante similar para niveles de SNR intermedios y altos. Este resultado avala, por un lado, la validez de la aproximación *log-max* en la que se fundamenta el modelo de enmascaramiento y, por otro, la viabilidad del reconocimiento/reconstrucción de espectros de voz incompletos.

A pesar del buen rendimiento ofrecido por las técnicas de reconstrucción cuando estas operan sobre máscaras oráculo, experimentalmente se ha comprobado que estas técnicas son muy sensibles a errores en las máscaras de segregación [121, 125, 126, 127, 224]. Dado que la estimación de estas máscaras suele realizarse, en la mayoría de las ocasiones, mediante la estimación de la SNR local y su posterior umbralización, los errores en la estimación de las máscaras pueden deberse a dos factores: una decisión incorrecta del umbral o una mala estimación del ruido acústico. La solución al primer problema es relativamente fácil: basta con elegir un umbral de SNR conservador que considere como no fiables a un alto porcentaje de elementos del espectro. La justificación subyacente de esta solución se debe a que los falsos positivos (ruido considerado como voz) afectan de forma más directa al rendimiento de las técnicas de reconstrucción (o, de forma equivalente, al reconocimiento mediante marginalización) que los falsos negativos (voz considerada como ruido) [249]. La solución al segundo problema, la mala

estimación de ruido, se torna más compleja. Así, como ejemplo, la mayoría de técnicas de estimación de ruido tienden a estimar incorrectamente la densidad de potencia espectral de los ruidos no estacionarios (p.ej. ruidos espontáneos de alta energía), lo que conlleva un gran número de errores (falsos positivos) en las máscaras de segregación.

A pesar de la dificultad práctica de una buena estimación del espectro del ruido, una alternativa factible que ha demostrado ser de gran utilidad de cara a robustecer las técnicas basadas en el paradigma de datos perdidos es el uso de máscaras continuas en oposición a las máscaras binarias [31, 225]. En lugar de contener decisiones de fiabilidad binarias $m \in \{0, 1\}$, las máscaras continuas consideran un grado de fiabilidad continuo para cada uno de los elementos del espectro; correspondiéndose esta fiabilidad con la probabilidad de no enmascaramiento $m = P(x \geq n)$. Al acarrear una decisión de fiabilidad suave, los errores de estimación no afectan de forma tan directa a la precisión del proceso de reconstrucción espectral, conduciendo ello a mejoras relativas (respecto al uso de máscaras binarias) de hasta el 25 % en precisión de reconocimiento [90, 225].

Frente al uso de máscaras de segregación, en este apartado proponemos incorporar las estimaciones de ruido directamente en el estimador de voz. Como se comprobará más adelante, esta forma de proceder proporciona una mayor flexibilidad por varias razones. En primer lugar, nos permite el uso de descripciones probabilísticas del ruido presente en la señal de voz. En particular, aquí modelaremos el ruido presente en cada frase de evaluación mediante un modelo de mezcla de gaussianas, siendo este modelo una representación más robusta de la distribución de los ruidos no estacionarios. De hecho, el propio modelo de enmascaramiento de la voz nos proporcionará un marco matemático elegante para la estimación del modelo de ruido. Este aspecto será tratado en la sección 4.3.3. Otra ventaja del algoritmo de reconstrucción propuesto es que, como consecuencia del propio proceso de estimación de las características de voz, éste proporciona medidas de fiabilidad para cada uno de los elementos del espectro. En este sentido, el algoritmo propuesto podría considerarse de forma alternativa como un método robusto de estimación de máscaras de segregación continuas.

La figura 4.5 muestra el diagrama de bloques del marco unificado de estimación que proponemos en esta sección. Primeramente un método iterativo basado en el algoritmo EM [66] se encarga de estimar el modelo de ruido (GMM) usando un criterio de máxima verosimilitud (ML). Las entradas a este algoritmo, tal y como se refleja en el diagrama, son dos: el espectro de voz ruidosa y un modelo probabilístico de las características de voz expresadas en el dominio log-Mel (también un GMM).

El GMM de ruido ajustado a los datos observados junto con el GMM de voz y la señal de voz distorsionada son empleados posteriormente por un estimador MMSE para reconstruir las regiones del espectro perdidas. Al contrario que la técnica TGI

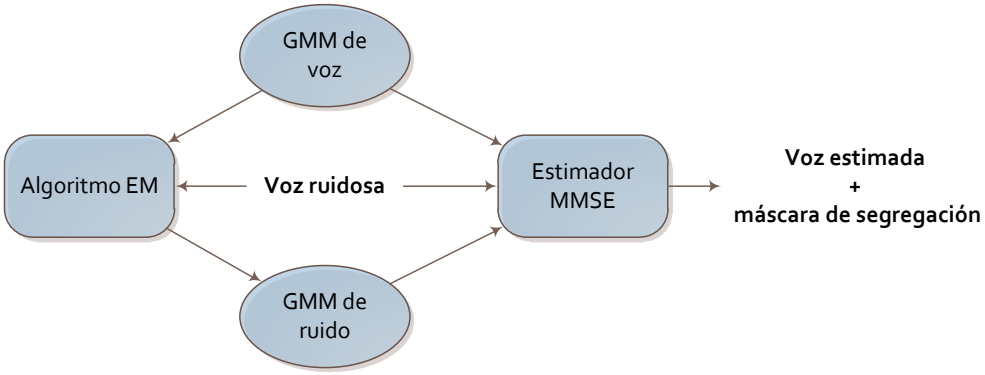

Figura 4.5: Esquema del marco unificado de estimación basado en el modelo de enmascaramiento de la voz.

estudiada en la sección anterior, el estimador de voz que aquí proponemos desconoce la segmentación del espectro en elementos fiables y perdidos, usando en cambio descripciones probabilísticas de la fiabilidad de cada elemento. Finalmente, como resultado de todo el proceso de estimación, se obtiene una versión realzada del espectro de la señal de voz y, opcionalmente, una máscara de segregación con la fiabilidad de cada elemento del espectro original.

Visto el esquema general de esta sección, en los siguientes apartados procederemos a derivar matemáticamente cada uno de los bloques representados en la figura 4.5. Empezaremos presentando el desarrollo matemático del estimador de las características de voz en el siguiente apartado. A esto seguirá, en la sección 4.3.2, el cómputo de las máscaras de segregación continuas basadas en el modelo de enmascaramiento. Continuaremos en la sección 4.3.3 con la derivación del algoritmo iterativo para la estimación del modelo de ruido. Finalmente, en la sección 4.3.4 efectuaremos una comparativa entre el método de estimación propuesto y otras técnicas similares.

4.3.1. Desarrollo del algoritmo de reconstrucción espectral

En este apartado volvemos a adoptar el criterio MMSE para calcular las estimas oportunas de los vectores de características de voz,

$$\hat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}] = \int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}) d\boldsymbol{x}. \tag{4.39}$$

De nuevo el paso que acarrea más dificultad en la derivación del estimador MMSE consiste en obtener una forma analítica para $p(\boldsymbol{x}|\boldsymbol{y})$. Encaminando nuestros pasos hacia este objetivo asumiremos, en primer lugar, que las características de voz en el domino

log-Mel se modelan mediante un GMM $\mathcal{M}_x$,

$$p(\boldsymbol{x}|\mathcal{M}_x) = \sum_{k_x=1}^{M_x} P(k_x|\mathcal{M}_x)\mathcal{N}_x(\boldsymbol{x}; \boldsymbol{\mu}_x^{(k_x)}, \boldsymbol{\Sigma}_x^{(k_x)}), \tag{4.40}$$

siendo M_x el número de gaussianas del GMM.

Asimismo, y como novedad del estimador propuesto en este apartado frente a la técnica de imputación TGI presentada en la sección 4.2, asumimos que el ruido que contamina la señal de voz observada puede modelarse mediante un segundo GMM $\mathcal{M}_n$ con M_n gaussianas,

$$p(\boldsymbol{n}|\mathcal{M}_n) = \sum_{k_n=1}^{M_n} P(k_n|\mathcal{M}_n)\mathcal{N}_n(\boldsymbol{n}; \boldsymbol{\mu}_n^{(k_n)}, \boldsymbol{\Sigma}_n^{(k_n)}) \tag{4.41}$$

que, como se verá en la sección 4.3.3, se estima para cada frase de evaluación.

Usando ambos modelos (4.40) y (4.41), la probabilidad a posteriori $p(\boldsymbol{x}|\boldsymbol{y})$ que aparece en la ecuación (4.39) se puede escribir como la probabilidad marginal de $p(\boldsymbol{x}, k_x, k_n|\boldsymbol{y}, \mathcal{M}_x, \mathcal{M}_n)$ respecto a los índices de las gaussianas:

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y}, \mathcal{M}_x, \mathcal{M}_n) &= \sum_{k_x=1}^{M_x}\sum_{k_n=1}^{M_n} p(\boldsymbol{x}, k_x, k_n|\boldsymbol{y}, \mathcal{M}_x, \mathcal{M}_n) \\ &= \sum_{k_x=1}^{M_x}\sum_{k_n=1}^{M_n} p(\boldsymbol{x}|\boldsymbol{y}, k_x, k_n)P(k_x, k_n|\boldsymbol{y}), \end{aligned} \tag{4.42}$$

donde se ha omitido la referencia a los GMMs $\mathcal{M}_x$ y $\mathcal{M}_n$ para una lectura más clara.

Luego la expresión resultante para el estimador MMSE de la ecuación (4.39) viene dada por

$$\hat{\boldsymbol{x}} = \sum_{k_x=1}^{M_x}\sum_{k_n=1}^{M_n} P(k_x, k_n|\boldsymbol{y}) \underbrace{\int \boldsymbol{x} p(\boldsymbol{x}|\boldsymbol{y}, k_x, k_n) d\boldsymbol{x}}_{\hat{\boldsymbol{x}}^{(k_x,k_n)}}. \tag{4.43}$$

Otra vez nos encontramos con que el estimador MMSE resultante consiste en una combinación lineal de ciertas estimas parciales $\hat{\boldsymbol{x}}^{(k_x,k_n)}$ ponderadas por sus respectivas probabilidades a posteriori $P(k_x, k_n|\boldsymbol{y})$[1]. Constatamos además la similitud de este estimador con las técnicas de combinación de modelos acústicos estudiadas en la sección 2.2.2.2: PMC [97, 103, 105] y las técnicas de factorización/descomposición de modelos propuestas en [229, 232, 233, 267]. Al igual que estas técnicas, nuestro estimador MMSE parte de dos modelos independientes: uno para la voz $\mathcal{M}_x$ y otro para el ruido $\mathcal{M}_n$.

[1] Debe notarse que $\sum_{k_x=1}^{M_x}\sum_{k_n=1}^{M_n} P(k_x, k_n|\boldsymbol{y}) = 1$.

En base a estos dos modelos se construye un tercer modelo $\mathcal{M}_y$ que representa la señal de voz ruidosa y que es el empleado para calcular las probabilidades de observación $p(\boldsymbol{y}|\mathcal{M}_y)$. Como se verá a continuación, el estimador propuesto no calcula de forma explícita los parámetros del modelo $\mathcal{M}_y$, sino que éste se expresa de forma factorizada en base a los modelos $\mathcal{M}_x$ y $\mathcal{M}_n$. Así, cada par $\langle k_x, k_n \rangle$ representa un estado (PDF) de $\mathcal{M}_y$, por lo que en total el modelo de voz distorsionada tiene $M_x \times M_n$ estados diferentes.

Continuando con la derivación matemática del estimador MMSE, seguimos por el cómputo de las probabilidades a posteriori $P(k_x, k_n|\boldsymbol{y})$. Tras la aplicación de la regla de Bayes, esta probabilidad se expresa como

$$\begin{aligned} P(k_x, k_n|\boldsymbol{y}) &= \frac{p(\boldsymbol{y}|k_x, k_n)P(k_x, k_n)}{p(\boldsymbol{y})} \\ &= \frac{p(\boldsymbol{y}|k_x, k_n)P(k_x)P(k_n)}{\sum_{k'_x=1}^{M_x}\sum_{k'_n=1}^{M_n} p(\boldsymbol{y}|k'_x, k'_n)P(k'_x)P(k'_n)}, \end{aligned} \tag{4.44}$$

donde se ha supuesto independencia estadística entre voz y ruido para simplificar la probabilidad $P(k_x, k_n)$ y expresarla como producto de sus factores.

En la ecuación (4.44) el único término que nos es desconocido es la probabilidad de observación $p(\boldsymbol{y}|k_x, k_n)$, ya que $P(k_x)$ y $P(k_n)$ son los pesos de las componentes en los modelos de voz y ruido, respectivamente. Para el cálculo de $p(\boldsymbol{y}|k_x, k_n)$, expresaremos esta probabilidad como la PDF marginal de $p(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{y}|k_x, k_n)$:

$$\begin{aligned} p(\boldsymbol{y}|k_x, k_n) &= \iint p(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{y}|k_x, k_n)d\boldsymbol{x}d\boldsymbol{n} \\ &= \iint p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n})p(\boldsymbol{x}|k_x)p(\boldsymbol{n}|k_n)d\boldsymbol{x}d\boldsymbol{n}, \end{aligned} \tag{4.45}$$

en esta ecuación se asume que $\boldsymbol{y}$ es independiente de las gaussianas k_x y k_n supuesto que $\boldsymbol{x}$ y $\boldsymbol{n}$ son conocidos.

Al contrario que la técnica de imputación descrita en la sección 4.2, en la técnica que aquí proponemos no se tiene constancia a priori de la fiabilidad de cada elemento del espectro. La única información disponible viene dada por el modelo de enmascaramiento de la ecuación (4.4) que nos indica que sólo una fuente, voz o ruido, domina en cada frecuencia. Esto es, cada elemento de la observación $\boldsymbol{y}$ equivale al máximo entre los elementos de los vectores $\boldsymbol{x}$ y $\boldsymbol{n}$ (aunque no sabemos cuál en cada uno de ellos). Por tanto, la probabilidad $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n})$ se puede expresar como

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n}) = \delta_{\text{máx}(\boldsymbol{x},\boldsymbol{n})}(\boldsymbol{y}). \tag{4.46}$$

Dado que la operación máx se aplica por separado a cada uno de los elementos de sus vectores argumento, la PDF anterior equivale al producto de las deltas aplicadas a

cada uno de los elementos por separado,

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{n}) = \prod_{i=1}^{D} \delta_{\text{máx}(x_i,n_i)}(y_i). \tag{4.47}$$

Las deltas de Dirac de la ecuación anterior pueden expresarse, de forma alternativa, como la siguiente función definida a trozos:

$$\delta_{\text{máx}(x_i,n_i)}(y_i) = \begin{cases} \infty, & (x_i = y_i) \wedge (n_i \leq x_i) \\ \infty, & (n_i = y_i) \wedge (x_i < n_i) \\ 0, & \text{en otro caso} \end{cases} \tag{4.48}$$

Como los tres casos de la función definida a trozos anterior son excluyentes entre sí, ésta se expresa de forma más conveniente para nuestros propósitos como la siguiente suma de términos más simples

$$\delta_{\text{máx}(x_i,n_i)}(y_i) = \delta_{x_i}(y_i)\mathbb{1}_{n_i \leq x_i} + \delta_{n_i}(y_i)\mathbb{1}_{x_i < n_i}. \tag{4.49}$$

Luego usando la ecuación anterior en (4.47), $p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{n})$ queda expresado como sigue

$$p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{n}) = \prod_{i=1}^{D} \left(\delta_{x_i}(y_i)\mathbb{1}_{n_i \leq x_i} + \delta_{n_i}(y_i)\mathbb{1}_{x_i < n_i}\right). \tag{4.50}$$

Esta ecuación equivale a evaluar las 2^D combinaciones posibles en las que se puede segregar la observación $\boldsymbol{y}$ en términos de considerar cada uno de sus elementos como voz o ruido:

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{n}) = {} & (\delta_{x_1}(y_1)\delta_{x_2}(y_2)\ldots\delta_{x_D}(y_D)\mathbb{1}_{n_1 \leq x_1}\mathbb{1}_{n_2 \leq x_2}\ldots\mathbb{1}_{n_D \leq x_D}) \\ & + (\delta_{x_1}(y_1)\delta_{x_2}(y_2)\ldots\delta_{n_D}(y_D)\mathbb{1}_{n_1 \leq x_1}\mathbb{1}_{n_2 \leq x_2}\ldots\mathbb{1}_{x_D < n_D}) + \ldots \\ & + (\delta_{n_1}(y_1)\delta_{n_2}(y_2)\ldots\delta_{n_D}(y_D)\mathbb{1}_{x_1 < n_1}\mathbb{1}_{x_2 < n_2}\ldots\mathbb{1}_{x_D < n_D}). \end{aligned} \tag{4.51}$$

Al usar (4.51) en la ecuación (4.45), la doble integral de marginalización se divide en 2^D sumandos, uno para cada una de las posibles combinaciones en las que se puede segregar la observación. Para una parametrización estándar de la voz con $D = 23$ filtros Mel, esto equivale a evaluar del orden de $2^{23} = 8.388.608$ integrales, lo que a todas luces es inviable computacionalmente hablando [229]. Para empeorar las cosas constatamos que estas integrales no tendrán solución analítica si los modelos $\mathcal{M}_x$ y $\mathcal{M}_n$ cuentan con matrices de covarianza no diagonales. En estos casos tendremos que recurrir a ciertas aproximaciones (p.ej. retener únicamente los elementos de la diagonal principal) para poder abordar el cálculo de dichas integrales.

Dado que el cálculo de $p(\boldsymbol{y}|k_x,k_n)$ se hace inviable en la forma que esta probabilidad se presenta en la ecuación (4.45), de aquí en adelante asumiremos independencia

estadística entre los elementos del vector $\boldsymbol{y}$ con objeto de obtener una solución analítica tratable para el cómputo de $p(\boldsymbol{y}|k_x, k_n)$:

$$p(\boldsymbol{y}|k_x, k_n) = \prod_{i=1}^{D} p(y_i|k_x, k_n). \tag{4.52}$$

De forma similar a la ecuación (4.45), aquí podemos expresar $p(y_i|k_x, k_n)$ como la distribución marginal de $p(x_i, n_i, y_i|k_x, k_n)$,

$$\begin{aligned} p(y_i|k_x, k_n) &= \iint p(x_i, n_i, y_i|k_x, k_n) dx_i dn_i \\ &= \iint p(y_i|x_i, n_i) p(x_i|k_x) p(n_i|k_n) dx_i dn_i. \end{aligned} \tag{4.53}$$

Aplicando el modelo de enmascaramiento (4.4) tenemos que $p(y_i|x_i, n_i)$ viene dada por

$$p\left(y_i|x_i, n_i\right) = \delta_{\text{máx}(x_i, ni)}(y_i) = \delta_{x_i}(y_i)\mathbb{1}_{n_i \leq x_i} + \delta_{n_i}(y_i)\mathbb{1}_{x_i < n_i}. \tag{4.54}$$

Finalmente sustituyendo el término $p(y_i|x_i, n_i)$ de la ecuación (4.53) por su valor de la ecuación (4.54), tenemos que la probabilidad de observación $p(y_i|k_x, k_n)$ es

$$\begin{aligned} &p(y_i|k_x, k_n) \\ &= \iint p(x_i|k_x) p(n_i|k_n) \delta_{x_i}(y_i) \mathbb{1}_{n_i \leq x_i} dx_i dn_i + \iint p(x_i|k_x) p(n_i|k_n) \delta_{n_i}(y_i) \mathbb{1}_{x_i < n_i} dx_i dn_i \\ &= p(y_i|k_x) \int_{-\infty}^{y_i} p(n_i|k_n) dn_i + p(y_i|k_n) \int_{-\infty}^{y_i} p(x_i|k_x) dx_i \\ &= \underbrace{\mathcal{N}_x\left(y_i; \mu_{x,i}^{(k_x)}, \sigma_{x,i}^{(k_x)}\right) \Phi_n\left(y_i; \mu_{n,i}^{(k_n)}, \sigma_{n,i}^{(k_n)}\right)}_{p(x_i = y_i, n_i \leq y_i|k_x, k_n)} + \underbrace{\mathcal{N}_n\left(y_i; \mu_{n,i}^{(k_n)}, \sigma_{n,i}^{(k_n)}\right) \Phi_x\left(y_i; \mu_{x,i}^{(k_x)}, \sigma_{x,i}^{(k_x)}\right)}_{p(n_i = y_i, x_i < y_i|k_x, k_n)}, \end{aligned} \tag{4.55}$$

donde se ha notado por $\mu_{x,i}^{(k_x)}$ y $\mu_{n,i}^{(k_n)}$ al elemento i-ésimo de las medias en las gaussianas k_x y k_n, respectivamente. De forma similar, $\sigma_{x,i}^{(k_x)}$ y $\sigma_{n,i}^{(k_n)}$ denotan a dicho elemento en los vectores con las desviaciones típicas de las gaussianas.

La expresión obtenida en la ecuación (4.55) para el cálculo de la probabilidad de observación $p(y_i|k_x, k_n)$ puede interpretarse fácilmente como sigue. De acuerdo al modelo de enmascaramiento de la ecuación (4.4), la observación y_i es el valor máximo entre x_i y n_i. Dicho con otras palabras esto equivale a decir que y_i es igual a x_i cuando n_i es menor o igual que x_i o, en caso contrario, y_i será igual a n_i cuando la energía del ruido exceda a la de la voz. Trasladado al dominio probabilístico el enunciado anterior supone evaluar dos probabilidades: $p(x_i = y_i, n_i \leq y_i|k_x, k_n)$ y $p(n_i = y_i, x_i < y_i|k_x, k_n)$.

La primera de ellas es la probabilidad de que la voz no hay sufrido ninguna distorsión por ser la magnitud del ruido menor. La segunda, por contra, evalúa la probabilidad de que la voz haya sido enmascarada por el ruido. Como se desconoce cuál de los dos casos se ha producido, las dos probabilidades anteriores se suman.

Tras haber derivado las expresiones matemáticas correspondientes a las probabilidades a posteriori $P(k_x, k_n|\boldsymbol{y})$ que intervienen en el estimador MMSE de la ecuación (4.43), a continuación analizaremos el cómputo de las estimas parciales $\hat{\boldsymbol{x}}^{(k_x,k_n)}$. Asumiendo de nuevo independencia estadística entre los distintos elementos de los vectores de características, las estimas anteriores equivalen al siguiente valor esperado

$$\hat{x}_i^{(k_x,k_n)} = \int x_i p(x_i|y_i, k_x, k_n) dx_i \equiv \iint x_i p(x_i, n_i|y_i, k_x, k_n) dx_i dn_i, \tag{4.56}$$

donde se ha expresado $p(x_i|y_i, k_x, k_n)$ como la distribución marginal de $p(x_i, n_i|y_i, k_x, k_n)$.

Recurriendo a la regla de Bayes, $p(x_i, n_i|y_i, k_x, k_n)$ se calcula como sigue

$$p(x_i, n_i|y_i, k_x, k_n) = \frac{p(y_i|x_i, n_i)p(x_i|k_x)p(n_i|k_n)}{p(y_i|k_x, k_n)}. \tag{4.57}$$

Todos los términos que aparecen en la ecuación anterior son conocidos: $p(y_i|x_i, n_i)$ viene dado por (4.54), $p(x_i|k_x)$ y $p(n_i|k_n)$ se pueden calcular directamente usando los modelos de voz y ruido, respectivamente, y $p(y_i|k_x, k_n)$ es la probabilidad de observación recogida en (4.55). Luego usando las expresiones obtenidas para estos términos, el valor esperado de la ecuación (4.56) se deriva de la siguiente forma

$$\begin{aligned}
\hat{x}_i^{(k_x,k_n)} &= \iint x_i \frac{p(y_i|x_i, n_i)p(x_i|k_x)p(n_i|k_n)}{p(y_i|k_x, k_n)} dx_i dn_i \\
&= \frac{1}{p(y_i|k_x, k_n)} \left[\iint x_i p(x_i|k_x)p(n_i|k_n)\delta_{x_i}(y_i)\mathbb{1}_{n_i \leq x_i} dx_i dn_i \right. \\
&\qquad \left. + \iint x_i p(x_i|k_x)p(n_i|k_n)\delta_{n_i}(y_i)\mathbb{1}_{x_i < x_i} dx_i dn_i \right] \\
&= \frac{1}{p(y_i|k_x, k_n)} \left[y_i p(y_i|k_x) \int_{-\infty}^{y_i} p(n_i|k_n) dn_i + p(y_i|k_n) \int_{-\infty}^{y_i} x_i p(x_i|k_x) dx_i \right].
\end{aligned} \tag{4.58}$$

Definiendo la probabilidad de presencia de voz (SPP, *Speech Presence Probability*) para el par de gaussianas $\langle k_x, k_n \rangle$ como

$$\begin{aligned}
w_i^{(k_x,k_n)} &= \frac{p(y_i|k_x) \int_{-\infty}^{y_i} p(n_i|k_n) dn_i}{p(y_i|k_x, k_n)} \\
&= \frac{\mathcal{N}_x\left(y_i; \mu_{x,i}^{(k_x)}, \sigma_{x,i}^{(k_x)}\right) \Phi_n\left(y_i; \mu_{n,i}^{(k_n)}, \sigma_{n,i}^{(k_n)}\right)}{p(y_i|k_x, k_n)}
\end{aligned} \tag{4.59}$$

y la probabilidad de enmascaramiento de la voz como

$$1 - w_i^{(k_x,k_n)} = \frac{p(y_i|k_n)\int_{-\infty}^{y_i} p(x_i|k_x)dx_i}{p(y_i|k_x,k_n)} = \frac{\mathcal{N}_n\left(y_i;\mu_{n,i}^{(k_n)},\sigma_{n,i}^{(k_n)}\right)\Phi_x\left(y_i;\mu_{x,i}^{(k_x)},\sigma_{x,i}^{(k_x)}\right)}{p(y_i|k_x,k_n)}. \tag{4.60}$$

Entonces la ecuación (4.58) puede reescribirse de forma más compacta como

$$\hat{x}_i^{(k_x,k_n)} = w_i^{(k_x,k_n)} y_i + \left(1 - w_i^{(k_x,k_n)}\right)\tilde{\mu}_{x,i}^{(k_x)}, \tag{4.61}$$

siendo $\tilde{\mu}_{x,i}^{(k_x)}$ la media de la gaussiana k_x-ésima cuando ésta se encuentra truncada a la derecha por el valor y_i (ver ecuación (B.3) del apéndice B).

Con esto finaliza el desarrollo matemático del algoritmo de reconstrucción espectral basado en el modelo de enmascaramiento con modelos de ruido. La interpretación del estimador obtenido es clara. En la ecuación (4.61) vemos que las estimas parciales $\hat{x}_i^{(k_x,k_n)}$ son combinación lineal de dos términos: y_i y $\tilde{\mu}_{x,i}^{(k_x)}$. El primer término, y_i, se corresponde con el nivel de energía observado y puede considerarse como la estimación de la energía de la voz para niveles de SNR altos. Por contra, $\tilde{\mu}_{x,i}^{(k_x)}$ equivale al valor de voz estimado cuando ésta ha sido enmascarada completamente por la energía del ruido. En este caso extremo, la única información que se conoce sobre el nivel de energía de la voz nos la proporciona el modelo de enmascaramiento: el nivel de energía se moverá en el intervalo $(-\infty, y_i]$. Los dos términos anteriores, y_i y $\tilde{\mu}_{x,i}^{(k_x)}$, van ponderados en (4.61) por las probabilidades de presencia de voz $w_i^{(k_x,k_n)}$ y de enmascaramiento $1 - w_i^{(k_x,k_n)}$, respectivamente.

Con objeto de facilitar la comparación posterior con otras técnica de reconstrucción similares, desarrollaremos a continuación la expresión final obtenida para la estimación MMSE de la característica i-ésima:

$$\begin{aligned}
\hat{x}_i &= \sum_{k_x=1}^{M_x}\sum_{k_n=1}^{M_n} P(k_x,k_n|\boldsymbol{y})\left[w_i^{(k_x,k_n)} y_i + \left(1 - w_i^{(k_x,k_n)}\right)\tilde{\mu}_{x,i}^{(k_x)}\right] \\
&= \underbrace{\left[\sum_{k_x=1}^{M_x}\sum_{k_n=1}^{M_n} P(k_x,k_n|\boldsymbol{y}) w_i^{(k_x,k_n)}\right]}_{m_i} y_i + \sum_{k_x=1}^{M_x}\sum_{k_n=1}^{M_n} P(k_x,k_n|\boldsymbol{y})\left(1 - w_i^{(k_x,k_n)}\right)\tilde{\mu}_{x,i}^{(k_x)} \\
&= m_i y_i + \sum_{k_x=1}^{M_x}\tilde{\mu}_{x,i}^{(k_x)} \underbrace{\sum_{k_n=1}^{M_n} P(k_x,k_n|\boldsymbol{y})}_{P(k_x|\boldsymbol{y})} - \sum_{k_x=1}^{M_x}\tilde{\mu}_{x,i}^{(k_x)} \underbrace{\sum_{k_n=1}^{M_n} P(k_x,k_n|\boldsymbol{y}) w_i^{(k_x,k_n)}}_{P(k_x|\boldsymbol{y})w_i^{(k_x)}} \\
&= m_i y_i + \sum_{k_x=1}^{M_x}\left(1 - w_i^{(k_x)}\right) P(k_x|\boldsymbol{y})\tilde{\mu}_{x,i}^{(k_x)}.
\end{aligned} \tag{4.62}$$

De aquí en adelante el estimador de la ecuación anterior se le conocerá como técnica de reconstrucción espectral basada en el modelo de enmascaramiento (MMSR, *Masking-Model based Spectral Reconstruction*). De nuevo constatamos que la estimación $\hat{x}_i$ es una combinación lineal del valor original y_i y de una estima parcial que representa el caso de enmascaramiento total de la voz. Esta última estima se calcula, a su vez, como combinación lineal de las medias de las gaussianas truncadas del GMM de voz $\mathcal{M}_x$.

4.3.1.1. Derivación alternativa

Aunque durante la derivación de la técnica MMSR que acabamos de presentar se ha considerado que el ruido se modela mediante una mezcla de gaussianas $\mathcal{M}_n$, debemos subrayar que dicha técnica también puede operar con estimaciones puntuales de las características del ruido. Así, a semejanza de otras técnicas de compensación y realce de la voz, en ciertos entornos puede que se disponga de una serie de estimas de la densidad de potencia espectral del ruido expresadas en el dominio log-Mel $(\hat{\boldsymbol{n}}_1, \ldots, \hat{\boldsymbol{n}}_T)$, una para cada instante de tiempo. Ante la incertidumbre inherente al proceso de estimación del ruido, podemos considerar, además, que cada estima $\hat{\boldsymbol{n}}_t$ tiene asociada una única PDF de evidencia $k_{n,t}$, ahora dependiente del tiempo, cuya media coincide con la propia estima $\boldsymbol{\mu}_{n,t} = \hat{\boldsymbol{n}}_t$ y cuya matriz de covarianza (diagonal) $\boldsymbol{\Sigma}_{n,t}$ modela el error del proceso de estimación,

$$p(\boldsymbol{n}_t|k_{n,t}) = \mathcal{N}_n(\boldsymbol{n}_t, \boldsymbol{\mu}_{n,t}, \boldsymbol{\Sigma}_{n,t}). \tag{4.63}$$

Bajo este nuevo modelo a priori del ruido, la expresión de la técnica MMSR queda de la siguiente forma

$$\begin{aligned}
\hat{x}_{t,i} &= \sum_{k_x=1}^{M_x} P(k_x, k_{n,t}|\boldsymbol{y}_t)\hat{x}_{t,i}^{(k_x,k_{n,t})} \\
&= \sum_{k_x=1}^{M_x} P(k_x, k_{n,t}|\boldsymbol{y}_t)\left[w_{t,i}^{(k_x,k_{n,t})} y_{t,i} + \left(1 - w_{t,i}^{(k_x,k_{n,t})}\right)\tilde{\mu}_{x,t,i}^{(k_x)}\right] \\
&= \underbrace{\left[\sum_{k_x=1}^{M_x} P(k_x, k_n|\boldsymbol{y}_t) w_{t,i}^{(k_x,k_{n,t})}\right]}_{m_{t,i}} y_{t,i} + \sum_{k_x=1}^{M_x} P(k_x, k_{n,t}|\boldsymbol{y}_t)\left(1 - w_{t,i}^{(k_x,k_{n,t})}\right)\tilde{\mu}_{x,t,i}^{(k_x)},
\end{aligned} \tag{4.64}$$

siendo $\tilde{\mu}_{x,t,i}^{(k_x)}$ la media de la gaussiana truncada al valor $y_{t,i}$.

Puede observarse que el estimador obtenido es muy similar al de la ecuación (4.62), radicando la mayor diferencia entre ambos en el modelo de ruido y su uso: en el estimador de la ecuación (4.62) el ruido se modela mediante un GMM común para toda

la frase de evaluación, mientras que en el estimador de la ecuación (4.64) el modelo de ruido es variante con el tiempo, ya que se dispone de una estima para cada instante de tiempo.

Es interesante analizar el comportamiento de esta nueva versión de la técnica MMSR en función de la fiabilidad de la estimación de ruido. Cuando la varianza de la estimación del ruido en (4.63) tiende a cero ($|\mathbf{\Sigma}_{n,t}| \to 0$), podemos considerar que se dispone de una especie de máscara oráculo. En tal caso, tenemos que $p(\boldsymbol{n}_t|k_{n,t}) \approx \delta_{\hat{\boldsymbol{n}}_t}(\boldsymbol{n}_t)$ y la probabilidad de observación $p(y_{t,i}|k_x, k_{n,t})$ de la ecuación (4.55) equivale a

$$\begin{aligned} p(y_{t,i}|k_x, k_{n,t}) &= p(y_{t,i}|k_x) \int_{-\infty}^{y_{t,i}} n_i p(n_i|k_{n,t}) dn_i + p(y_{t,i}|k_{n,t}) \int_{-\infty}^{y_{t,i}} x_i p(x_i|k_x) dx_i \\ &= p(y_{t,i}|k_x)\Phi(y_{t,i}|k_{n,t}) + p(y_{t,i}|k_{n,t})\Phi(y_{t,i}|k_x). \end{aligned} \tag{4.65}$$

Dado que la PDF de evidencia asociada al ruido es una delta de Dirac, $p(y_{t,i}|k_{n,t})$ será otra delta,

$$p(y_{t,i}|k_{n,t}) = \delta_{\hat{n}_{t,i}}(y_{t,i}) \tag{4.66}$$

y la función de probabilidad acumulada, $\Phi(y_{t,i}|k_{n,t})$, se corresponde con una función escalón centrada en $\hat{n}_{t,i}$,

$$\Phi(y_{t,i}|k_{n,t}) = \mathcal{U}_{\hat{n}_{t,i}}(y_{t,i}) = \begin{cases} 0, & y_{t,i} < \hat{n}_{t,i} \\ 1, & y_{t,i} \geq \hat{n}_{t,i} \end{cases}. \tag{4.67}$$

Usando las dos ecuaciones anteriores, la probabilidad de observación $p(y_{t,i}|k_x, k_{n,t})$ definida en (4.65) se puede expresar como una función definida a trozos en función de las magnitudes relativas de la observación $y_{t,i}$ y la estima de ruido $\hat{n}_{t,i}$,

$$p(y_{t,i}|k_x, k_{n,t}) = \begin{cases} 0, & y_{t,i} < \hat{n}_{t,i} \\ p(y_{t,i}|k_x) + \delta_{\hat{n}_{t,i}}(y_{t,i})\Phi(y_{t,i}|k_x) \approx \delta_{\hat{n}_{t,i}}, & y_{t,i} = \hat{n}_{t,i} \\ p(y_{t,i}|k_x), & y_{t,i} > \hat{n}_{t,i} \end{cases}. \tag{4.68}$$

Estudiemos por separado cada uno de estos tres casos. En primer lugar, al considerar la PDF de evidencia del ruido como una delta de Dirac, implícitamente asumimos que la estimación del ruido $\hat{n}_{t,i}$ es perfecta. Esto implica que el primer caso de la ecuación anterior, $y_{t,i} < \hat{n}_{t,i}$, no puede darse[1] y, por tanto, la probabilidad de observación es nula. El segundo caso ($y_{t,i} = \hat{n}_{t,i}$) se presenta cuando la voz es enmascarada por el ruido y, por consiguiente, la observación $y_{t,i}$ es igual a la energía estimada del ruido $\hat{n}_{t,i}$. En tal situación, la probabilidad de observación $p(y_{t,i}|k_x, k_{n,t})$ nos dice que las probabilidades del modelo de voz son despreciables y que ésta viene dominada por la

[1] Bajo el supuesto de que la relación entre fases entre voz y ruido es nula.

probabilidad del modelo de ruido $\delta_{\hat{n}_{t,i}}(y_{t,i})$, lo que equivale a decir que ese elemento del espectro no es fiable. El cálculo de las probabilidades a posteriori $P(k_x, k_{n,t}|\boldsymbol{y}_t)$, no obstante, no se verá afectado por estos valores no fiables, ya que la contribución de las probabilidades $\delta_{\hat{n}_{t,i}}(y_{t,i})$ se verá cancelado al aparecer en el numerador y el denominador de la ecuación (4.44). En el caso límite en el que el vector observado $\boldsymbol{y}_t$ coincida con el ruido estimado $\hat{\boldsymbol{n}}_t$ y todos sus elementos sean no fiables, las probabilidades a posteriori serán aproximadamente iguales a las probabilidades a priori de las gaussianas del modelo de voz $P(k_x, k_{n,t}|\boldsymbol{y}_t) \approx P(k_x)$. Si el segundo caso de la función por trozos anterior equivale al de un elemento perdido, el último caso ($y_{t,i} > \hat{n}_{t,i}$) es el de un elemento fiable en el que la energía de la voz es mucho mayor que la del ruido. En tal caso, la probabilidad de observación viene determinada únicamente por la gaussiana del modelo de voz $p(y_{t,i}|k_x, k_{n,t}) = p(y_{t,i}|k_x)$.

Al igual que hemos analizado el comportamiento de las probabilidades de observación $p(y_{t,i}|k_x, k_{n,t})$ cuando la PDF de evidencia asociada a la estimación del ruido es $\delta_{\hat{n}_{t,i}}(y_{t,i})$, también resulta interesante estudiar el comportamiento de las estimas parciales $\hat{x}_{t,i}^{(k_x,k_{n,t})}$ que aparecen en el estimador MMSE de la ecuación (4.64) en dicho caso. Recordemos que el valor de $\hat{x}_{t,i}^{(k_x,k_{n,t})}$ viene dado por la ecuación (4.61), por lo que si sustituimos los términos en los que aparece la PDF del ruido en dicha expresión por $\delta_{\hat{n}_{t,i}}(y_{t,i})$ y la CDF por $\mathcal{U}_{\hat{n}_{t,i}}(y_{t,i})$ resulta

$$\begin{aligned}\hat{x}_{t,i}^{(k_x,k_{n,t})} &= w_{t,i}^{(k_x,k_{n,t})} y_{t,i} + \left(1 - w_{t,i}^{(k_x,k_{n,t})}\right) \tilde{\mu}_{x,t,i}^{(k_x)} \\ &= \frac{p(y_{t,i}|k_x)\mathcal{U}_{\hat{n}_{t,i}}(y_{t,i})}{p(y_{t,i}|k_x, k_{n,t})} y_{t,i} + \frac{\delta_{\hat{n}_{t,i}}(y_{t,i})\Phi(y_{t,i}|k_x)}{p(y_{t,i}|k_x, k_{n,t})} \tilde{\mu}_{x,t,i}^{(k_x)}.\end{aligned} \tag{4.69}$$

Considerando de nuevo los distintos casos que se obtienen en función de los valores relativos de $y_{t,i}$ y $\hat{n}_{t,i}$ tenemos lo siguiente:

$$\hat{x}_{t,i}^{(k_x,k_{n,t})} = \begin{cases} \infty, & y_{t,i} < \hat{n}_{t,i} \\ \underbrace{\dfrac{p(y_{t,i}|k_x)}{\delta_{\hat{n}_{t,i}}(y_{t,i})}}_{\approx 0} y_{t,i} + \underbrace{\dfrac{\delta_{\hat{n}_{t,i}}(y_{t,i})\Phi(y_{t,i}|k_x)}{\delta_{\hat{n}_{t,i}}(y_{t,i})}}_{\approx 1} \tilde{\mu}_{x,t,i}^{(k_x)}, & y_{t,i} = \hat{n}_{t,i} \\ y_{t,i}, & y_{t,i} > \hat{n}_{t,i}. \end{cases} \tag{4.70}$$

Exceptuando el primer caso, que bajo la suposición de certidumbre absoluta en la estimación de $\hat{n}_{t,i}$ no puede darse, los otros dos casos recogidos en la ecuación anterior coinciden con las situaciones en las que el elemento del espectro se encuentra perdido (segundo caso) y el elemento es fiable (tercer caso). Cuando la observación $y_{t,i}$ es completamente fiable, el valor estimado $\hat{x}_{t,i}^{(k_x,k_{n,t})}$ coincide con la propia observación, como no podría ser de otra forma. Si, por contra, la energía del ruido enmascara a

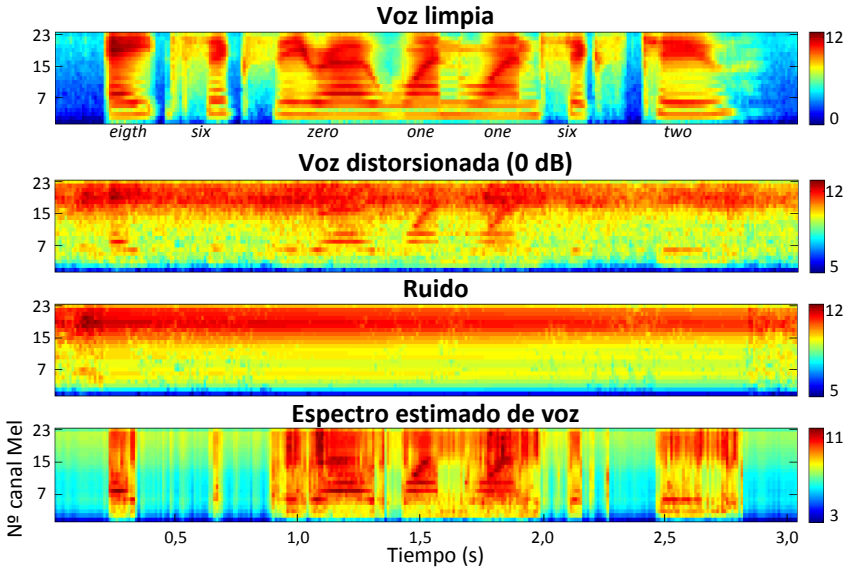

Figura 4.6: Ejemplo de reconstrucción espectral usando ruido estimado.

la de la voz, entonces el valor estimado equivale a la media de la gaussiana truncada $\tilde{\mu}_{x,t,i}^{(k_x)}$ que aparece recogida en el segundo caso de la ecuación previa. En pocas palabras, lo que demuestran las ecuaciones (4.68) y (4.70) es la equivalencia del algoritmo de reconstrucción propuesto en este apartado con la técnica de imputación TGI estudiada en la sección 4.2, cuando ambos enfoques trabajan en condiciones ideales (ruido y máscaras oráculo).

Para concluir este apartado, en la figura 4.6 mostramos un ejemplo del proceso de reconstrucción espectral llevado a cabo por la técnica MMSR que hemos presentado en este apartado. En la figura se muestran cuatro espectrogramas log-Mel correspondientes a la frase *eight six zero one one six two* (ocho seis cero uno uno seis dos) de la base de datos Aurora2 [141]: espectrograma de la frase sin distorsionar (primera gráfica), espectrograma de la frase distorsionada por ruido aditivo del tipo *subway* a 0 dB de SNR (segunda gráfica), estimación de ruido usada por el algoritmo de reconstrucción (tercera gráfica) y espectro reconstruido obtenido por el algoritmo propuesto (última gráfica). La estima de ruido de la tercera gráfica se ha calculado mediante interpolación lineal de las dos medias parciales calculadas para las 20 primeras y 20 últimas tramas. Después de la interpolación lineal, el valor de ruido estimado se ha acotado superiormente para que éste no supere en ningún caso el valor de la observación ruidosa. Además de dicha estima, al algoritmo de reconstrucción también se le ha suministrado una matriz de covarianza fija e invariante al tiempo (no mostrada en la gráfica) que modela el error de la estimación del ruido. Esta matriz contiene en su diagonal principal las varianzas

de las 20 primeras y las 20 últimas tramas de la frase. Tomando como entrada el espectro distorsionado de voz, el ruido estimado junto con su varianza y un modelo de voz (GMM con 32 gaussianas), el algoritmo de reconstrucción devuelve a su salida el espectro notado como "Espectro estimado de voz". Incluso a una SNR tan baja como la usada en el ejemplo (0 dB), comprobamos visualmente que el algoritmo de reconstrucción es capaz de mitigar parcialmente la distorsión producida por el ruido aditivo en la señal de voz. En comparación con la reconstrucción mostrada en la figura 4.4, el espectro estimado que aquí mostramos contiene más distorsiones, pero debe notarse que aquí empleamos ruido estimado mientras que en la figura 4.4 usábamos máscaras oráculo.

4.3.2. Estimación de la fiabilidad de los elementos del espectro

La técnica MMSR de reconstrucción espectral que acabamos de presentar puede considerarse, de forma alternativa, como una técnica robusta para la segregación del espectro en sus componentes fiables y no fiables. Para lograr esta funcionalidad, basta considerar de nuevo las fórmulas de estimación finalmente obtenidas. En la versión del estimador que usa GMMs para modelar la distribución del ruido, la expresión obtenida es la que aparece recogida en (4.62), la cual volvemos a reproducir aquí por claridad

$$\hat{x}_i = \underbrace{\left[\sum_{k_x=1}^{M_x}\sum_{k_n=1}^{M_n} P(k_x,k_n|\boldsymbol{y}) w_i^{(k_x,k_n)}\right]}_{m_i} y_i + \sum_{k_x=1}^{M_x}\sum_{k_n=1}^{M_n} P(k_x,k_n|\boldsymbol{y})\left(1-w_i^{(k_x,k_n)}\right)\tilde{\mu}_{x,i}^{(k_x)}, \tag{4.71}$$

donde se ha omitido el índice temporal a fin de simplificar la expresión resultante.

Para la versión que usa estimaciones de ruido calculadas para cada instante de tiempo, la expresión final es la que aparece en la ecuación (4.64),

$$\hat{x}_{t,i} = \underbrace{\left[\sum_{k_x=1}^{M_x} P(k_x,k_n|\boldsymbol{y}_t) w_{t,i}^{(k_x,k_{n,t})}\right]}_{m_{t,i}} y_{t,i} + \sum_{k_x=1}^{M_x} P(k_x,k_{n,t}|\boldsymbol{y}_t)\left(1-w_{t,i}^{(k_x,k_{n,t})}\right)\tilde{\mu}_{x,t,i}^{(k_x)}, \tag{4.72}$$

expresión en la que sí aparece explícitamente el índice temporal t, ya que las estimas de ruido $\hat{\boldsymbol{n}}_t$ varían con el tiempo.

En ambos casos observamos que el valor estimado consiste en una combinación lineal de la observación original y de una sumatoria en la que aparecen las medias de las gaussianas truncadas del GMM. El peso en la combinación lineal de la observación

y_i (alternativamente $y_{t,i}$) al que hemos notado por m_i ($m_{t,i}$ en la ecuación (4.72)) puede considerarse como una máscara de segregación continua, ya que coincide con la probabilidad a posteriori de que la energía de la voz sea dominante, $m_i = P(x_i \geq n_i|\boldsymbol{y}, \mathcal{M}_x, \mathcal{M}_n)$. Esto puede comprobarse fácilmente desarrollando esta probabilidad. Para el estimador de la ecuación (4.71) tenemos que $P(x_i \geq n_i|\boldsymbol{y}, \mathcal{M}_x, \mathcal{M}_n)$ puede expresarse como (de nuevo se omite el índice temporal por claridad),

$$P(x_i \geq n_i|\boldsymbol{y}, \mathcal{M}_x, \mathcal{M}_n) = \sum_{k_x=1}^{M_x} \sum_{k_n=1}^{M_n} P(k_x, k_n|\boldsymbol{y}) P(x_i \geq n_i|\boldsymbol{y}, k_x, k_n), \tag{4.73}$$

donde $P(k_x, k_n|\boldsymbol{y})$ viene dada por la ecuación (4.44).

Para calcular el término $P(x_i \geq n_i|\boldsymbol{y}, k_x, k_n)$ asumimos, en primer lugar, independencia entre los elementos del vector, luego $P(x_i \geq n_i|y_i, k_x, k_n) = P(x_i \geq n_i|\boldsymbol{y}, k_x, k_n)$. Esta última probabilidad coincide, como veremos a continuación, con el término $w_i^{(k_x,k_n)}$ que aparece en la ecuación (4.71). Bajo las restricciones que nos impone el modelo de enmascaramiento de la ecuación (4.4), $P(x_i \geq n_i|y_i, k_x, k_n)$ es igual al siguiente cociente de probabilidades

$$\begin{aligned} P(x_i \geq n_i|y_i, k_x, k_n) &= \frac{p(x_i = y_i, n_i \leq y_i|k_x, k_n)}{p(x_i = y_i, n_i \leq y_i|k_x, k_n) + p(n_i = y_i, x_i < y_i|k_x, k_n)} \\ &= \frac{p(x_i = y_i, n_i \leq y_i|k_x, k_n)}{p(y_i|k_x, k_n)}. \end{aligned} \tag{4.74}$$

Desarrollando los términos que aparecen en la ecuación anterior tenemos que

$$p(x_i = y_i, n_i \leq y_i|k_x, k_n) = p(x_i = y_i|k_x) P(n_i \leq y_i|k_n) = p(y_i|k_x)\Phi(y_i|k_n), \tag{4.75}$$

$$p(n_i = y_i, x_i < y_i|k_x, k_n) = p(n_i = y_i|k_n) P(x_i < y_i|k_n) = p(y_i|k_n)\Phi(y_i|k_x). \tag{4.76}$$

Por lo tanto,

$$P(x_i \geq n_i|y_i, k_x, k_n) = \frac{p(y_i|k_x)\Phi(y_i|k_n)}{p(y_i|k_x)\Phi(y_i|k_n) + p(y_i|k_n)\Phi(y_i|k_x)}. \tag{4.77}$$

Como podemos comprobar, la ecuación anterior coincide con la definición de $w_i^{(k_x,k_n)}$ dada en (4.59). Luego hemos comprobado que las probabilidades $P(x_i \geq n_i|\boldsymbol{y}, \mathcal{M}_x, \mathcal{M}_n)$ actúan como una máscara continua que segrega el espectro en sus componentes fiables (dominadas por la energía de la voz) y no fiables (dominadas por el ruido). Al contrario que las máscaras binarias usadas en la sección 4.2, la máscara asociada con esta probabilidad es continua, por lo que es de esperar que los errores en la estimación del ruido tengan un efecto menor en ella. Para uso posterior, a continuación reescribimos

la expresión empleada para el cálculo de dichas máscaras:

$$m_i = \sum_{k_x=1}^{M_x} \sum_{k_n=1}^{M_n} P(k_x, k_n|\boldsymbol{y}) w_i^{(k_x,k_n)}. \tag{4.78}$$

Acabamos de derivar la expresión oportuna para el cálculo de las máscaras de segregación asociadas al estimador de la ecuación (4.71). Para la segunda versión del estimador dada por la ecuación (4.72), el proceso es muy similar, por lo que no lo reproduciremos aquí, evitando con ello redundancias innecesarias. Sí que hacemos constar explícitamente la ecuación final para el cálculo de la máscara de segregación:

$$m_{t,i} = \sum_{k_x=1}^{M_x} P(k_x, k_{n,t}|\boldsymbol{y}_t) w_{t,i}^{(k_x,k_{n,t})}. \tag{4.79}$$

En comparación con otras técnicas de estimación de máscaras de segregación, el método que acabamos de proponer presenta varias características que lo hacen muy atractivo. En primer lugar tenemos que es una técnica libre de parámetros experimentales. Esto contrasta con otras técnicas de estimación que requieren conjuntos de validación para hallar experimentalmente el valor óptimo de algunos de sus parámetros (p.ej. las técnicas estudiadas en la sección 2.2.5.5 que aplican una compresión sigmoidal a las estimas locales de SNR). Por otra parte puede observarse que dicho método se basa en un marco estadístico bayesiano que emplea información a priori sobre la voz y el ruido. Desde este punto de vista es plausible esperar que modelos de fuente más precisos redunden en una mejor estimación de estas máscaras. Asimismo, la teoría bayesiana en la que se fundamenta el método propuesto ofrece un marco flexible para la incorporación de otra información que pueda ser de utilidad para una segregación más fiel del espectro (p.ej. *pitch* del locutor, localización espacial, ...).

Finalmente, la figura 4.7 muestra un ejemplo de máscara de segregación estimada por el método descrito en este apartado. De las dos variantes del método de estimación, se ha empleado la que usa estimas puntuales de ruido para cada instante de tiempo. El modelo de ruido empleado coincide con el usado para generar los espectros de la figura 4.6, a saber, interpolación lineal de la estimas de ruido calculadas para las primeras y últimas tramas de la elocución junto con una matriz de covarianza diagonal y fija también obtenida a partir de las primeras y últimas tramas. Además de la máscara de segregación estimada, la figura 4.7 muestra la máscara oráculo (ideal) como referencia. Ésta última se ha calculado aplicando un umbral de 7 dB a la SNR local obtenida a partir de los espectros de voz limpio y ruidoso de la figura. Comparando la máscara estimada con la oráculo constatamos la precisión alcanzada por el método propuesto, incluso a SNRs tan bajas como la usada en este ejemplo.

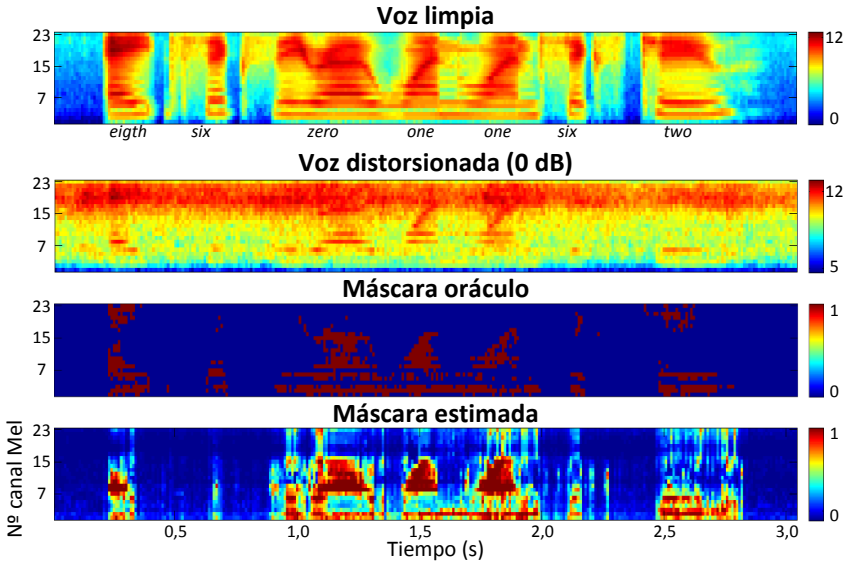

Figura 4.7: Máscara de segregación continua estimada a partir del espectro de voz distorsionada.

4.3.3. Estimación iterativa del modelo de ruido

En los apartados anteriores hemos estudiado cómo, a partir del modelo de enmascaramiento de la voz, podemos desarrollar convenientemente estimadores que nos permitan reconstruir las regiones distorsionadas del espectro y, opcionalmente, calcular la fiabilidad de las características observadas. Estos estimadores se derivaban siguiendo un enfoque bayesiano que emplea de información a priori sobre la distribución en el dominio log-Mel de las características de voz y ruido. En particular, veíamos que la técnica de reconstrucción MMSR emplea sendos GMMs para modelar las probabilidades a priori de la voz $p(\boldsymbol{x}|\mathcal{M}_x)$ y el ruido $p(\boldsymbol{n}|\mathcal{M}_n)$. Mientras que el GMM de voz $\mathcal{M}_x$ puede estimarse sin demasiada dificultad usando voz sin distorsionar, el cálculo de los parámetros del modelo de ruido $\mathcal{M}_n$ se hace más complejo, puesto que debe ser estimado para cada frase de *test* por separado.

La estimación del modelo $\mathcal{M}_n$ conlleva dos problemas destacables. El primero de ellos lo supone segregar los elementos del espectro observado en ruido y voz. Así, mientras que los elementos dominados por el ruido pueden emplearse directamente para estimar los parámetros del modelo $\mathcal{M}_n$, los elementos dominados por la voz deben tratarse como elementos perdidos en este proceso. El segundo problema que nos encontramos es la falta de datos para estimar de forma robusta los parámetros del modelo. En este sentido debemos decir que, en líneas generales, el número de componentes M_n del modelo de ruido deberá ser proporcional a la variabilidad acústica del ruido

y a la duración de la elocución que se reconoce, pero mucho menor que el número de gaussianas M_x del GMM de voz .

Los problemas anteriores han supuesto una barrera práctica para el uso generalizado de modelos de ruido complejos en las técnicas de compensación. En este apartado veremos cómo el modelo de enmascaramiento nos puede ayudar en esta tarea, al permitir solventar parcialmente ambos problemas. En particular, la meta que aquí nos proponemos es el desarrollo de un algoritmo iterativo para la estimación del modelo de ruido $\mathcal{M}_n$, usando para ello el soporte matemático que nos brinda el modelo de enmascaramiento de la ecuación (4.4). Aparte de esto, para derivar dicho algoritmo en este apartado también nos apoyaremos en el algoritmo iterativo introducido en la sección 2.3.2. En todo momento debe quedar claro que el desarrollo que aquí se presenta es aplicable únicamente al caso de modelar el ruido mediante un GMM, pero no a cuando se disponen de una serie de estimas específicas de ruido $\hat{\boldsymbol{n}}_1, \ldots, \hat{\boldsymbol{n}}_T$.

Antes de pasar a una presentación más formal del algoritmo de estimación que proponemos, empezaremos por describir de manera informal la idea subyacente. Para ilustrar esta idea, consideraremos los espectros de voz que aparecen en la figura 4.7: voz limpia, distorsionada a 0 dB, máscara oráculo y máscara estimada. Suponiendo que se tuviese un conocimiento perfecto sobre la segregación del espectro en elementos de voz y ruido, se podría emplear la información recogida en la máscara oráculo para calcular los parámetros del modelo de ruido. Esto es, las medias y varianzas del modelo se calcularían a partir de los elementos del espectro identificados como ruido en la máscara.

La situación anterior, no obstante, no es realista, ya que en la práctica no se suele disponer de la máscara oráculo. Como una aproximación más realista, se puede usar la segregación recogida en la máscara estimada de la figura 4.7. En esta situación ya no tenemos elementos fiables y no fiables, sino que cada elemento del espectro tiene asignado un grado de enmascaramiento continuo. Supuesto, por simplificar la exposición, que el modelo de ruido constara de una única gaussiana, la media de ésta se obtendría promediando las energías de ruido estimadas para cada instante de tiempo. Como el ruido puede quedar enmascarado por la voz, análogamente a lo que se ha hecho en la técnica MMSR, las estimas de ruido vendrían a ser una combinación lineal de los valores de energía observados, $y_{t,i}$, y de ciertas estimas parciales de ruido, $\tilde{\mu}_{n,t,i}$, que representan la situación en la que la voz enmascara totalmente el ruido. Promediando estas estimas se obtendría la media de la gaussiana deseada,

$$\mu_{n,i} = \frac{1}{T}\sum_{t=1}^{T}\left[(1-m_{t,i})y_{t,i} + m_{t,i}\tilde{\mu}_{n,t,i}\right]. \tag{4.80}$$

donde $m_{t,i} \in [0,1]$ indica el valor de la máscara para el canal en frecuencia i e instante

de tiempo t y $\tilde{\mu}_{n,t,i}$ es la media de una gaussiana truncada a la derecha por el valor $y_{t,i}$. Esta media se calcula usando una estimación previa del modelo de ruido.

Al igual que hemos derivado intuitivamente la ecuación para la estimación de la media del modelo de ruido, la expresión para el cálculo de la covarianza podría obtenerse de una forma similar. Extendiendo el razonamiento al caso de contar varias gaussianas, podemos pensar en ajustar el GMM a los datos observados usando un algoritmo EM modificado. Este algoritmo emplea la información recogida en la máscara de segregación para determinar la fiabilidad de los datos. Partiendo de una estimación inicial del modelo, $\mathcal{M}_n^{(0)}$, el algoritmo calcularía la máscara de segregación y, a partir de ella, reestimaría los parámetros del modelo dando lugar a $\mathcal{M}_n^{(1)}$. A partir de este nuevo modelo, se procede de nuevo a calcular la máscara de segregación y a reestimar los parámetros del modelo para dan lugar a una versión actualizada $\mathcal{M}_n^{(2)}$. Este procedimiento iterativo se repite hasta que el algoritmo converja.

De manera formal, el objetivo perseguido en este apartado es hallar los parámetros del GMM de ruido $\mathcal{M}_n$ que, junto con el modelo de voz $\mathcal{M}_x$, maximice la probabilidad de observación de la voz ruidosa $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$. Este modelo, por tanto, será estimado por separado para cada elocución usando un criterio de máxima verosimilitud (ML),

$$\hat{\mathcal{M}}_n = \underset{\mathcal{M}_n}{\operatorname{argmax}}\, p(\boldsymbol{Y}|\mathcal{M}_n, \mathcal{M}_x). \tag{4.81}$$

Los parámetros que definen el GMM del ruido vienen dados por $\mathcal{M}_n = \{\langle \pi_n^{(1)}, \boldsymbol{\mu}_n^{(1)}, \boldsymbol{\Sigma}_n^{(1)} \rangle, \ldots, \langle \pi_n^{(M_n)}, \boldsymbol{\mu}_n^{(M_n)}, \boldsymbol{\Sigma}_n^{(M_n)} \rangle\}$, siendo M_n el número de gaussianas del modelo, $\pi_n^{(k_n)}$ las probabilidades a priori (pesos) de las gaussianas, $\boldsymbol{\mu}_n^{(k_n)}$ las medias y, por último, $\boldsymbol{\Sigma}_n^{(k_n)}$ las matrices de covarianza. Estas matrices serán diagonales ya que, como hemos visto, el modelo de enmascaramiento conduce a la evaluación de CDFs sin solución analítica en caso contrario,

$$\boldsymbol{\Sigma}_n^{(k_n)} = \operatorname{diag}\left(\boldsymbol{\sigma}_n^{(k_n)^2}\right) = \operatorname{diag}\left(\sigma_{n,1}^{(k_n)^2}, \ldots, \sigma_{n,D}^{(k_n)^2}\right). \tag{4.82}$$

En la sección 2.3.2 vimos que la optimización directa de la ecuación (4.81) es matemáticamente inviable, teniendo que recurrir en la práctica a algoritmos iterativos para su cómputo. Al igual que el algoritmo presentado en dicha sección para la estimación del GMM de ruido en el contexto de la técnica VTS, en este apartado adoptaremos el algoritmo EM [66] para el ajuste del GMM de ruido. Notando como $\mathcal{M}_n$ a la hipótesis actual sobre el GMM de ruido y $\hat{\mathcal{M}}_n$ la versión actualizada que maximiza la probabilidad de observación, entonces la función auxiliar $\mathcal{Q}(\mathcal{M}_n, \hat{\mathcal{M}}_n)$ requerida por el algoritmo EM se define como

$$\mathcal{Q}(\mathcal{M}_n, \hat{\mathcal{M}}_n) = \mathbb{E}\left[\log p(\boldsymbol{Y}, \boldsymbol{M}_x, \boldsymbol{M}_n|\hat{\mathcal{M}}_n, \mathcal{M}_x)|\boldsymbol{Y}, \mathcal{M}_n, \mathcal{M}_x\right], \tag{4.83}$$

donde $(\boldsymbol{Y}, \boldsymbol{M}_x, \boldsymbol{M}_n)$ es la tripleta que, en nomenclatura EM, define el conjunto completo de datos compuesto por las observaciones $\boldsymbol{Y}$ y los conjuntos ocultos $\boldsymbol{M}_x$ y $\boldsymbol{M}_n$ con los índices de las gaussianas de los modelos $\mathcal{M}_x$ y $\hat{\mathcal{M}}_n$, respectivamente. Estos últimos conjuntos indican, para cada instante de tiempo, qué componentes de $\mathcal{M}_x$ y $\hat{\mathcal{M}}_n$ han generado el vector de voz observado. Debemos remarcar que el algoritmo EM nos asegura que el método iterativo que define converge, del tal forma que $p(\boldsymbol{Y}|\hat{\mathcal{M}}_n, \mathcal{M}_x) \geq p(\boldsymbol{Y}|\mathcal{M}_n, \mathcal{M}_x)$.

La expresión para la función auxiliar de la ecuación (4.83) se expande como sigue (ver p.ej. [66])

$$\begin{aligned}
\mathcal{Q}(\mathcal{M}_n, \hat{\mathcal{M}}_n) &= \mathbb{E}\left[\log p(\boldsymbol{Y}, \boldsymbol{M}_x, \boldsymbol{M}_n|\hat{\mathcal{M}}_n, \mathcal{M}_x)\Big|\, \boldsymbol{Y}, \mathcal{M}_n, \mathcal{M}_x\right] \\
&= \sum_{t=1}^{T}\sum_{k_x=1}^{M_x}\sum_{k_n=1}^{M_n} P(k_x, k_n|\boldsymbol{y}_t, \mathcal{M}_n, \mathcal{M}_x) \log p(\boldsymbol{y}_t, k_x, k_n|\hat{\mathcal{M}}_n, \mathcal{M}_x) \\
&= \sum_{t=1}^{T}\sum_{k_x=1}^{M_x}\sum_{k_n=1}^{M_n} P(k_x, k_n|\boldsymbol{y}_t, \mathcal{M}_n, \mathcal{M}_x) \Big[\log p(\boldsymbol{y}_t|k_x, k_n, \hat{\mathcal{M}}_n, \mathcal{M}_x) \\
&\qquad + \log p(k_n|\hat{\mathcal{M}}_n) + \log p(k_x|\mathcal{M}_x)\Big], \qquad (4.84)
\end{aligned}$$

donde la probabilidad a posteriori $P(k_x, k_n|\boldsymbol{y}_t, \mathcal{M}_n, \mathcal{M}_x)$ se calcula usando la ecuación (4.44) y la estimación previa del modelo de ruido $\mathcal{M}_n$.

Podemos expresar de forma más compacta esta ecuación omitiendo siempre que sea posible y no lleve a confusión la mención explícita a $\mathcal{M}_x$ y $\hat{\mathcal{M}}_n$. Asimismo, de aquí en adelante notaremos como $\gamma_t^{(k_x,k_n)} = P(k_x, k_n|\boldsymbol{y}_t, \mathcal{M}_n, \mathcal{M}_x)$ y $\hat{\pi}_n^{(k_n)} = p(k_n|\hat{\mathcal{M}}_n)$. También hemos de observar que la probabilidad a priori $p(k_x|\mathcal{M}_x)$ es independiente del modelo de ruido y, por tanto, no afecta a la optimización de sus parámetros. Por último, para el cálculo de $p(\boldsymbol{y}_t|k_x, k_n, \hat{\mathcal{M}}_n, \mathcal{M}_x)$ supondremos independencia estadística entre los elementos del vector $\boldsymbol{y}_t$, tal y como aparece recogido en la ecuación (4.52). Teniendo en cuenta todo esto, la expresión resultante para la función auxiliar del algoritmo EM es

$$\begin{aligned}
\mathcal{Q}(\mathcal{M}_n, \hat{\mathcal{M}}_n) &= \sum_{t=1}^{T}\sum_{k_x=1}^{M_x}\sum_{k_n=1}^{M_n} \gamma_t^{(k_x,k_n)} \left[\log \prod_{i=1}^{D} p(y_{t,i}|k_x, k_n) + \log \hat{\pi}_n^{(k_n)}\right] \\
&= \sum_{t=1}^{T}\sum_{k_x=1}^{M_x}\sum_{k_n=1}^{M_n} \gamma_t^{(k_x,k_n)} \left[\sum_{i=1}^{D} \log p(y_{t,i}|k_x, k_n) + \log \hat{\pi}_n^{(k_n)}\right]. \qquad (4.85)
\end{aligned}$$

Ésta es la expresión de la función auxiliar que finalmente emplearemos para hallar el valor de los parámetros del modelo de ruido, a saber, $\hat{\mu}_{n,i}^{(k_n)}$, $\hat{\sigma}_{n,i}^{(k_n)}$ y $\hat{\pi}_n^{(k_n)}$ ($k_n = 1, \ldots, M_n; i = 1, \ldots, D$). Para calcular estos parámetros, simplemente derivamos la función $\mathcal{Q}(\mathcal{M}_n, \hat{\mathcal{M}}_n)$ respecto al parámetro que queremos estimar e igualamos a cero.

En este apartado únicamente mostraremos las expresiones finales para la reestimación de los parámetros del modelo, omitiendo, por tanto, los pasos intermedios que nos conducen hasta ellas. El desarrollo detallado del algoritmo EM puede consultarse en el apéndice C. Antes de presentar las ecuaciones finalmente obtenidas, definimos a continuación algunos términos que aparecen en el algoritmo EM propuesto y que no hemos definido antes. En primer lugar, la probabilidad a posteriori de la gaussiana k_n-ésima se define como

$$\gamma_t^{(k_n)} = \sum_{k_x=1}^{M_x} \gamma_t^{(k_x,k_n)}. \tag{4.86}$$

Igualmente definimos a partir de (4.78) la máscara de segregación calculada a partir de la gaussiana k_n-ésima del GMM como sigue

$$m_{t,i}^{(k_n)} = \sum_{k_x=1}^{M_x} \gamma_t^{(k_x,k_n)} w_{t,i}^{(k_x,k_n)}, \tag{4.87}$$

donde $w_{t,i}^{(k_x,k_n)}$ define la probabilidad de no enmascaramiento de la voz dadas el par de gaussianas k_x y k_n. Esta probabilidad se calcula usando la ecuación (4.59).

Debido al enmascaramiento que se produce entre los espectros de voz y ruido, habrá situaciones en las que para calcular los parámetros del modelo de ruido (medias y varianzas) tendremos que trabajar con valores estimados. En consecuencia, introducimos la definición del valor estimado para la energía del ruido supuesto que éste fue generado por la gaussiana k_n-ésima del GMM como:

$$\tilde{\mu}_{n,t,i}^{(k_n)} \equiv \mathbb{E}\left[n_{t,i} | n_{t,i} < y_{t,i}, k_n\right] = \mu_{n,i}^{(k_n)} - \sigma_{n,i}^{(k_n)} \frac{\mathcal{N}(\overline{y}_{t,i}^{(k_n)})}{\Phi(\overline{y}_{t,i}^{(k_n)})}, \tag{4.88}$$

donde

$$\overline{y}_{t,i}^{(k_n)} = \frac{y_{t,i} - \mu_{n,i}^{(k_n)}}{\sigma_{n,i}^{(k_n)}}. \tag{4.89}$$

Podemos observar que el valor estimado coincide con la media de la gaussiana k_n-ésima del GMM, supuesto que ésta se trunca superiormente por la observación $y_{t,i}$. Asimismo debe notarse que este valor se calcula usando los parámetros del GMM estimado en la iteración anterior del algoritmo EM, esto es, $\mathcal{M}_n$.

De forma indisolublemente asociada con la estima de la ecuación (4.88), tenemos la varianza de dicha estimación (ver ecuación (B.14)),

$$\tilde{\sigma}_{n,t,i}^{(k_n)^2} \equiv \mathrm{Var}\left[n_{t,i} | n_{t,i} < y_{t,i}, k_n\right]. \tag{4.90}$$

A partir de las variables anteriores, el algoritmo EM procede iterativamente a ajustar los valores de las medias, varianzas y pesos de las gaussianas del GMM, usando para ello los valores obtenidos en la iteración anterior y las características observadas correspondientes a la voz distorsionada. Este ciclo de actualización de los parámetros del modelo se repite mientras que el algoritmo no converja. Más adelante comentaremos algunos aspectos sobre los criterios de parada del algoritmo. A continuación presentamos el algoritmo definitivo para la estimación del modelo de ruido:

- **Inicialización:** estimar $\mathcal{M}_n$ a partir de los N primeros y N últimos segundos de la elocución (suponemos que el inicio y final de la frase es ruido).
- **Bucle principal EM**: repetir los siguientes pasos mientras que el algoritmo no converja.
  - *Paso E*: calcular el valor de $\gamma_t^{(k_n)}$, $m_{t,i}^{(k_n)}$, $\tilde{\mu}_{n,t,i}^{(k_n)}$ y $\tilde{\sigma}_{n,t,i}^{(k_n)^2}$ mediante las ecuaciones (4.86), (4.87), (4.88) y (4.90), usando para ello el modelo $\mathcal{M}_n$ estimado en la iteración anterior.
  - *Paso M*: los parámetros del GMM actualizado $\hat{\mathcal{M}}_n$ vienen dados por las expresiones

$$\begin{aligned}
\hat{\mu}_{n,i}^{(k_n)} &= \frac{\sum_{t=1}^{T} m_{t,i}^{(k_n)} \tilde{\mu}_{n,t,i}^{(k_n)} + \left(\gamma_t^{(k_n)} - m_{t,i}^{(k_n)}\right) y_{t,i}}{\sum_{t=1}^{T} \gamma_t^{(k_n)}}, \\
\hat{\sigma}_{n,i}^{(k_n)^2} &= \frac{\sum_{t=1}^{T} m_{t,i}^{(k_n)} \left[\tilde{\sigma}_{n,t,i}^{(k_n)^2} + \left(\tilde{\mu}_{n,t,i}^{(k_n)} - \hat{\mu}_{n,i}^{(k_n)}\right)^2\right] + \left(\gamma_t^{(k_n)} - m_{t,i}^{(k_n)}\right) \left(y_{t,i} - \hat{\mu}_{n,i}^{(k_n)}\right)^2}{\sum_{t=1}^{T} \gamma_t^{(k_n)}}, \\
\hat{\pi}_n^{(k_n)} &= \frac{1}{T} \sum_{t=1}^{T} \gamma_t^{(k_n)}.
\end{aligned}$$

  - *Actualización:* $\mathcal{M}_n = \hat{\mathcal{M}}_n$.

En el algoritmo anterior, la estimación del modelo inicial de ruido se puede efectuar de varias maneras. Podemos, por ejemplo, estimar la posición inicial de las medias y las varianzas del modelo usando el algoritmo de las k-medias [178]. Otra opción, que es la aquí adoptamos, es usar el propio algoritmo EM para estimar un modelo de mezcla de gaussianas a partir de los primeros y últimos segundos de la frase de *test*. Como estos segmentos de la frase contienen únicamente ruido (es la suposición básica en la que nos basamos), el algoritmo EM usado para ajustar el GMM sería el propuesto originalmente por Dempster en [66].

Otro aspecto que merece la pena comentar del algoritmo EM que se ha presentado es el criterio de parada del mismo. Éste puede ser tan simple como un número de iteraciones máximo o puede tener en cuenta la precisión del GMM estimado en el

modelado de los datos observados. En lo que respecta a este último criterio de parada, debemos decir que la medida natural de la bondad del ajuste del GMM la proporciona la propia verosimilitud de los datos observados que pretendemos maximizar

$$\begin{aligned} L(\boldsymbol{Y}|\hat{\mathcal{M}}_n, \mathcal{M}_x) &= \log p(\boldsymbol{Y}|\hat{\mathcal{M}}_n, \mathcal{M}_x) \\ &= \sum_{t=1}^{T} \log \sum_{k_x=1}^{M_x} \sum_{k_n=1}^{M_n} p(\boldsymbol{y}_t|k_n, k_x, \hat{\mathcal{M}}_n, \mathcal{M}_x). \end{aligned} \tag{4.91}$$

El algoritmo EM nos asegura que la probabilidad anterior no decrece y, por lo tanto, $L(\boldsymbol{Y}|\hat{\mathcal{M}}_n, \mathcal{M}_x) \geq L(\boldsymbol{Y}|\mathcal{M}_n, \mathcal{M}_x)$. Por tanto, podemos usar la diferencia entre las probabilidades de observación calculadas para dos iteraciones consecutivas como criterio de parada, esto es, si $L(\boldsymbol{Y}|\hat{\mathcal{M}}_n, \mathcal{M}_x) - L(\boldsymbol{Y}|\mathcal{M}_n, \mathcal{M}_x) < \epsilon$, entonces el algoritmo EM ha convergido. En esta tesis usaremos una estrategia híbrida de parada que combina un número de iteraciones máximo (típicamente 15) con un umbral de convergencia $\epsilon = 10^{-4}$.

Merece la pena hacer un inciso en uno de los casos especiales que se deducen del algoritmo EM presentado anteriormente: nos referimos al caso en el que el GMM cuenta con una única gaussiana. Por su fácil interpretación, a continuación estudiaremos cómo se simplifican las fórmulas de reestimación de los parámetros del GMM para $M_n = 1$. En primer lugar observamos que las probabilidades a posteriori de las gaussianas del GMM son siempre uno, es decir, $\gamma_t^{(k_n)} = 1$. Asimismo, al haber una única gaussiana, la máscara de segregación $m_{t,i}^{(k_n)}$ coincide con la máscara global $m_{t,i}$. Si aplicamos estos dos cambios, podemos comprobar que la fórmula de reestimación de las medias se expresa ahora de la siguiente forma

$$\hat{\mu}_{n,i} = \frac{1}{T} \sum_{t=1}^{T} m_{t,i} \tilde{\mu}_{n,t,i} + (1 - m_{t,i})\, y_{t,i}, \tag{4.92}$$

que es la misma expresión que intuitivamente derivábamos en la ecuación (4.80).

Lo mismo ocurre con la fórmula de reestimación de las varianzas del GMM,

$$\hat{\sigma}_{n,i}^2 = \frac{1}{T} \sum_{t=1}^{T} m_{t,i} \left[\tilde{\sigma}_{n,t,i}^2 + (\tilde{\mu}_{n,t,i} - \hat{\mu}_{n,i})^2 \right] + (1 - m_{t,i}) \left(y_{t,i} - \hat{\mu}_{n,i} \right)^2. \tag{4.93}$$

Frente a la expresión clásica para el cálculo de la varianza muestral, observamos que la ecuación anterior cuenta además con la expresión $\tilde{\sigma}_{n,t,i}^2 + (\tilde{\mu}_{n,t,i} - \hat{\mu}_{n,i})^2$ multiplicada por el valor de la máscara. Esta expresión modela el error cuadrático de la observación cuando la voz es dominante. En dicha situación, en el cálculo de la varianza se deberá trabajar con las estimaciones de ruido $\tilde{\mu}_{n,t,i}$ y no con los valores observados $y_{t,i}$. Es por ello que dicha expresión cuenta con dos sumandos: $\tilde{\sigma}_{n,t,i}^2$ y $(\tilde{\mu}_{n,t,i} - \hat{\mu}_{n,i})^2$. El primero es

la varianza asociada a la estimación $\tilde{\mu}_{n,t,i}$ y se relaciona con la probabilidad acumulada de la cola de la gaussiana en el intervalo $(-\infty, y_{t,i}]$. El segundo término nos indica el error cuadrático cometido al aproximar el valor estimado $\tilde{\mu}_{n,t,i}$ por la media de la gaussiana $\hat{\mu}_{n,i}$.

Finalmente, como no podría ser de otro modo, la probabilidad a priori de la única gaussiana del modelo será la unidad.

4.3.4. Análisis comparativo

La técnica de reconstrucción espectral MMSR que hemos presentado en este sección es similar en algunos aspectos a otras técnicas de compensación de características que podemos encontrar en la literatura. A fin de destacar los aspectos relevantes de cada técnica, en este apartado realizaremos un estudio comparativo de las similitudes y diferencias entre nuestra propuesta y otras técnicas de compensación bien conocidas. En particular, centraremos este estudio en la comparación con las técnicas de imputación basadas en el paradigma de datos perdidos [88, 90, 126, 224] y en la técnica de compensación VTS [199, 239].

4.3.4.1. Comparativa con las técnicas de imputación

Comenzamos analizando las similitudes y diferencias entre nuestra técnica MMSR y las técnicas de imputación basadas en el paradigma de datos perdidos. Dentro de este grupo de técnicas, en primer lugar estudiaremos aquellas diseñadas para trabajar con máscaras de segregación binarias, para posteriormente dirigir nuestra mirada al estudio de las técnicas que trabajan con máscaras continuas. En cualquiera de los dos casos, usen máscaras binaria o máscaras continuas, todas las técnicas de imputación se basan, al igual que nuestra técnica de reconstrucción, en el modelo de enmascaramiento de la voz de la ecuación (4.4).

La diferencia más significativa entre nuestra propuesta MMSR y las técnicas de imputación es el tipo de información a priori que maneja. Mientras que las técnicas de imputación como TGI usan máscaras para segmentar el espectro en regiones fiables y perdidas, MMSR usa modelos a priori del ruido. A pesar de estas diferencias, en este apartado demostraremos que ambos enfoques son equivalentes cuando funcionan en condiciones ideales, es decir, con máscaras oráculo las técnicas de imputación y con estimaciones de ruido sin error la técnica MMSR. Para demostrar tal afirmación, partiremos de la formulación desarrollada para MMSR en la sección 4.3.1 y, mediante las suposiciones oportunas, veremos cómo se deducen las expresiones de las distintas técnicas de imputación (p.ej. TGI).

Como veíamos en la ecuación (4.43), la técnica MMSR se deriva de una estimación MMSE en la que intervienen probabilidades de observación $P(k_x, k_n|\boldsymbol{y})$ y los valores esperados $\hat{x}_i^{(k_x,k_n)}$. Tal y como muestra la ecuación (4.44), las probabilidades a posteriori para esta técnica dependen de las probabilidades de observación $p(\boldsymbol{y}|k_x, k_n)$ que, a su vez, pueden calcularse como

$$p(\boldsymbol{y}|k_x, k_n) = \prod_{i=1}^{D} \iint p(y_i|x_i, n_i)p(x_i|k_x)p(n_i|k_n)dx_i dn_i. \tag{4.94}$$

El punto clave de esta ecuación, como veíamos, es $p(y_i|x_i, n_i)$. Dicha probabilidad limita los valores que y_i puede tomar de forma que, según el modelo de enmascaramiento de la voz, su valor es (ver ecuación (4.54))

$$p\left(y_i|x_i, n_i\right) = \delta_{x_i}(y_i)\mathbb{1}_{n_i \leq x_i} + \delta_{n_i}(y_i)\mathbb{1}_{x_i < n_i}. \tag{4.95}$$

Hasta aquí todas las expresiones que hemos expuesto se refieren a la técnica MMSR. En las técnicas de imputación, no obstante, la situación cambia, ya que disponemos de una máscara (binaria o continua) que nos proporciona información a priori sobre la segmentación del espectro. Esta información puede ser incorporada en la ecuación anterior quedando ésta de la siguiente forma

$$p(y_i|x_i, n_i) = m_i\delta_{x_i}(y_i)\mathbb{1}_{n_i \leq x_i} + (1 - m_i)\delta_{n_i}(y_i)\mathbb{1}_{x_i < n_i}, \tag{4.96}$$

siendo m_i el valor de la máscara para el elemento i-ésimo .

Introduciendo la ecuación anterior en el desarrollo matemático de MMSR podemos derivar la mayor parte de las técnicas de imputación propuestas en la literatura. En el supuesto de usar máscaras binarias, las probabilidades de observación $p(\boldsymbol{y}|k_x, k_n)$ empleadas por el estimador MMSE se transforman en

$$\begin{aligned} p(\boldsymbol{y}|k_x, k_n) &= \prod_{i \in \boldsymbol{s}_r} p(y_i|k_x) \int_{-\infty}^{y_i} p(n_i|k_n)dn_i \times \prod_{j \in \boldsymbol{s}_u} p(y_j|k_x) \int_{-\infty}^{y_j} p(x_j|k_x)dx_j \\ &= \varphi_n(\boldsymbol{y}) \prod_{i \in \boldsymbol{s}_r} p(y_i|k_x) \prod_{j \in \boldsymbol{s}_u} \int_{-\infty}^{y_j} p(x_j|k_x)dx_j, \end{aligned} \tag{4.97}$$

donde $\boldsymbol{s}_r$ y $\boldsymbol{s}_u$ son los conjuntos con los índices de los elementos fiables y no fiables del vector $\boldsymbol{y}$, respectivamente, y $\varphi_n(\boldsymbol{y})$ es el siguiente producto que sólo depende del modelo de ruido,

$$\varphi_n(\boldsymbol{y}) = \prod_{i \in \boldsymbol{s}_r} \int_{-\infty}^{y_i} p(n_i|k_n)dn_i \prod_{j \in \boldsymbol{s}_u} p(y_j|k_n). \tag{4.98}$$

Para demostrar la equivalencia entre las técnicas de imputación y la técnica MMSR, tenemos que asumir bien que el modelo de ruido tiene una única componente ($M_n = 1$) o

bien que los parámetros de la única gaussiana son variables en el tiempo. En cualquiera de estos dos casos, $\varphi_n(\boldsymbol{y})$ se comportará como una constante que no tendrá ningún efecto en el cálculo de las probabilidades a posteriori $P(k_x, k_n|\boldsymbol{y})$. En este caso, la probabilidad de observación de la ecuación (4.97) será proporcional a

$$\begin{aligned} p(\boldsymbol{y}|k_x, k_n) &\propto \prod_{i\in\boldsymbol{s}_r} p(y_i|k_x) \prod_{j\in\boldsymbol{s}_u} \int_{-\infty}^{y_j} p(x_j|k_x)dx_j \\ &= p(\boldsymbol{y}_r|k_x) \prod_{j\in\boldsymbol{s}_u} \Phi(y_j|k_x) \end{aligned} \tag{4.99}$$

que, como vemos, es el mismo resultado que se obtuvo para la técnica de imputación TGI en la ecuación (4.32). Con esto se demuestra que partiendo de la formulación desarrollada para la técnica MMSR, se puede llegar a la obtenida para TGI sin más que introducir una segmentación a priori del espectro.

Procediendo de igual forma que en el cálculo de las probabilidades de observación, podemos demostrar fácilmente que los valores esperados $\hat{x}_i^{(k_x,k_n)}$ calculados por las técnicas de imputación pueden derivarse de la formulación desarrollada para MMSR. De nuevo basta con sustituir los términos $p(y_i|x_i, n_i)$ que aparecen en el desarrollo de MMSR por el valor dado (4.96). En tal caso, los valores esperados anteriores resultan ser

$$\hat{x}_i^{(k_x,k_n)} = \begin{cases} y_i, & m_i = 1 \\ \tilde{\mu}_{x,i}^{(k_x)}, & m_i = 0 \end{cases}, \tag{4.100}$$

dicha ecuación coincide con la obtenida para TGI en la sección (4.2), pero usando matrices de covarianza diagonales en lugar de matrices completas.

Al comparar las ecuaciones (4.97) y (4.100) que acabamos de derivar para la técnica de imputación TGI, con las ecuaciones (4.55) y (4.61) relativas a MMSR, constamos la siguiente diferencia entre ambos enfoques. Mientras el uso de máscaras binarias en la técnica de imputación implica una segregación ruda del espectro en elementos perdidos y fiables, el uso de modelos a priori del ruido en MMSR supone una segmentación suave en la que cada elemento del espectro tiene asociada una probabilidad de enmascaramiento. Aunque la segmentación ruda realizada por las máscaras binarias es, en teoría, lo óptimo desde el punto de vista del CASA [271], en situaciones reales en donde la máscara de segregación ha de ser estimada, el uso de segmentaciones probabilísticas otorga una mayor robustez a las técnicas de reconstrucción frente a errores en la máscara [31].

De forma similar a la hecho para la imputación basada máscaras binarias, las técnicas basadas máscaras continuas pueden considerarse como casos especiales de la técnica MMSR. Para demostrar tal afirmación, partimos igualmente de la formulación desarrollada para MMSR, pero volviendo a emplear el cómputo modificado de las

probabilidades $p(y_i|x_i, n_i)$ dado en (4.96). Usando estas probabilidades modificadas y máscaras continuas, la probabilidad de observación $p(\boldsymbol{y}|k_x, k_n)$ dada por la ecuación (4.94) para MMSR se transforma en

$$p(\boldsymbol{y}|k_x, k_n) = \prod_{i=1}^{D} m_i p(y_i|k_x)\Phi(y_i|k_n) + (1 - m_i)p(y_i|k_n)\Phi(y_i|k_x). \tag{4.101}$$

Procediendo de forma análoga, puede demostrarse que las estimas parciales de voz se calculan como

$$\hat{x}_i^{(k_x,k_n)} = m_i y_i + (1 - m_i)\tilde{\mu}_{x,i}^{(k_x)}. \tag{4.102}$$

Por tanto, el valor de voz finalmente estimado por las técnicas de imputación es (ver [90, 225])

$$\hat{x}_i = m_i y_i + (1 - m_i) \sum_{k_x=1}^{M_x} \sum_{k_n=1}^{M_n} P(k_x, k_n|\boldsymbol{y})\tilde{\mu}_{x,i}^{(k_x)}. \tag{4.103}$$

Comparando la ecuación anterior con la obtenida para la técnica MMSR en la ecuación (4.62) observamos dos diferencias. En primer lugar, las técnicas de imputación requieren de una segmentación a priori del espectro, mientras que en MMSR esa segmentación se calcula automáticamente a partir de las estimas de ruido. De hecho en la sección 4.3.2 veíamos que el método de reconstrucción proporciona sus propias máscaras de segmentación que pueden emplearse en combinación, por ejemplo, con las técnicas de imputación estudiadas. La segunda diferencia es más sutil que la primera y la apreciamos comparando las ecuaciones (4.101) y (4.55). Mientras que en la técnica MMSR las probabilidades de observación $p(y_i|k_x, k_n)$ sólo dependen de las probabilidades dadas por los modelos de voz y ruido, en la imputación usando máscaras continuas también influye el valor de la máscara. Esto genera cierta redundancia al considerar, por un lado, las probabilidades dadas por el modelo de ruido y, por otro, la información de la máscara.

4.3.4.2. Comparativa con VTS

Para concluir el análisis de la técnica MMSR, no podemos obviar la comparación con VTS, que puede considerarse, a día de hoy, como el estado del arte dentro de las técnicas de compensación de características. De las múltiples versiones y extensiones que han ido surgido de esta técnica, nos centraremos en la versión propuesta originalmente por Moreno en [199] por ser la más conocida.

En primer lugar, observamos que MMSR y VTS se basan en modelos analíticos que describen la forma en la que se distorsiona la voz por acción del ruido. Esto hace

que ambas técnicas sean muy eficientes y que puedan funcionar en entornos desconocidos siempre que, como se ha dicho varias veces, se disponen de estimaciones de las características del ruido presente en dicho entorno. Otra característica que comparten MMSR y VTS es que ambas son, en su concepción, estimadores MMSE de las características de voz. Para derivar los estimadores correspondientes, MMSR y VTS emplean GMMs como modelos modelos de fuente, aunque también podrían emplearse modelos más precisos (p.ej. HMMs). Como último punto en común, debemos comentar la gran flexibilidad que ofrecen los modelos de distorsión de los que se derivan ambas técnicas: el desarrollo en series vectoriales de Taylor (VTS) y el modelo de enmascaramiento (MMSR). Además de aplicados a la compensación de características, estos modelos de distorsión han sido usados para derivar distintas técnicas de adaptación modelos [9, 267], o como base para desarrollar algoritmos iterativos para la estimación de los modelos de ruido requeridos por VTS y MMSR (ver secciones 2.3.2 y 4.3.3).

En cuanto a las diferencias entre VTS y MMSR, la primera de ellas es obvia: como se ha dicho, los modelos de distorsión en los que se basan son diferentes, aunque se derivan de la misma función de interacción, a saber, $\boldsymbol{y} = \log(e^{\boldsymbol{x}} + e^{\boldsymbol{n}})$.

Otro punto en el que se diferencian ambas técnicas es la forma final que toma el modelo de voz distorsionada $\mathcal{M}_y$ que estiman. Suponiendo que se parte de sendos GMMs para modelar voz y ruido, cada PDF del modelo $\mathcal{M}_y$ equivaldrá a la combinación de una PDF del modelo de voz limpia y otra del modelo de ruido. En la técnica VTS, tal y como aparece reflejado en la figura 2.9, cada PDF de $\mathcal{M}_y$ se aproxima mediante una gaussiana, incluso si con ello se incurre en un gran error debido a la no linealidad generada por el ruido. En el caso de la técnica MMSR, las PDFs del modelo de voz distorsionada se aproximan mediante una combinación de las PDFs y CDFs de los modelos originales. Como se aprecia en la figura 4.2, la PDF resultante no es gaussiana y presenta una característica no lineal muy pronunciada para SNRs intermedias. Consideramos que el poder expresar el modelo $\mathcal{M}_y$ como mezcla de PDFs y CDFs resulta muy atractivo, ya que abre la puerta al uso de otras funciones de distribución de probabilidad continuas para el modelado de las fuentes sonoras (voz y/o ruido). La única restricción que deben cumplir estas funciones de probabilidad es que permitan el cálculo analítico de su PDF y su CDF.

La última diferencia que queremos hacer notar en relación a las técnicas MMSR y VTS se refiere al dominio en el que se aplican. A lo largo de todo esto capítulo hemos comentado que el modelo de enmascaramiento de la ecuación (4.4) se expresa de forma natural en el dominio log-Mel, o en cualquier otro dominio que suponga una transformación de compresión (logaritmo o raíces de diferente orden) del espectro obtenido tras aplicar una FFT. La única restricción que se impone para que este modelo

sea útil es que no puede haber ninguna otra transformación involucrada en el cálculo, como es el caso del cepstrum. Esto no ocurre así en la técnica VTS, la cual puede aplicarse indistintamente en ambos dominios (log-Mel y cepstrum). De hecho lo usual es aplicar esta última en el dominio cepstral y aprovechar el uso de matrices de covarianza diagonales en los modelos.

4.4. Resumen

Este capítulo ha tratado sobre el modelo de enmascaramiento de la voz y su aplicación a la compensación de las características de voz distorsionadas por ruido acústico. En la sección 4.1 dicho modelo se ha derivado como aproximación de la función de interacción exacta entre dos fuentes sonoras. De acuerdo a este modelo, la interacción entre dos fuentes (p.ej. voz y ruido) en el dominio log-Mel se simplifica a un problema de enmascaramiento, esto es, la energía de la voz domina al ruido o viceversa. Esta visión alternativa del efecto del ruido sobre las características de voz permite reformular el problema de compensación de características como dos problemas nuevos: estimación de la fiabilidad (grado de enmascaramiento) de los elementos del espectro y estimación de aquellos enmascarados por el ruido.

El primer intento para abordar la reconstrucción de espectros incompletos se ha centrado únicamente en el segundo problema de los dos mencionados. Así, en la sección 4.2 veíamos que la técnica TGI asume el conocimiento a priori sobre la segmentación del espectro en regiones enmascaradas por el ruido y regiones que no han sufrido distorsión alguna. En base a esta información, TGI deriva el estimador oportuno para reconstruir la información perdida en las regiones enmascaradas, dado que se conoce el espectro de la voz en las regiones sin enmascarar. Para llevar a cabo esta tarea, TGI explota la información recogida en modelos a priori de la voz en forma de correlaciones entre las regiones enmascaradas y sin enmascarar.

Por otra parte, la técnica MMSR propuesta en la sección 4.3 aborda ambos problemas de forma conjunta al emplear modelos estadísticos de ruido en lugar de máscaras de segregación. Esto le permite a MMSR estimar de forma conjunta la energía de la voz en las regiones enmascaradas y la fiabilidad de los elementos del espectro. Así, partiendo de sendos modelos (GMMs) para la voz y el ruido, esta técnica es capaz de, por un lado, reconstruir el espectro de la voz en aquellas regiones enmascaradas por el ruido (sección 4.3.1) y, por otro, calcular una máscara de segregación que indique el grado de fiabilidad de cada elemento del espectro (sección 4.3.2). Asimismo, en la sección 4.3.3 se vio cómo, a partir de la formulación matemática desarrollada para la técnica MMSR, se puede derivar un algoritmo iterativo para la estimación de los parámetros de los

GMMs de ruido que esta técnica emplea. La idea intuitiva detrás del algoritmo se basa en la observación de que podemos emplear aquellas regiones del espectro identificadas como ruido en la máscara de segregación para estimar los parámetros del modelo.

CAPÍTULO 5

Modelado temporal y tratamiento de la incertidumbre

Las técnicas de compensación de características que se han propuesto en los capítulos previos muestran varios atributos que las hacen muy atractivas de cara a su implantación en sistemas de reconocimiento reales. Para empezar, estas técnicas suelen ser mucho más eficientes, en tiempo y memoria, que las técnicas de adaptación. En efecto, mientras que la adaptación del sistema de reconocimiento a la voz de un locutor suele efectuarse durante una primera fase personalización del sistema (conocida como *enrollment*), la adaptación al ruido acústico requiere modificar continuamente los parámetros del modelo acústico para que éste refleje las características cambiantes del ruido. Hacer esto en tiempo real, como puede suponerse, es inviable en sistemas de reconocimiento de gran vocabulario que cuentan con decenas de miles de parámetros. En cambio, las técnicas de compensación suelen conllevar un coste computacional menor al usar modelos de voz más simples que los usados por el reconocedor. Asimismo, otra de las ventajas de la compensación es su independencia del reconocedor, lo que implica que de cara a éste la etapa de compensación es transparente.

A pesar de estas ventajas de la compensación de características, las técnicas de adaptación suelen ofrecer un mayor rendimiento. Así en repetidas ocasiones se ha demostrado experimentalmente que al implementar una misma técnica (p.ej. VTS [9, 199] o Algonquin [94, 95, 165]) tanto en su versión orientada a la adaptación de modelos como en su versión de compensación de características, la versión diseñada para adaptar los modelos ofrece mejores resultados de reconocimiento (ver p.ej. [172]). Las razones de este comportamiento las podemos encontrar en los dos aspectos clave que se detallan a continuación.

El primero de ellos es el modelado acústico empleado ambos enfoques. En este sentido observamos que mientras las técnicas de compensación suelen usar modelos de voz relativamente simples (p.ej. GMMs con unos pocos cientos de gaussianas), los modelos acústicos del reconocedor suelen ser bastante más complejos y, posiblemente, entrenados en base a algún criterio discriminatorio entre las unidades acústicas. Asimismo, los GMMs de voz usados por las técnicas de compensación suelen modelar únicamente la distribución en frecuencia de la voz, mientras que una componente importante de los HMMs del reconocedor es el modelado temporal. A lo largo de este capítulo estudiaremos más detenidamente este detalle.

La segunda razón tiene que ver con el carácter inherentemente aproximado de las técnicas de compensación. En particular vemos que estas técnicas condensan la información recogida en la distribución a posteriori $p(\boldsymbol{x}|\boldsymbol{y})$ en un único valor de tendencia central (p.ej. la media en el caso de la estimación MMSE o la moda en la estimación MAP), generando con ello una pérdida importante de información que pueda ser útil para el reconocimiento. En las técnicas de adaptación, por contra, la voz distorsionada se descodifica usando la distribución $p(\boldsymbol{y})$ al completo. Esto último puede considerarse, desde el punto de vista estadístico, como la estrategia óptima de reconocimiento para la señal de voz distorsionada.

A fin de mejorar el rendimiento de las técnicas de compensación propuestas en los capítulos anteriores y equipararlo al alcanzado por la adaptación de modelos, en este capítulo nos planteamos la mejora del modelado estadístico de las características de voz incluyendo correlaciones temporales, así como el tratamiento de la incertidumbre del proceso de estimación en el reconocedor. La primera parte de este capítulo se dedicará al estudio del modelado y modelado de la redundancia temporal de la voz, presentando dos propuestas para alcanzar tal fin. La primera propuesta se basará en un modelado de las correlaciones temporales de orden corto presentes en la señal de voz. Como veremos, esta idea se plasmará en un esquema basado en una ventana deslizante diseñada para contener segmentos de voz, del orden de la longitud de un fonema. Estos segmentos serán los que posteriormente se usarán para estimar los modelos de voz oportunos (GMMs o diccionarios VQ). Frente al modelado de las correlaciones corto plazo, la segunda propuesta que describiremos empleará modelos ocultos de Markov para representar la distribución de la señal de voz tanto en frecuencia como en tiempo.

La segunda parte de este capítulo estará dedicada al estudio del tratamiento, por parte del reconocedor, de la incertidumbre de las estimas de voz. Como puede suponerse, los espectros de voz obtenidos por las técnicas de estimación MMSE propuestas contendrán errores debidos a factores tales como la SNR de la señal observada, la aleatoriedad del ruido, la precisión de las estimas de ruido empleadas y la bondad de ajuste

del modelo de voz empleado. Hasta ahora esta imprecisión era desconocida para el reconocedor, pues éste veía únicamente estimas puntuales de la señal voz. Para que el reconocedor tenga constancia explícita de la incertidumbre asociada al proceso de estimación, se propondrán distintos métodos para el cómputo de medidas de incertidumbre y su uso dentro del reconocedor. El objetivo que perseguimos es que los vectores de características peor estimados tengan un peso menor en la etapa de descodificación.

5.1. Modelado temporal de la voz

Una de las características básicas de la voz que entronca con los fundamentos de la audición humana es su redundancia. En presencia de ruido, la redundancia de la voz ayuda a desambiguar posibles casos de confusión entre distintas unidades acústicas y, de esta forma, llegar a comprender el mensaje contenido en esta señal. Así lo han demostrado numerosos experimentos psicoacústicos en los que la inteligibilidad de la voz apenas se ve mermada pese a sufrir severas distorsiones como la supresión de bandas de frecuencias exclusivas [15, 93, 273, 274].

Hasta ahora, en las técnicas de compensación que hemos presentado únicamente habíamos considerado la redundancia en frecuencia de la voz. No obstante, la redundancia temporal es una información que también ha demostrado ser muy útil de cara a su uso en el procesado digital de señales de voz degradadas. En este sentido, trabajos como el de Ephraim y Malah [83] fueron pioneros en la aplicación de modelos de fuente más complejos en el campo del realce estadístico de señales de voz. Desde entonces una multitud de trabajos [39, 122, 127, 161, 252] ha confirmado que el uso de la información temporal redunda en una mejora consistente de las estimas de voz obtenidas por las técnicas de compensación de características. No sólo la información temporal ha resultado ser útil para la compensación de las características usadas por los reconocedores de voz, sino que esta información es también esencial para la mitigación de las pérdidas producidas durante la transmisión de la voz [38, 47, 48, 92, 118, 148, 214, 231].

Con objeto de ejemplificar el modo en el que las correlaciones temporales pueden ser útiles para compensar las características de voz, en la figura 5.1 mostramos el resultado de aplicar dos versiones de la técnica TGI (ver sección 4.2) sobre un espectro de voz ruidosa: cuando sólo se explotan las correlaciones en frecuencia (*reconstrucción en frecuencia*) y cuando se explotan ambas correlaciones usando el enfoque que presentaremos en la sección 5.1.1 (*reconstrucción en tiempo-frecuencia*). Recordemos que esta técnica tiene por cometido la estimación de las regiones del espectro afectadas por el ruido, supuesto que la localización de éstas es conocida. Para ello, en el ejemplo de la figura empleamos una máscara oráculo que identifica las regiones fiables del espectro

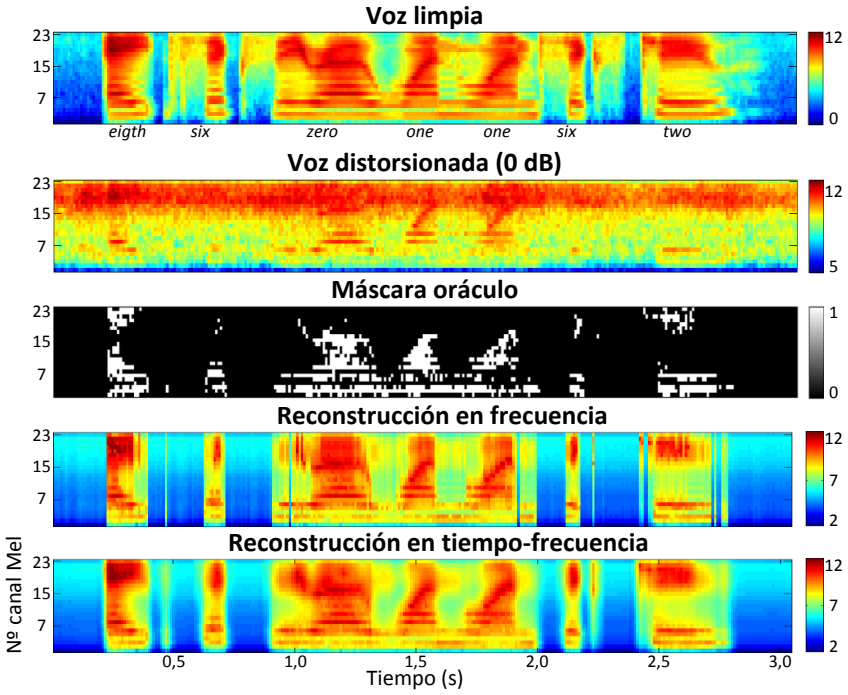

Figura 5.1: Mejora del proceso de reconstrucción espectral por la explotación de las correlaciones temporales de la voz.

con el color blanco, y las regiones afectadas por el ruido con el color negro.

A la vista de los espectros de la figura podemos constatar lo siguiente. Primero, la distorsión producida por el ruido en la voz se traduce en una pérdida de la información codificada en esta señal (por ejemplo las altas frecuencias de los fonemas /s/ localizados en torno a los instantes de tiempo 0,57 s y 2,07 s). Esta pérdida de información, que puede llegar a ser particularmente importante a bajas SNRs, es irrecuperable con las técnicas de compensación descritas hasta ahora, salvo que se impongan otras restricciones adicionales en el proceso de reconstrucción espectral. En este sentido, las restricciones temporales y las que impone el modelo de lenguaje pueden ser especialmente útiles para esta tarea.

En segundo lugar observamos que el espectro reconstruido empleando únicamente correlaciones en frecuencia presenta un notable bandeado vertical con discontinuidades entre vectores de características vecinos[1] (p.ej. en torno al último dígito de la elocución en el instante de tiempo 2,5 s). Llevadas al extremo, estas discontinuidades hacen que en el espectro estimado aparezcan líneas verticales cuando algún vector de características

[1]Estas discontinuidades pueden generar errores de inserción y sustitución durante la etapa de reconocimiento.

con ciertos elementos fiables está rodeado por regiones completamente perdidas (p.ej. en 0,47 y 2,23 s) o al contrario, esto es, regiones perdidas rodeadas por voz (p.ej. en los instantes de tiempo 0,98; 1,92; y 2,72 s). De nuevo estas discontinuidades se podrían aliviar imponiendo restricciones de suavidad temporal en el proceso de reconstrucción espectral.

También constatamos que las discontinuidades del espectro estimado son más pronunciadas en las regiones menos fiables del espectro original. Por ejemplo, la reconstrucción obtenida para el dígito *zero* es más suave que la calculada para el dígito *two*, conteniendo este último dígito una menor proporción de elementos fiables. En la sección 5.2 veremos que esto se debe a que la varianza de la estimación MMSE es mayor cuanto menor sea el número de elementos fiables. A fin de reducir esta varianza, una solución lógica es explotar la redundancia temporal de la voz, de forma que al estimar las características ruidosas se tenga en cuenta, por ejemplo, las características fiables cercanas (en tiempo o frecuencia).

Finalmente, la última gráfica de la figura 5.1 muestra el espectro estimado por la técnica TGI cuando ésta explota tanto las correlaciones en frecuencia como en tiempo durante la reconstrucción. Como se puede apreciar, el uso de ambas correlaciones redunda en una estimación más correcta del espectro de voz original, así como en una reconstrucción más suave del mismo.

Una vez vistas las bondades de emplear información sobre la evolución temporal de la voz durante la compensación de características, en los siguientes apartados describiremos los dos enfoques que aquí adoptamos para tal propósito: modelado de las correlaciones temporales de corta duración y modelado HMM.

5.1.1. Modelado de las correlaciones temporales de orden corto

En este aparatado nos planteamos la extensión del modelado de la voz que se venía utilizando hasta ahora para que, en lugar de representar vectores de características aislados, los modelos estadísticos representen pequeños segmentos de voz cuya longitud sea similar a la de un fonema. En este sentido cabe destacar que una de las estrategias más simples y efectivas para lograr este objetivo, consiste en modelar conjuntamente los parámetros estáticos y dinámicos de la voz. Como veremos en el capítulo 6, la inclusión de las características dinámicas en los modelos de voz supone una mejora relativa media de aproximadamente el 5 % en tasa de palabras correctamente reconocidas.

En lugar de modelar conjuntamente las características estáticas y dinámicas, en esta sección introducimos un modelado alternativo basado en la representación de seg-

mentos de voz de corta duración con τ vectores de características consecutivos. Estos segmentos (o parches) se obtienen aplicando un mecanismo de ventana deslizante sobre la elocución de entrada,

$$\boldsymbol{z}_t = \left(\boldsymbol{x}_{\delta\cdot(t-1)+1}^\top, \ldots, \boldsymbol{x}_{\delta\cdot(t-1)+\tau}^\top\right)^\top, \tag{5.1}$$

donde δ es el desplazamiento o incremento de la ventana. Por ejemplo, si $\tau = 3$ y $\delta = 2$ entonces $\boldsymbol{z}_1 = (\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \boldsymbol{x}_3^\top)^\top$, $\boldsymbol{z}_2 = (\boldsymbol{x}_3^\top, \boldsymbol{x}_4^\top, \boldsymbol{x}_5^\top)^\top$, etc. Como podemos comprobar, dos segmentos consecutivos $\boldsymbol{z}_{t+1}$ y $\boldsymbol{z}_t$ se solapan $\tau - \delta$ tramas.

Los nuevos vectores de características $\boldsymbol{z}_t$ definidos en la ecuación (5.1) pueden usarse ahora para entrenar modelos de mezcla de PDFs (p.ej. GMMs o diccionarios VQ) de la forma usual. No obstante, la alta dimensión de los supervectores $\boldsymbol{z}_t$ puede acarrear una mala estimación de los parámetros del modelo estadístico[1]. Para evitar esto, recurrimos a un análisis de componentes principales (PCA, *Principal Component Analysis*) [36, 150] a fin de reducir la dimensión de los vectores $\boldsymbol{z}_t$ antes de entrenar los modelos a priori de la voz. PCA emplea una transformación lineal $\boldsymbol{P}$ para proyectar $\boldsymbol{z}_t$ a un espacio ortogonal de menor dimensión que el original, consiguiendo con ello reducir el número de elementos de los vectores de entrenamiento sufriendo pérdida de información mínima. La transformación lineal que PCA define es

$$\boldsymbol{z}'_t = \boldsymbol{P}^\top (\boldsymbol{z}_t - \boldsymbol{\mu}_z)\,. \tag{5.2}$$

En la ecuación anterior, $\boldsymbol{\mu}_z$ denota la media de los vectores de entrenamiento en el espacio $Z = \mathbb{R}^n$ de los segmentos de voz,

$$\boldsymbol{\mu}_z = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{z}_t. \tag{5.3}$$

Por otra parte, el operador lineal $\boldsymbol{P}$ empleado por PCA se obtiene a partir de la matriz de covarianza global de los datos de entrenamiento. Esta matriz se calcula como sigue:

$$\boldsymbol{\Sigma}_z = \frac{1}{T-1}\sum_{t=1}^{T} (\boldsymbol{z}_t - \boldsymbol{\mu}_z)\,(\boldsymbol{z}_t - \boldsymbol{\mu}_z)^\top\,. \tag{5.4}$$

Al ser $\boldsymbol{\Sigma}_z$ simétrica, la podemos expresar en base a su matriz de autovectores $\boldsymbol{V}$ y a la matriz diagonal autovalores $\boldsymbol{\Lambda}$,

$$\boldsymbol{\Sigma}_z = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^\top. \tag{5.5}$$

[1] La dimensión de $\boldsymbol{z}_t$ es $n = \tau \times D$, donde D es la dimensión de los vectores de características originales, típicamente 23 si se trata de características log-Mel o 13 si son coeficientes cepstrales.

Finalmente, asumiendo que los elementos de la matriz $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ se encuentran ordenados de mayor a menor, $\boldsymbol{P}$ equivale a la submatriz con las primeras n^* columnas de $\boldsymbol{V}$. Suponiendo que el porcentaje de varianza del espacio original Z que se retiene se denota por ρ (típicamente $\rho = 0,95$), entonces n^* viene dado por

$$n^* = \text{máx}\left\{n : (n = 1, \ldots, \tau \cdot D) \wedge (\phi(n) \leq \rho)\right\}, \tag{5.6}$$

donde $\phi(n)$ es el porcentaje de varianza que se retiene con los n primeros autovectores,

$$\phi(n) = \frac{\sum_{i=1}^{n} \lambda_i}{\text{tr}(\boldsymbol{\Lambda})} \tag{5.7}$$

y $\text{tr}(\Lambda)$ denota la traza de la matriz Λ

Tras reducir la dimensión de los datos de entrenamiento con PCA, podemos proceder a entrenar un modelo estadístico que ahora representará tanto la distribución en frecuencia de la voz, como sus variaciones temporales de corta duración. En el caso de las técnicas de compensación basadas en datos estéreo de la sección 3.2, esto supone entrenar sendos diccionarios VQ: uno para las grabaciones de voz sin distorsionar y otro para las grabaciones degradadas. En base a estos diccionarios y a las grabaciones estéreo disponibles, se estiman los parámetros de las transformaciones de compensación empleadas por esas técnicas. Así, por ejemplo, se calcularían las medias y matrices de covarianza de las subregiones contenidas en cada celda VQ. Por último, tras compensar las características de voz en el espacio PCA, se aplicaría la siguiente transformación inversa para expresar los parámetros de voz en el dominio Z de los segmentos de voz:

$$\hat{\boldsymbol{z}}_t = \boldsymbol{P}\hat{\boldsymbol{z}}'_t + \boldsymbol{\mu}_z, \tag{5.8}$$

siendo $\hat{\boldsymbol{z}}'_t$ el vector de características estimado en el espacio de características $Z' = \mathbb{R}^{n^*}$ definido por la proyección PCA.

En el procedimiento que acabamos de describir, la compensación de la voz se efectúa en el espacio Z', pues es en este espacio en el que se expresan los parámetros de los diccionarios VQ. Un esquema de compensación alternativo consistiría en, previamente al cálculo de las estimas de voz correspondientes, expresar los parámetros de los modelos de voz en el espacio Z original. Es decir, el espacio Z' únicamente se emplea para entrenar más robustamente el GMM, mientras que la compensación se sigue realizando en el espacio original Z. Suponiendo sin pérdida de generalidad que el modelo de voz fuese un GMM con M componentes, los parámetros $(\boldsymbol{\mu}_{z'}^{(k)}, \boldsymbol{\Sigma}_{z'}^{(k)})$ de la gaussiana k-ésima en Z' se transformarían a Z de la siguiente forma:

$$\boldsymbol{\mu}_z^{(k)} = \boldsymbol{P}\boldsymbol{\mu}_{z'}^{(k)} + \boldsymbol{\mu}_z, \tag{5.9}$$

$$\boldsymbol{\Sigma}_z^{(k)} = \boldsymbol{P}\boldsymbol{\Sigma}_{z'}^{(k)}\boldsymbol{P}^\top + \boldsymbol{\Sigma}_r. \tag{5.10}$$

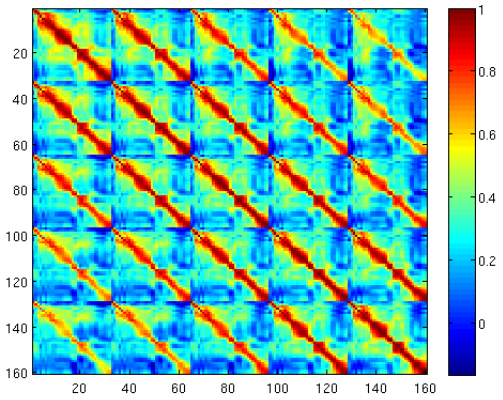

Figura 5.2: Correlaciones estimadas para segmentos cortos de características log-Mel de longitud 5.

La matriz $\boldsymbol{\Sigma}_r$ de la expresión anterior representa la varianza residual del conjunto de datos de entrenamiento no considerada por la técnica PCA. Dicha matriz, que se emplea para evitar que el rango de $\boldsymbol{\Sigma}_z^{(k)}$ sea deficiente, se calcula de la siguiente forma

$$\boldsymbol{\Sigma}_r = \boldsymbol{R}\boldsymbol{R}^\top \boldsymbol{\Sigma}_z \boldsymbol{R}\boldsymbol{R}^\top, \tag{5.11}$$

siendo $\boldsymbol{R}$ la submatriz con los $n - n^*$ autovectores restantes de $\boldsymbol{\Sigma}_z$ no considerados en $\boldsymbol{P}$, de forma que, $\boldsymbol{V} = [\boldsymbol{P}\boldsymbol{R}]$.

A modo de ejemplo, la figura 5.2 muestra la matriz de correlación obtenida tras aplicar la transformación de la ecuación (5.10) a una gaussiana entrenada con características PCA. Los segmentos de voz de partida consisten en secuencias de $\tau = 5$ vectores con $D = 32$ características log-Mel obtenidas siguiendo el esquema de ventana deslizante de la ecuación (5.1) con un desplazamiento de la ventana igual a $\delta = 1$ vector, por lo que $\boldsymbol{z}_t = (x_t, x_{t+1}, \ldots, x_{t+159})^\top$. En la figura se puede apreciar la gran correlación que existe entre características log-Mel próximas, tanto en frecuencia como en tiempo. Conforme la distancia entre dos elementos aumenta, la correlación disminuye.

Este procedimiento alternativo para el cómputo de los valores estimados de voz es, de hecho, el único procedimiento viable cuando consideramos las técnicas de reconstrucción espectral estudiadas en el capítulo 4. Como comentábamos en dicho capítulo, el modelo de enmascaramiento en el que se basan estas técnicas segrega el espectro de la señal observada en distintas regiones dominadas bien por la energía de la voz, o bien por la del ruido. Dicha segregación se expresa de forma fácil y natural en el dominio del banco de filtros, pero no así en otros dominios que supongan una transformación de las características log-Mel, como por ejemplo el cepstrum o el dominio PCA que

acabamos de describir. Por lo tanto, en estas técnicas sólo se aplicará el análisis PCA durante la fase de entrenamiento de los modelos de voz correspondientes. Tras esta fase, la transformación aplicada por PCA será deshecha para expresar los modelos en el dominio original.

Después de aplicar las técnicas de compensación oportunas sobre los distintos segmentos de voz $\boldsymbol{z}_t$ de las frases de *test*, obtendremos una serie de supervectores estimados $\hat{\boldsymbol{z}}_t$ los cuales, a su vez, se compondrán de una serie de vectores de características estimados $\hat{\boldsymbol{z}}_t = (\hat{\boldsymbol{x}}^\top_{\delta\cdot(t-1)+1}, \ldots, \hat{\boldsymbol{x}}^\top_{\delta\cdot(t-1)+\tau})^\top$. Como consecuencia del solapamiento que se produce entre segmentos consecutivos, se da el hecho de que una misma observación $\boldsymbol{y}_t$ puede tener asociadas varias estimas $\hat{\boldsymbol{x}}_t(l)$ $(l = 1, \ldots, L)$. Para obtener el valor estimado final $\hat{\boldsymbol{x}}_t$, la estimas anteriores se combinan linealmente en función de su valor de confianza c_l,

$$\hat{\boldsymbol{x}}_t = \sum_{l=1}^{L} c_l \hat{\boldsymbol{x}}_t(l). \tag{5.12}$$

Para que la estima sea consistente, los valores de confianza deben cumplir la restricción $\sum_{l=1}^{L} c_l = 1$. Asimismo, los segmentos de voz mejor estimados deberían contar con valores de confianza más altos. La cuestión de cómo se estiman los valores de confianza será estudiada en detalle en la sección 5.2.

5.1.2. Modelado de la voz usando modelos ocultos de Markov

Los modelos ocultos de Markov son una de las herramientas más conocidas en el área del reconocimiento automático del habla. Desde sus primeros inicios a finales de los años 60 del siglo pasado, los HMMs no sólo han demostrado su potencia y versatilidad en el modelado acústico de la voz [220], sino también en otras tareas donde la evolución temporal de la señal de interés es relevante. En este apartado se emplearán los HMMs como sustitutos más potentes de los modelos a priori de voz empleados por las técnicas de compensación descritas en los capítulos 3 y 4. En todo momento debe quedar claro que HMMs más complejos, como los usados por el propio reconocedor de voz, repercutirán en estimaciones más precisas de las características degradadas de voz [184]. Por tanto, su uso está más que recomendado en sistema de reconocimiento comerciales. En nuestro caso, los HMMs que emplearemos serán versiones aumentadas de los modelos de voz previamente usados (GMMs o diccionarios VQs) a los que se les ha agregado una matriz con las probabilidades de transición entre los estados del HMM. Por tanto, cada estado del HMM se corresponderá con una componente (gaussiana o celda VQ) del modelo de voz original y la topología

del modelo será ergódica (completamente conectada). Estos HMMs simplificados nos servirán como prueba de concepto para justificar la validez del enfoque de compensación que aquí proponemos.

El HMM usado para compensar la voz ruidosa se define mediante la tupla $\mathcal{M}_x = \langle \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi} \rangle$. Los elementos a_{ij} de la matriz $\boldsymbol{A}$ definen las probabilidades de transición entre los estados s_i y s_j $(1 \leq i, j \leq M)$ del HMM, $a_{ij} = P(q_{t+1} = s_j | q_t = s_i)$, donde $(q_1, q_2, \ldots, q_T)$ define la secuencia oculta de estados. El cómputo de estas probabilidades se efectúa promediando las probabilidades de transición entre estados consecutivos q_t y q_{t+1} para el conjunto de datos de entrenamiento:

$$\begin{aligned} a_{ij} &= \frac{\sum_{t=1}^{T} p\left(\boldsymbol{x}_{t+1} | q_{t+1} = s_j\right) p\left(\boldsymbol{x}_t | q_t = s_i\right) P(s_j) P(s_i)}{\sum_{t'=1}^{T} p\left(\boldsymbol{x}_{t'} | q_{t'} = s_i\right) P(s_i)} \\ &= \frac{\sum_{t=1}^{T} p\left(\boldsymbol{x}_{t+1} | q_{t+1} = s_j\right) p\left(\boldsymbol{x}_t | q_t = s_i\right) \pi_j \pi_i}{\sum_{t'=1}^{T} p\left(\boldsymbol{x}_{t'} | q_{t'} = s_i\right) \pi_i}. \end{aligned} \tag{5.13}$$

El vector $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_M)$ contiene las probabilidades a priori de los estados del modelo. Estas probabilidades se corresponderán con los pesos de las componentes del modelo original del que se parte. Por último, $\boldsymbol{B} = (\langle \boldsymbol{\mu}_x^{(1)}, \boldsymbol{\Sigma}_x^{(1)} \rangle, \ldots, \langle \boldsymbol{\mu}_x^{(M)}, \boldsymbol{\Sigma}_x^{(M)} \rangle)$ contiene los parámetros que modelan las M regiones (gaussianas o celdas VQ) en las que se divide el espacio de características. A partir de estos parámetros pueden computarse las probabilidades de observación del HMM.

Para aplicar el modelo de Markov anterior al proceso de compensación de la voz ruidosa, reescribimos la expresión del estimador MMSE descrito en la sección 3.1 para que aparezca reflejada de forma explícita la evolución temporal de la señal de entrada [122, 127, 214],

$$\hat{\boldsymbol{x}}_t = \sum_{k=1}^{M} P\left(q_t = s_k | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_T, \mathcal{M}_x\right) \hat{\boldsymbol{x}}_t^{(k)}. \tag{5.14}$$

En la ecuación anterior, $\hat{\boldsymbol{x}}_t^{(k)} = \mathbb{E}[\boldsymbol{x}_t | \boldsymbol{y}_t, k]$ denota la estima parcial del vector de características para el estado k-ésimo del HMM. En el cálculo de dicho valor esperado no haremos consideraciones sobre la evolución temporal de la voz y, por consiguiente, $\hat{\boldsymbol{x}}_t^{(k)}$ se calcula según las indicaciones dadas para cada técnica de compensación en particular. En cambio, en el cómputo de la probabilidad a posteriori $\gamma_t^{(k)} = P(q_t = s_k | \boldsymbol{y}_1, \ldots, \boldsymbol{y}_T, \mathcal{M}_x)$, sí aparece reflejada de forma explícita la probabilidad del estado k en el instante de tiempo t respecto a la secuencia de vectores observados $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$. Esta probabilidad puede calcularse de manera eficiente mediante el algoritmo de avance-retroceso (*forward-backward algorithm*) [220],

$$\gamma_t^{(k)} = \frac{\alpha_t^{(k)} \beta_t^{(k)}}{\sum_{k'=1}^{M} \alpha_t^{(k')} \beta_t^{(k')}}, \tag{5.15}$$

donde $\alpha_t^{(k)}$ y $\beta_t^{(k)}$ son las probabilidades de avance y retroceso, respectivamente, definidas como:

$$\alpha_t^{(k)} = p\left(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t, q_t = s_k | \mathcal{M}_x\right), \tag{5.16}$$

$$\beta_t^{(k)} = p\left(\boldsymbol{y}_{t+1}, \ldots, \boldsymbol{y}_T | q_t = s_k, \mathcal{M}_x\right). \tag{5.17}$$

Estas probabilidades se calculan, a su vez, de forma recursiva,

$$\alpha_t^{(k)} = \left[\sum_{j=1}^{M} \alpha_{t-1}^{(j)} a_{jk}\right] p\left(\boldsymbol{y}_t | s_k\right), \tag{5.18}$$

$$\beta_t^{(k)} = \sum_{j=1}^{M} a_{kj} p\left(\boldsymbol{y}_{t+1} | s_j\right) \beta_{t+1}^{(j)} \tag{5.19}$$

y su inicialización para los estados $k = 1, \ldots, M$ del HMM es,

$$\alpha_1^{(k)} = \pi_k p\left(\boldsymbol{y}_1 | s_k\right). \tag{5.20}$$

$$\beta_T^{(k)} = 1 \tag{5.21}$$

Al igual que en el caso de las estimas parciales $\hat{\boldsymbol{x}}_t^{(k)}$, el cómputo de las probabilidades de observación $p(\boldsymbol{y}_t|s_k)$ atenderá a las particularidades de cada técnica de compensación en concreto. Por ejemplo, en el caso de las técnicas basadas en datos estéreo, esta probabilidad se aproxima por $p(\boldsymbol{y}_t|s_k) \approx P(k_y^*(t)|k_x)$, siendo $k_y^*(t)$ la celda VQ a la que pertenece la observación $\boldsymbol{y}_t$ y $k_x \equiv s_k$. Aplicando la regla de Bayes podemos expresar esta probabilidad como

$$P(k_y|k_x) = \frac{P(k_x|k_y)P(k_x)}{P(k_y)}, \tag{5.22}$$

para $1 \leq k_x \leq M_x$ y $1 \leq k_y \leq M_y$. Todas las probabilidades de la ecuación anterior son conocidas, tal y como se expuso en el capítulo 3.

En las técnicas de reconstrucción basadas en el paradigma de datos perdidos, TGI y MMSR, para calcular esta probabilidad deberemos suministrar al estimador cierta información a priori sobre el ruido, ya sea una máscara de segregación (TGI) o una estimación de la potencia del ruido (MMSR). Por ejemplo, en la técnica TGI la probabilidad anterior equivale a $p(\boldsymbol{y}_t|s_k) \equiv p(\boldsymbol{y}_t|k) = p(\boldsymbol{y}_{t,u}, \boldsymbol{y}_{t,r}|k)$, cuya expresión viene dada por la ecuación (4.32). En el caso de la técnica MMSR, esta probabilidad se calcula teniendo en cuenta las probabilidades del modelo de ruido, tal y como se recoge en la sección 4.3.1.

5.2. Tratamiento de la incertidumbre de la estimación

Consideremos de nuevo la figura 5.1 con distintas reconstrucciones obtenidas por la técnica TGI para un espectrograma de voz ruidosa. Como ya se comentó anteriormente, podemos constatar que el espectro reconstruido usando únicamente correlaciones en frecuencia (*reconstrucción en frecuencia*) presenta un mayor número de discontinuidades en las regiones menos fiables del espectro original. Esto se justifica por la correlación inversa existente entre la incertidumbre durante el cálculo de $P(k|\boldsymbol{y})$ (las probabilidades a posteriori de las gaussianas del modelo de voz) y el número de elementos fiables del vector observado. Dicho de otra forma, cuanto menor sea el número de elementos fiables de la observación $\boldsymbol{y}$, más uniforme tenderán a ser las probabilidades $P(k|\boldsymbol{y})$ $(k = 1, \ldots, M)$.

A fin de comprobar empíricamente esta afirmación, en la figura 5.3 mostramos la distribución de las probabilidades $P(k|\boldsymbol{y})$ ($M = 256$ gaussianas) para el ejemplo de la figura 5.1. Para facilitar la visualización de las probabilidades, en la figura se han representado las probabilidades acumuladas asociadas a $P(k|\boldsymbol{y})$, después de ser ordenadas de mayor a menor. Basándonos en la gráfica de probabilidades acumuladas, podemos reafirmar lo dicho anteriormente: si la proporción de elementos fiables es alta, entonces la entropía asociada a $P(k|\boldsymbol{y})$ es pequeña, existiendo una gaussiana k^* cuya probabilidad a posteriori es cercana a 1, $P(k^*|\boldsymbol{y}) \approx 1$, y la del resto 0 (p.ej. en torno a los instantes de tiempo 0,30; 1,20; 1,50; 1,85 y 2,55 s) . En cambio, si la proporción de elementos fiables es pequeña o nula, entonces la entropía asociada a la distribución de probabilidad discreta $P(k|\boldsymbol{y})$ será alta y, por consiguiente, también lo será la varianza de la distribución a posteriori $p(\boldsymbol{x}|\boldsymbol{y})$ en la que se basa nuestra estimación MMSE (p.ej. en los periodos de silencio iniciales y finales de las frases).

Partiendo de los comentarios anteriores, se puede concluir que el conocimiento sobre la distribución $p(\boldsymbol{x}|\boldsymbol{y})$ nos da pistas sobre la fiabilidad de los vectores de características estimados. Hasta ahora se ha considerado que todas las estimas de voz obtenidas por las técnicas de compensación son igualmente fiables de cara al reconocimiento. No obstante, en la figura 5.3 hemos visto indicios de que esto no es así. En líneas generales podemos decir que la fiabilidad de la estimación MMSE dependerá de ciertos factores como el nivel de SNR local, la aleatoriedad del ruido, la precisión de los modelos de voz empleados y, para ciertas técnicas, el error cometido en la estimación del modelo de ruido o la máscara de segregación.

Dado que las estimas de voz obtenidas son aproximaciones, parece razonable modificar el algoritmo de descodificación integrado en el reconocedor para que éste considere

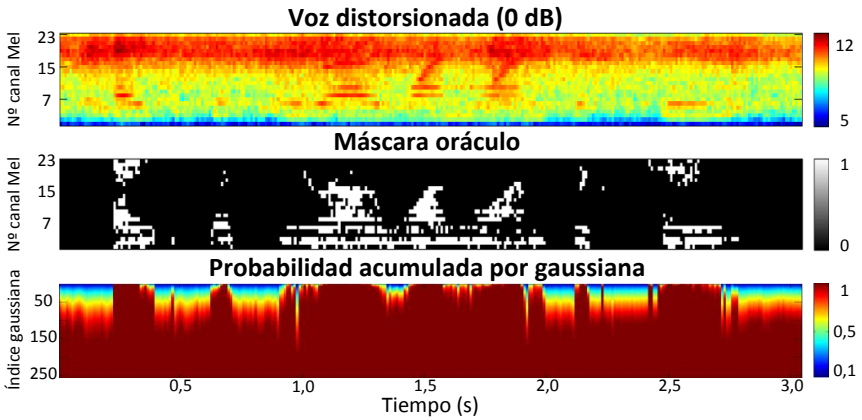

Figura 5.3: Relación entre la fiabilidad del espectro observado y la distribución a posteriori de las gaussianas del GMM.

la fiabilidad de cada estima. De esta forma, en lugar de tratar los valores estimados como deterministas, consideraremos que estos tienen asociada cierta función de probabilidad que modela la evidencia. En base a esta evidencia, el reconocedor otorgará un mayor peso en la etapa de descodificación a los vectores mejor estimados, mientras que aquellos escasamente fiables serán simplemente ignorados.

De la descripción anterior podemos deducir que el problema que pretendemos abordar consta de dos fases bien diferenciadas: (i) estimación de la incertidumbre (fiabilidad) de los valores estimados y (ii) propagación de la incertidumbre al reconocedor. Con respecto a la primera fase, en el siguiente apartado veremos que la propia estimación MMSE nos ofrece mecanismos para estimar de forma sencilla medidas de la incertidumbre de cada reconstrucción. Por otro lado, el uso de dichas medidas por parte del reconocedor requerirá modificar el cómputo de las probabilidades de observación. De entre el conjunto de técnicas propuestas en la literatura para este propósito, nosotros nos centraremos aquí en dos de ellas: descodificación *soft-data* y algoritmo ponderado de Viterbi. En los siguientes apartados estudiaremos con mayor profundidad estas dos propuestas.

5.2.1. Descodificación *soft-data*

Consideremos el esquema clásico de reconocimiento basado en HMMs entrenados con voz limpia donde cada estado s del modelo se representa mediante GMM de la

forma

$$p(\boldsymbol{x}_t|s) = \sum_{k=1}^{M} P(k|s)\mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_s^{(k)}, \boldsymbol{\Sigma}_s^{(k)}), \tag{5.23}$$

denotando q un estado dado del HMM.

En presencia de errores de estimación, los modelos anteriores no modelarán con propiedad las características compensadas, lo que producirá discrepancias que harán que la precisión del sistema de reconocimiento se resienta. Una estrategia propuesta por Morris en [203, 204] para reducir esta discrepancia consiste en reemplazar el cálculo de las probabilidades de observación $p(\boldsymbol{x}_t|s)$ por su valor esperado $\hat{p}(\boldsymbol{x}_t|s)$. Para calcular este valor esperado, supondremos que la técnica de compensación aplicada proporciona una variable aleatoria $\boldsymbol{x}_t \sim p(\boldsymbol{x}_t|\boldsymbol{y}_t, \Lambda_t)$ (Λ_t indica cualquier información a priori distinta de la observación $\boldsymbol{y}_t$) en lugar de un valor determinista $\hat{\boldsymbol{x}}_t$. La PDF de evidencia $p(\boldsymbol{x}_t|\boldsymbol{y}_t, \Lambda_t)$ que proporciona el algoritmo de compensación puede emplearse entonces para calcular el valor esperado $\hat{p}(\boldsymbol{x}_t|s)$ como sigue,

$$\hat{p}(\boldsymbol{x}_t|s) = \mathbb{E}[p(\boldsymbol{x}_t|s)|\boldsymbol{y}_t, \Lambda_t] = \int p(\boldsymbol{x}_t|s)p(\boldsymbol{x}_t|\boldsymbol{y}_t, \Lambda_t)d\boldsymbol{x}_t. \tag{5.24}$$

La ecuación anterior nos permite jugar con diferentes distribuciones de evidencia para los valores estimados. En el caso de que $\boldsymbol{x}_t$ sea determinista, su PDF de evidencia degenerará a una delta de Dirac centrada en la estima $\hat{\boldsymbol{x}}_t$, $p(\boldsymbol{x}_t|\boldsymbol{y}_t, \Lambda_t) = \delta_{\hat{\boldsymbol{x}}_t}(\boldsymbol{x}_t)$, y la ecuación (5.24) será

$$\hat{p}(\boldsymbol{x}_t|s) = \int p(\boldsymbol{x}_t|s)\delta_{\hat{\boldsymbol{x}}_t}(\boldsymbol{x}_t)d\boldsymbol{x}_t = p(\hat{\boldsymbol{x}}_t|s), \tag{5.25}$$

lo que equivale a calcular las probabilidades de observación con el GMM de partida.

En este trabajo consideraremos que la evidencia de la observación viene dada por una distribución normal multivariante, $p(\boldsymbol{x}_t|\boldsymbol{y}_t, \Lambda_t) = \mathcal{N}(\boldsymbol{x}_t; \hat{\boldsymbol{x}}_t, \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_t})$, siendo $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_t}$ la matriz de covarianza asociada con la estimación $\hat{\boldsymbol{x}}_t$. En este caso, es posible probar (ver p.ej. [203]) que la regla de descodificación con datos inciertos se traduce en

$$\begin{aligned}
\hat{p}(\boldsymbol{x}_t|s) &= \int p(\boldsymbol{x}_t|s)p(\boldsymbol{x}_t|\boldsymbol{y}_t, \Lambda_t)d\boldsymbol{x}_t \\
&= \sum_{k=1}^{M} P(k|s) \int \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_s^{(k)}, \boldsymbol{\Sigma}_s^{(k)})\mathcal{N}(\boldsymbol{x}_t; \hat{\boldsymbol{x}}_t, \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_t})d\boldsymbol{x}_t \\
&= \sum_{k=1}^{M} P(k|s)\mathcal{N}\left(\hat{\boldsymbol{x}}_t; \boldsymbol{\mu}_s^{(k)}, \boldsymbol{\Sigma}_s^{(k)} + \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_t}\right).
\end{aligned} \tag{5.26}$$

Esta regla es conocida en la literatura de diferentes formas: técnica *soft-data* [48, 122, 203, 204, 215], descodificación con incertidumbre de la observación [19, 177, 252] o,

simplemente, reconocimiento con incertidumbre de la estimación [35]. Podemos apreciar que el cómputo de las probabilidades de observación de $\hat{\boldsymbol{x}}_t$ es similar al caso determinista, salvo por la aparición de la varianza de la estima, $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_t}$, que incrementa la varianza de las gaussianas de la mezcla. Para valores de varianza pequeños, $|\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_t}| \to 0$, nos encontramos de nuevo con el caso determinista. Por otra parte, si el error de la estimación tiende a infinito, $|\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_t}| \to \infty$, la probabilidad de observación de $\hat{\boldsymbol{x}}_t$ será aproximadamente equiprobable. En dicho caso la descodificación vendrá guiada por las probabilidades de transición del HMM, despreciándose las observaciones.

Estudiada la forma en la que la etapa de reconocimiento se modifica para aceptar valores inciertos, la única tarea que nos resta es el cómputo de las matrices de covarianza de la estimación. En el caso improbable de que se disponga de las grabaciones de voz limpia asociadas a las ruidosas compensadas, podemos calcular una suerte de varianza oráculo de la siguiente forma:

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_t} = (\hat{\boldsymbol{x}}_t - \boldsymbol{x}_t)\,(\hat{\boldsymbol{x}}_t - \boldsymbol{x}_t)^\top\,, \tag{5.27}$$

siendo $\boldsymbol{x}_t$ el vector de características extraído de las grabaciones sin degradar.

Aunque la suposición en la que se fundamentan no sea realista, el uso de estas varianzas oráculo nos servirá para delimitar el techo superior del rendimiento del reconocedor cuando éste trabaja con datos inciertos. En la práctica, las matrices $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_t}$ tendrán que ser estimadas a partir de la propia señal distorsionada. Varias son las propuestas que podemos encontrar en la literatura en este sentido: desde la estimación de la incertidumbre basándose en la posición de los formantes [142], pasando por su cálculo mediante una función polinómica del nivel de SNR de la señal [19] o su deducción a partir del propio proceso de realce de características [35, 69, 72, 122, 126, 252]. En este trabajo seguiremos este último enfoque, siendo $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_t}$ la varianza de la estimación MMSE,

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_t} = \mathrm{Var}\left[\boldsymbol{x}_t \,|\boldsymbol{y}_t, \Lambda_t, \mathcal{M}_x\right] = \mathbb{E}\left[\boldsymbol{x}_t\boldsymbol{x}_t^\top \middle|\, \boldsymbol{y}_t, \Lambda_t, \mathcal{M}_x\right] - \hat{\boldsymbol{x}}_t\hat{\boldsymbol{x}}_t^\top. \tag{5.28}$$

Desarrollando convenientemente el término $\mathbb{E}[\boldsymbol{x}_t\boldsymbol{x}_t^\top|\boldsymbol{y}_t, \Lambda_t, \mathcal{M}_x]$ en base a las pro-

babilidades del modelo de voz tenemos que

$$\begin{aligned}\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_t} &= \left[\sum_{k=1}^{M} P(k|\boldsymbol{y}_t,\Lambda_t,\mathcal{M}_x)\underbrace{\int \boldsymbol{x}_t\boldsymbol{x}_t^\top p(\boldsymbol{x}_t|\boldsymbol{y}_t,\Lambda_t,\mathcal{M}_x)d\boldsymbol{x}_t}_{=\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}^{(k)}}+\hat{\boldsymbol{x}}^{(k)}\hat{\boldsymbol{x}}^{(k)\top}}\right] - \hat{\boldsymbol{x}}_t\hat{\boldsymbol{x}}_t^\top \\ &= \sum_{k=1}^{M} P(k|\boldsymbol{y}_t,\Lambda_t,\mathcal{M}_x)\left(\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}^{(k)}} + \hat{\boldsymbol{x}}^{(k)}\hat{\boldsymbol{x}}^{(k)\top}\right) - \hat{\boldsymbol{x}}_t\hat{\boldsymbol{x}}_t^\top \\ &= \sum_{k=1}^{M} P(k|\boldsymbol{y}_t,\Lambda_t,\mathcal{M}_x)\left[\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}^{(k)}} + \left(\hat{\boldsymbol{x}}^{(k)} - \hat{\boldsymbol{x}}_t\right)\left(\hat{\boldsymbol{x}}^{(k)} - \hat{\boldsymbol{x}}_t\right)^\top\right]. \end{aligned} \tag{5.29}$$

Los vectores $\hat{\boldsymbol{x}}^{(k)}$ de la ecuación anterior son las estimas parciales calculadas por el estimador MMSE para cada región (gaussiana o celda VQ) del modelo de voz. Por otra lado, $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}^{(k)}}$ es la varianza asociada a la reconstrucción $\hat{\boldsymbol{x}}^{(k)}$. Para calcular esta matriz nos apoyaremos en las suposiciones en las que se fundamenta cada técnica de compensación en particular. Así, en las técnicas del capítulo 3, $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}^{(k)}}$ se estimará usando las grabaciones estéreo disponibles. En cambio, en las técnicas de reconstrucción espectral descritas en el capítulo 4, $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}^{(k)}}$ la deduciremos del propio estimador MMSE. Para el caso del método de imputación TGI, los elementos de $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}^{(k)}}$ correspondientes a las componentes fiables del espectro tendrán una varianza nula, ya que estos valores son deterministas (asumimos que la máscara de segregación no contiene errores). Por otra parte, asumiendo independencia estadística entre las distintas características de voz, la varianza de los valores imputados se calcula recurriendo a la expresión dada en (B.14). En el caso de la técnica MMSR, el procedimiento es análogo al de la técnica TGI, obteniendo una expresión similar a la que aparece en (C.29).

En función del dominio en el que se compense la voz, la matriz de covarianza dada en (5.29) deberá transformarse o no para representar la varianza de la estimación en el dominio de los parámetros del reconocedor. Esto ocurre, por ejemplo, si el proceso de compensación inicial se efectúa en el dominio log-Mel (p.ej. en las técnicas TGI y MMSR) y el reconocedor trabaja con parámetros cepstrales. En este caso, la PDF de evidencia final para los vectores de características vendría dada por

$$\boldsymbol{x}_t^c \sim \mathcal{N}\left(\boldsymbol{C}\hat{\boldsymbol{x}}_t, \boldsymbol{C}\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_t}\boldsymbol{C}^\top\right), \tag{5.30}$$

siendo $\boldsymbol{C}$ el operador lineal que implementa la DCT y el superíndice c denota el dominio cepstral.

Debemos advertir que hasta ahora no se ha realizado distinción alguna entre el tipo de característica estimada: estática o dinámica. En las técnicas de compensación basadas en datos estéreo esta distinción es innecesaria, pues al no emplear ningún modelo

de distorsión de la voz la compensación se realiza en bloques, que contendrán tanto los parámetros estáticos como los dinámicos (velocidades y aceleraciones). Por contra, las técnicas de reconstrucción del capítulo 4 sólo estiman las características estáticas perdidas. A partir de valores reconstruidos se computarán las oportunas características dinámicas, sin embargo no disponemos de una estadística que nos permita estimar su varianza. En este caso recurriremos a la estrategia propuesta en [213], donde estas varianzas se calculan como combinación lineal de las varianzas obtenidas para los parámetros estáticos.

5.2.2. Algoritmo ponderado de Viterbi

Una estrategia alternativa para la propagación de la incertidumbre de la estimación $\hat{\boldsymbol{x}}_t$ al reconocedor de voz, consiste en emplear un factor exponencial de pesado $\rho_t \in [0, 1]$ que, aplicado sobre las probabilidades de observación $p(\hat{\boldsymbol{x}}_t|q)$, reduzca la contribución de los valores con mayor incertidumbre. Considerando este factor de pesado, la ecuación de actualización de las probabilidades por estado que emplea el algoritmo de Viterbi se modifica de la forma,

$$\phi_t(s_j) = \max_{s_i} \{\phi_{t-1}(s_i)a_{ij}\}\, p(\hat{\boldsymbol{x}}_t|s_j)^{\rho_t}, \tag{5.31}$$

donde s_i y s_j son dos estados del HMM, a_{ij} y $p(\hat{\boldsymbol{x}}_t|s_j)$ denotan a las probabilidades de transición y observación, respectivamente, y, por último, $\phi_t(s_j)$ proporciona la probabilidad del mejor camino de reconocimiento para el estado s_j en el instante de tiempo t. La única diferencia entre la descodificación usual por Viterbi y la de la ecuación anterior reside en el peso exponencial ρ_t, que sería $\rho_t = 1$ en el algoritmo de Viterbi normal.

A esta estrategia alternativa de reconocimiento con incertidumbre se le conoce con el nombre de algoritmo ponderado de Viterbi (WVA, *Weighted Viterbi Algorithm*) [48, 119, 122, 126, 278]. Además de su simplicidad (conlleva sólo una multiplicación en el dominio de las probabilidades logarítmicas), el algoritmo WVA ha demostrado ser una excelente técnica de reconocimiento con incertidumbre, superando incluso a otras técnicas más complejas como la presentada en el apartado anterior [48, 122]. Aunque existan diferentes extensiones a este algoritmo que consideran el uso de diferentes pesos $\rho_{t,i}$ $(i = 1, \ldots, D)$, uno para cada característica del vector estimado, aquí nos restringiremos al caso de partida, esto es, sólo consideraremos un factor de pesado ρ_t por cada vector $\hat{\boldsymbol{x}}_t$. La estimación y uso de diferentes pesos queda, pues, como trabajo futuro.

La mayor dificultad de WVA consiste en cómo determinar los factores de pesado ρ_t. Como ya comentamos en el apartado anterior, podemos considerar que la fiabilidad de una cierta estimación $\hat{\boldsymbol{x}}_t$ está íntimamente relacionada con la distribución de la

evidencia dada por $p(\boldsymbol{x}_t|\boldsymbol{y}_t, \Lambda_t)$. Si esta PDF se aproxima a una delta, entonces consideraremos que la estimación obtenida está libre de error y, por tanto, su fiabilidad será alta ($\rho_t \approx 1$). Si, en cambio, la PDF de evidencia tiende a una distribución uniforme, entonces el error esperado de $\hat{\boldsymbol{x}}_t$ será grande y su fiabilidad baja ($\rho_t \approx 0$). En este apartado proponemos dos criterios distintos para medir la incertidumbre asociada a la estimación: el error cuadrático medio (MSE, *Mean Squared Error*) de la estima y la entropía de la PDF de evidencia.

El error cuadrático medio de $\hat{\boldsymbol{x}}_t$ se define como la traza de la matriz de covarianza asociada a dicha estima. En la ecuación (5.29) obtuvimos una expresión analítica para el cálculo de dicha matriz. Por tanto, el MSE de $\hat{\boldsymbol{x}}_t$, al que notaremos por ϵ_t, viene dado por

$$\epsilon_t \equiv \text{MSE}(\hat{\boldsymbol{x}}_t) = \text{tr}(\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_t}). \tag{5.32}$$

El rango en el que se distribuye este error es $\epsilon_t \in [0, \infty)$. Para transformarlo al rango $[0, 1]$ y así obtener el factor de pesado ρ_t que buscamos, usaremos una función sigmoidal de la forma

$$\rho_t = 1 - \frac{1}{1 + e^{-a(\epsilon_t - b)}}, \tag{5.33}$$

siendo a la pendiente de la sigmoide y b su centro. Estos dos parámetros se determinan experimentalmente usando conjuntos de validación.

Como hemos mencionado anteriormente, la función de entropía nos proporciona una forma alternativa de medir la incertidumbre asociada a la estimación bhx_t. Para evitar el problema de medir la entropía asociada a una PDF, usaremos una versión discreta de la PDF de evidencia $p(\boldsymbol{x}_t|\boldsymbol{y}_t, \Lambda_t)$. De esta forma, evitaremos además el tener que decidir la forma paramétrica concreta para esta PDF, consiguiendo con ello un cómputo más simple y eficiente de la entropía requerida. En particular, la discretización que aplicaremos será la derivada de considerar las M regiones (asociadas a un diccionario VQ o a un GMM) del espacio de voz limpio. Así, utilizaremos la distribución $P(k|\boldsymbol{y}_t, \Lambda_t)$ (donde $k = 1, \ldots, M$ recorre la distintas regiones del espacio) para computar la siguiente entropía,

$$H(\hat{\boldsymbol{x}}_t) = -\sum_{k=1}^{M} P(k|\boldsymbol{y}_t, \Lambda_t) \log_2 P(k|\boldsymbol{y}_t, \Lambda_t) \tag{5.34}$$

Finalmente, el valor de fiabilidad ρ_t usado por el algoritmo WVA se calcula aplicando de nuevo una compresión sigmoidal sobre $H(\hat{\boldsymbol{x}}_t)$. Esta vez el rango de la entropía será $H(\hat{\boldsymbol{x}}_t) \in [0, H_{max}]$, siendo $H_{max} = \log_2(M)$ la entropía de la distribución uniforme.

5.3. Resumen

En este capítulo se ha investigado dos aspectos clave para las técnicas de compensación como son el uso de mejores modelos a priori de voz y el procesado de la incertidumbre derivada del proceso de estimación. En relación al primer punto, se ha estudiado la forma de modelar conjuntamente tanto la distribución en frecuencia como la evolución temporal de la voz, resultando este estudio en dos propuesta alternativas. La primera de ellas se basa en un modelado mediante modelos de mezclas de PDFs (p.ej. GMMs) de segmentos de voz de duración corta. Por otra parte, la segunda propuesta se ha basado en representar la voz mediante modelos ocultos de Markov. Ambas propuestas resultan en una mejora notable de la calidad del espectro de voz estimado, reduciendo el número de errores cometidos al procesar los vectores de características de forma independiente.

La segunda mitad del capítulo ha estado dedicada a la estimación de la fiabilidad de la reconstrucción y su posterior tratamiento por el reconocedor. Hemos visto que a partir de la distribución a posteriori de la voz podemos derivar medidas que nos indiquen la fiabilidad predicha del valor estimado. En particular, dos son las medidas en las que hemos centrado nuestra atención: la varianza del estimador MMSE y la entropía de la distribución a posteriori. Una vez calculadas estas medidas, el siguiente paso consiste en modificar el motor de reconocimiento para que pondere, en función de su fiabilidad, el impacto de cada observación en el proceso de descodificación. De nuevo, dos han sido los esquemas que hemos estudiado para esta tarea. El primero de ellos, denominado reconocimiento *soft-data*, usa la varianza de la estimación MMSE para incrementar las varianzas de los gaussianas que componen los estados del modelo acústico. La segunda estrategia, conocida como algoritmo WVA, emplea un factor exponencial de pesado al que se elevan las probabilidades de observación calculadas por el reconocedor. En resumen, ambas estrategias implementan un mecanismo efectivo para reducir la contribución al reconocimiento de aquellos valores estimados menos fiables.

CAPÍTULO 6

Evaluación

En este capítulo presentamos la evaluación experimental de las técnicas de compensación de características propuestas en los capítulos anteriores. Para ello se llevarán a cabo diversos experimentos de reconocimiento de voz en condiciones ruidosas (con ruido aditivo y convolutivo, principalmente) en los que el rendimiento de dichas técnicas se medirá en función de la tasa de palabras correctamente reconocidas. Asimismo, estos experimentos nos servirán como base para realizar un análisis comparativo entre las técnicas que aquí proponemos y otras técnicas similares que se encuentran en la literatura, destacando de forma razonada los pros y contras de cada una de ellas.

La organización de este capítulo es la siguiente. En la sección 6.1 describimos el marco experimental en el que nos hemos basado para desarrollar la evaluación experimental de las técnicas propuestas. Esto incluirá la descripción de las bases de datos empleadas, el tipo de parametrización de la voz usada y, por último, detalles relativos a los modelos acústicos y del lenguaje del reconocedor. Seguidamente la sección 6.2 expondrá los resultados obtenidos para los experimentos de reconocimiento. Finalmente, en la sección 6.3 se presentará un resumen de los resultados obtenidos.

6.1. Marco experimental

Con el fin de evaluar las técnicas propuestas, se hace necesario establecer un marco experimental que nos permita evaluarlas en una variedad de contextos. Estos deberían incluir, a ser posible, tareas de reconocimiento con distinta dificultad, como por ejemplo reconocimiento de palabras aisladas, tareas de medio vocabulario y tareas de gran

vocabulario o habla espontánea. Asimismo, dado que esta tesis versa sobre reconocimiento robusto, el marco experimental debería contar con un gran abanico de contextos acústicos diferentes, intentando que estos reflejen el uso cotidiano que se espera del sistema de reconocimiento. Aunque no lo abordamos aquí directamente, también debería quedar reflejada la variedad acústica en lo que respecta al locutor. En esta tesis hemos intentando, en la medida de lo posible, cubrir una gran variedad de usos posibles de los sistemas de reconocimiento.

En los siguientes apartados se describen las bases de datos empleadas durante la evaluación, así como la representación de la voz y los detalles sobre el *back-end* usados en cada caso.

6.1.1. Bases de datos

La evaluación de las técnicas que se proponen en esta tesis se ha realizado usando dos bases de datos bien conocidas en la literatura: Aurora2 [141] y Aurora4 [140]. Ambas contienen elocuciones en inglés con distintos tipos de ruido añadidos artificialmente. No obstante, mientras que Aurora2 define una tarea de reconocimiento de secuencias de dígitos, Aurora4 contiene frases del periódico Wall Street Journal. Las dos bases de datos fueron liberadas inicialmente por el grupo de trabajo STQ AURORA del Instituto Europeo de Estándares de Telecomunicaciones (ETSI, *European Telecommunications Standards Institute*) con objeto de evaluar los estándares DSR de reconocimiento distribuido de voz. Desde entonces estas bases de datos se han convertido en estándar *de facto* al ser empleadas de manera amplia por un gran número de grupos de investigación alrededor del mundo. En este sentido, ambas bases de datos se constituyen como un marco de comparación inigualable para evaluar los progresos de la ciencia, permitiendo a los distintos grupos comparar sus propuestas con el estado del arte bajo un entorno de trabajo unificado.

Además de las dos bases de datos anteriores, para evaluar ciertas técnicas de esta tesis en un mayor rango de condiciones ruidosas, hemos definido un nueva base de datos a la que hemos denominado como Aurora2 ampliada. Esta nueva base de datos, como veremos, se obtiene tras degradar artificialmente las grabaciones de voz limpia de Aurora2 con distintos ruidos adquiridos por nuestro grupo de investigación de la Universidad de Granada.

6.1.1.1. Aurora2

Aurora2 define una tarea consistente en el reconocimiento de voz continua de cadenas de hasta 7 dígitos en inglés pronunciados por adultos de EE.UU. El vocabulario

empleado consta de 11 palabras, una para cada dígito excepto el cero, que admite dos pronunciaciones ("zero" y "oh"). Esta base de datos es una versión revisada de la base de datos TIDigits [169] después de aplicar las siguientes operaciones:

- Submuestreo. La frecuencia de muestreo inicial de 20 KHz con la que fueron adquiridas originalmente las grabaciones de TIDigits, fue diezmada hasta los 8 KHz a fin de reflejar la frecuencia con la que normalmente se trabaja en comunicaciones digitales por voz.

- Filtrado. Sobre la señal muestreada a 8 KHz se aplica un filtrado lineal que simula las características en frecuencia de los terminales y equipos en el área de comunicaciones. Este filtrado adicional se realizó con el estándar G.712 propuesto por la ITU [7].

- Adición de ruido acústico. A fin de evaluar las prestaciones del reconocimiento de voz en condiciones de ruido, además de la referencia limpia, Aurora2 incluya versiones de las elocuciones contaminadas artificialmente con 8 tipos de ruido acústico y distintos niveles de SNR (20, 15, 10, 5, 0 y -5 dB).

Las señales obtenidas en la etapa anterior se dividen en dos conjuntos: uno de entrenamiento y otro de evaluación. Para la fase de entrenamiento se disponen de 8440 frases grabadas en condiciones de alta SNR por un total de 55 locutores masculinos y 55 femeninos, todos ellos adultos. Este grupo de frases define el conjunto de entrenamiento limpio de Aurora2. Además del conjunto limpio, se define un segundo conjunto de entrenamiento denominado multiestilo o multicondición donde las 8440 frases del conjunto limpio son divididas en 20 subconjuntos con 422 frases cada uno. Las frases de cada subconjunto se contaminan con un tipo de ruido aditivo de entre cuatro posibles (ruido del metro, gente hablando, motor de un coche y sala de exposiciones) y una SNR de entre cinco valores posibles (20 dB, 15 dB, 10 dB, 5 dB o señal limpia).

El conjunto de evaluación de Aurora2 se compone de 4004 frases pronunciadas por 52 hombres y 52 mujeres. Estas frases se dividen en 4 grupos con 1001 frases cada uno, lo que supone un total de 32.883 dígitos para reconocer por grupo. Las frases de cada grupo se contaminan ahora con distintos tipos de ruido y valores de SNR que reflejan entornos acústicos potenciales para las aplicaciones de reconocimiento automático del habla. Los tipos de ruido acústico empleados son ruido del metro (*subway*), gente hablando (*babble*), motor de un coche (*car*), sala de exposiciones (*exhibition*), restaurante (*restaurant*), ruido de calle (*street*), aeropuerto (*airport*) y ruido adquirido en una estación de trenes (*train station*). En cuanto a los niveles de SNR empleados estos son 20,

15, 10, 5, 0 y -5 dB, junto con una referencia limpia donde la señal no se distorsiona. Usando estos ruidos y SNRs, se definen 3 subconjuntos de evaluación: A, B y C.

En el subconjunto A (*set A*) cada uno de los 4 grupos de 1001 frases de *test* se contamina con un ruido de los utilizados en las secuencias de entrenamiento multicondición. Los valores de SNR empleados son los 7 comentados anteriormente. En total el subconjunto A contiene 28 grupos de 1001 frases cada uno. El subconjunto B (*set B*) se obtiene de forma similar, pero esta vez se emplean 4 ruidos no considerados en el entrenamiento multicondición (restaurante, calle, aeropuerto y estación de tren). Por último, en el subconjunto de evaluación C (*set C*) sólo se consideran 2 grupos de 1001 frases de los 4 posibles. Para contaminar estas frases se emplean 2 ruidos (metro y calle) y los 7 niveles de SNR comentados antes. Además del ruido aditivo, en este subconjunto de *test* se considera un filtrado convolutivo de característica MIRS que emula el comportamiento de los terminales de telecomunicaciones GSM. Este subconjunto tendrá pues los efectos del ruido aditivo y del ruido convolucional juntos.

6.1.1.2. Aurora2 ampliada

Uno de los problemas que presenta Aurora2 de cara a la evaluación de las técnicas de reconocimiento robusto de voz, es la reducida variabilidad de los ruidos que contiene, siendo estos además bastante similares entre sí. Además de su reducida variabilidad, otro problema con el que nos encontramos es la corta duración de algunos de los ruidos incluidos en esta base de datos[1]. Estas limitaciones pueden suponer una traba a la hora de evaluar algunas de las técnicas de compensación propuestas en esta tesis. En particular, las técnicas basadas en grabaciones estéreo requieren un gran volumen de datos de entrenamiento para estimar de forma robusta las transformaciones que aplican. A fin de evaluar estas técnicas en una mayor variabilidad de entornos acústicos que los que encontramos en Aurora2, en el transcurso de esta tesis se diseñó una nueva base de datos degradando las grabaciones limpias de Aurora2 con distintos ruidos reales grabados por nuestro grupo de investigación de la Universidad de Granada. En los siguientes párrafos describimos en detalle la forma en la que se diseñó esta nueva base de datos.

Para los propósitos de este trabajo se seleccionaron los 8440 ficheros del conjunto entrenamiento limpio de Aurora2 y los 4004 ficheros sin distorsionar del conjunto de evaluación *set* A. La simulación de diferentes ambientes acústicos se realizó añadiendo artificialmente distintos tipos de ruido a los conjuntos de ficheros de voz anteriores.

[1]La duración media de los ruidos incluidos en Aurora2 es de 125 segundos, siendo el ruido de sala de exposiciones el más corto con una duración de sólo 19 segundos y el más largo el ruido de restaurante con una duración de 286 segundos

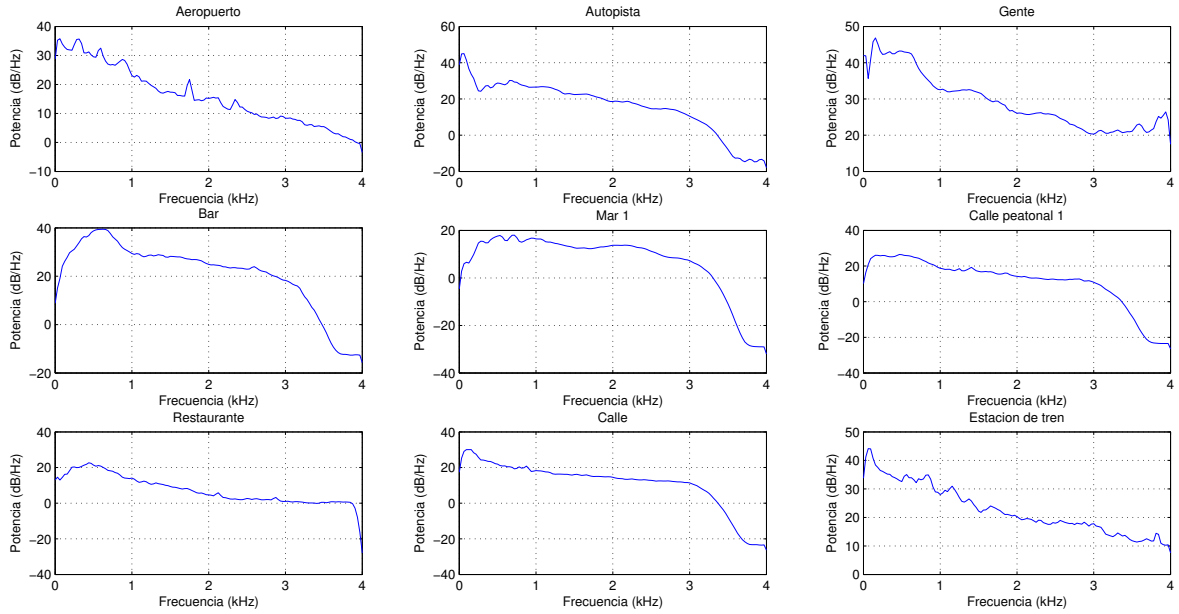

Figura 6.1: Densidad de potencia espectral de los ruidos usados en el entrenamiento y en el conjunto de evaluación A de la base de datos Aurora2 ampliada.

El conjunto de ruidos acústicos seleccionados consta de nueve ambientes: aeropuerto (grabación de una terminal de un aeropuerto), autopista (ruido de coches en un tramo de una autopista), gente (grabación de gente hablando), bar (grabación de un bar), mar 1 (sonido de una playa en un día de pocas olas), calle peatonal 1 (calle peatonal muy transitada), restaurante (gente hablando en un restaurante), calle (calle con tráfico) y estación de tren (ambiente de una estación de tren). La densidad de potencia espectral de estos ruidos se muestra en la figura 6.1. Cada una de las grabaciones, cuya longitud va desde los 180 segundos a los 530 segundos (longitud media de 282 segundos), se dividió en dos partes: dos tercios de su longitud se emplearon para entrenamiento y el tercio restante se reservó para contaminar los ficheros de voz de evaluación.

Usando los ficheros de voz del conjunto de entrenamiento limpio de Aurora2 y los ruidos reservados para entrenamiento, se definieron 55 conjuntos de entrenamiento distintos cada uno con 8440 ficheros. El primer conjunto contiene los ficheros de voz tal cual sin distorsionar. El resto de los 54 conjuntos se generaron contaminando artificialmente los 8440 ficheros con cada uno de los 9 ruidos a una SNR determinada de entre 6 valores posibles (20, 15, 10, 5, 0 y -5 dB). La adición de ruido a los ficheros de voz se realizó siguiendo la recomendación ITU P.56 [5] con el software asociado.

A partir de los 4004 ficheros de voz limpia del *set* A de Aurora2 se definieron 2 conjuntos de evaluación. El primer conjunto, al que nos referiremos también como *set*

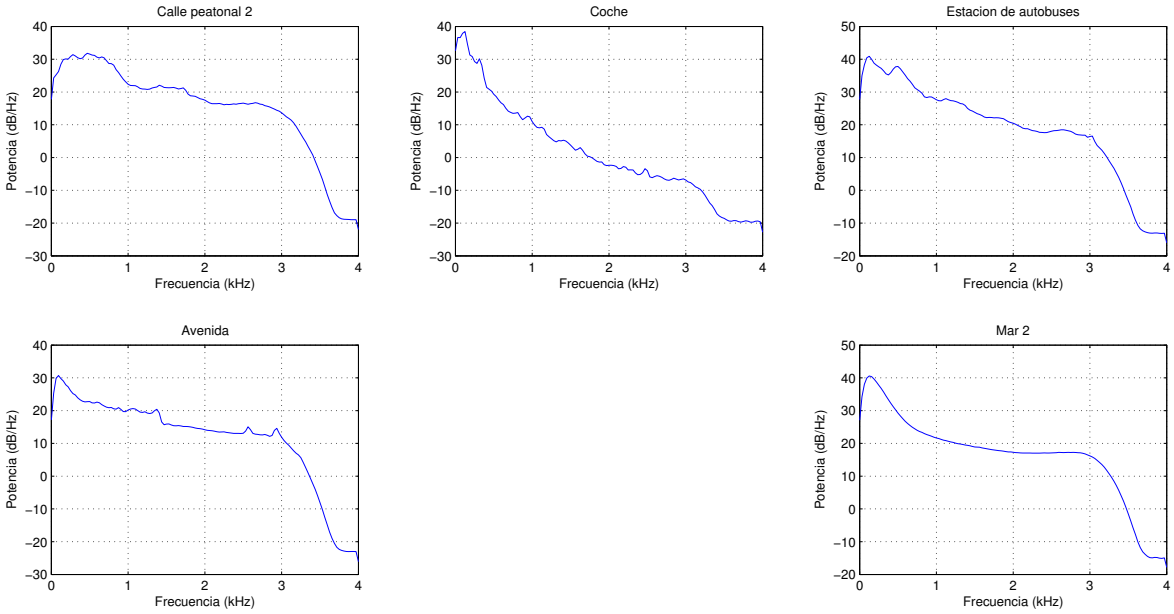

Figura 6.2: Densidad de potencia espectral de los ruidos usados en el conjunto de evaluación B de la base de datos Aurora2 ampliada.

A, se construye contaminando los ficheros de voz con los ruidos y SNRs consideradas en el entrenamiento, disponiendo de 55 condiciones de evaluación con 4004 ficheros cada una. El objetivo de este conjunto evaluación es el estudio del rendimiento de las técnicas propuestas en condiciones acústicas similares a las consideradas durante el entrenamiento. El segundo conjunto de evaluación, que notaremos como *set* B, se crea de la misma forma que el conjunto anterior, pero considerando diferentes ruidos y valores de SNRs que los usados en el entrenamiento. Los tipos de ruido utilizados para este segundo conjunto de *test* son cuatro: calle peatonal 2 (grabación de una calle peatonal), coche (grabación en el interior de un vehículo circulando por una autovía), estación de autobuses (interior de una estación de autobuses), avenida (avenida con mucho tráfico y gente andando) y mar 2 (grabación en una playa en un día de mar embravecida). La figura 6.2 muestra la densidad de potencia espectral de estos ruidos. Por otra parte, se han considerado 5 valores de SNR distintos a los empleados anteriormente (17,5, 12,5, 7,5, 2,5 y -2,5 dB), lo que hace un total de 25 condiciones acústicas diferentes. El objetivo de este segundo conjunto de evaluación es el estudio del comportamiento de las técnicas propuestas en condiciones acústicas no abordadas durante el entrenamiento.

6.1.1.3. Aurora4

Aurora4 define una tarea de reconocimiento de mediano/gran vocabulario (5000 palabras) consistente en frases dictadas en inglés del periódico Wall Street Journal. Al igual que Aurora2, en esta base de datos se definen dos conjuntos de entrenamiento, limpio y multicondición, y varios conjuntos de evaluación. El conjunto de datos de entrenamiento limpio está formado por 7138 frases (14 horas de duración en total) de 83 locutores grabadas con un micrófono de proximidad y sin ruido ambiental. Existe también un conjunto de datos de entrenamiento multicondición que se crea dividiendo las 7318 frases en dos subconjuntos. Las frases del primer subconjunto se graban con el micrófono de proximidad, mientras que el resto de ellas son adquiridas con un micrófono de peor calidad. Estos subconjuntos contienen, a su vez, ruido acústico añadido de forma artificial de entre los presentes en las frases de *test* a una SNR que oscila entre los 10 dB y 20 dB (elegida de forma aleatoria).

Los conjuntos de evaluación incluidos en Aurora4 se generan a partir de un conjunto inicial de 330 frases (aproximadamente 40 minutos de duración en total correspondientes a 5353 palabras) muestreadas a 16 KHz y filtradas según la característica P.341 de la ITU [6]. De estas 330 frases se tienen, a su vez, distintas versiones grabadas bien por un micrófono Sennheiser de proximidad, bien por otros 18 micrófonos de peor calidad. A partir de las grabaciones hechas con el micrófono Sennheiser se generan 6 subconjuntos de evaluación obtenidos tras sumar de manera artificial los siguientes tipos de ruido: motor de un coche (*set* T-02), gente hablando (*set* T-03), restaurante (*set* T-04), ruido de calle (*set* T-05), aeropuerto (*set* T-06) y ruido de un tren (*set* T-07). La SNR elegida para degradar las frases se escoge aleatoriamente entre 5 dB y 15 dB a incrementos de 1 dB. Además de los 6 subconjuntos anteriores, se dispone también de un séptimo conjunto con las 330 frases originales sin distorsionar (*set* T-01).

A fin de tener en cuenta posibles distorsiones debidas al canal, a los 7 conjuntos de evaluación anteriores hay que sumar otros 7 más obtenidos de la misma forma, pero esta vez usando las grabaciones realizadas con los 18 micrófonos restantes (*sets* T-08 a T-14) .

6.1.2. Representación de la voz

El método de extracción de características de la voz que hemos empleado en la evaluación experimental es el propuesto por el Instituto Europeo de Telecomunicaciones en su estándar ETSI ES 201 108 DSR front-end (ETSI FE) [1]. El estándar ETSI FE implementa un análisis cepstral básico en donde no se incluye ningún módulo específico para robustecer el sistema de reconocimiento frente a degradaciones debidas al ruido.

El esquema seguido por este estándar para calcular las características de voz empleadas por el sistema de reconocimiento es el siguiente. En primer lugar, se muestrea la señal de voz a 8 KHz y sobre ésta se aplica una compensación de la componente continua además de un filtrado de preénfasis con un factor $\mu = 0{,}97$. Posteriormente, la señal es segmentada en tramas de 25 ms (200 muestras) desplazadas cada 10 ms usando una ventana de Hamming, la cual supone un buen compromiso entre la resolución espectral y el rizado lateral. Tras aplicar la ventana de Hamming, al vector se le añaden ceros hasta completar los 256 valores, sobre los cuales se aplica una FFT de 256 puntos para calcular la magnitud del espectro de la señal.

Ya en el dominio espectral, la señal pasa por un banco de 23 filtros triangulares distribuidos uniformemente en escala Mel desde 64 Hz hasta 4 KHz. Usando logaritmos naturales y mediante una transformada discreta del coseno se obtienen 13 coeficientes cepstrales en escala Mel (MFCC, *Mel Frequency Cepstral Coefficients*). Por tanto, las características estáticas de la voz son los 13 primeros coeficientes cepstrales $(C0, C1, \ldots, C12)$. Finalmente, el vector de características se amplia con los coeficientes dinámicos correspondientes (velocidades y aceleraciones), alcanzando una dimensión final de 39 elementos.

Adicionalmente, en algunos experimentos se utilizará el método avanzado de extracción de características ETSI ES 202 050 DSR advanced front-end (ETSI AFE) [2]. La principal diferencia entre este *front-end* avanzado y el básico, ETSI FE, es la inclusión de métodos de procesamiento robusto que hacen que los parámetros extraídos sean más inmunes al ruido acústico. En concreto, las principales técnicas de procesamiento robusto que se incluyen son: doble filtrado de Wiener para reducción de ruido acústico, reducción dinámica del ruido en función de la SNR de la señal (procesado de la forma de onda) y, por último, ecualización ciega de los parámetros cepstrales extraídos mediante la técnica BEQ (*Blind EQualization*, ecualización ciega).

6.1.3. Reconocedor de voz

Los reconocedores de voz empleados en la evaluación experimental se basan en modelos ocultos de Markov de habla continua entrenados sólo con voz limpia (sin distorsionar). Para su construcción se ha empleado los *scripts* incluidos en las respectivas bases de datos y el paquete de software HTK [279] en su versión 3.4. En los siguientes apartados se describen los detalles del reconocedor empleado en cada base de datos.

6.1.3.1. Aurora2

En Aurora2 se dispone de un HMM para cada una de las once palabras[1] que conforman el vocabulario de esta base de datos. Cada HMM cuenta con los siguientes parámetros:

- Se establecen 16 estados por palabra (18 estados si se consideran los dos nodos de enlace al principio y final de cada HMM).
- Se usa una topología de Bakis (topología de izquierda a derecha con un salto máximo permitido de un estado).
- Cada estado se modela mediante una mezcla de 3 gaussianas multivariantes (de dimensión 39) con matriz de covarianza diagonal.

Adicionalmente, se definen dos modelos de pausa, uno para el comienzo y final de la frase y otro para los silencios entre las palabras. El primero consta de 3 estados con una mezcla de 6 gaussianas por estado. El segundo consiste simplemente en un estado ligado (compartiendo sus parámetros) al estado intermedio del modelo de silencio para comienzo y fin de frase. El entrenamiento de los HMMs se realiza aplicando el algoritmo de reestimación de Baum-Welch en varias iteraciones.

6.1.3.2. Aurora4

El modelado acústico en esta base de datos se basa en HMMs dependientes del contexto (trifonemas de palabras cruzadas) entrenados sobre una base inicial de 45 fonemas. Cada modelo cuenta con tres estados y 8 gaussianas por estado. La única excepción ocurre en los modelos de silencio donde se emplean 16 gaussianas por estado. Para reducir el número de parámetros y entrenar de manera robusta los diferentes modelos, se emplean árboles de regresión para agrupar los estados que son estadísticamente similares entre sí. El modelo del lenguaje consiste en una bigramática con 5000 palabras.

6.2. Evaluación de las técnicas propuestas

A lo largo de esta sección evaluaremos las técnicas propuestas en esta tesis para mejorar el reconocimiento automático de voz en condiciones de ruido acústico. La

[1]El vocabulario consiste en los diez dígitos en inglés, teniendo el 0 dos pronunciaciones asociadas: 'zero' y 'oh'.

evaluación experimental consistirá en distintos experimentos de reconocimiento de voz sobre las bases de datos Aurora2 y Aurora4, principalmente. A fin de decidir qué técnicas de las propuestas proporcionan mejores resultados y si estos son estadísticamente significativos, en el siguiente apartado comenzaremos presentando las medidas de rendimiento adoptadas y los procedimientos de contraste de hipótesis oportunos.

6.2.1. Criterios de evaluación

Existen varias medidas complementarias que nos permiten evaluar el rendimiento de un sistema de reconocimiento automático del habla. La primera de ellas es el porcentaje de palabras correctamente reconocidas: cuanto mayor sea este porcentaje, mayor será el rendimiento del sistema en cuestión. Una segunda medida del rendimiento del sistema que podemos considerar es la complejidad computacional en tiempo y/o memoria del mismo. Así, un sistema que proporcione buenas tasas de reconocimiento de palabras, pero cuya complejidad sea muy elevada, difícilmente podrá ser llevado a la práctica o aplicado a tareas que demanden requisitos de tiempo real. Aunque ambas medidas son importantes de cara a la implantación de un sistema de reconocimiento real, en este capítulo nos centraremos fundamentalmente en la primera de ellas, esto es, mediremos el rendimiento del sistema en función del número de errores cometidos por el mismo. En cuanto a la complejidad computacional, en los capítulos anteriores pueden encontrarse algunos análisis y comparativas que se consideraron relevantes.

6.2.1.1. Precisión del reconocedor

La precisión del reconocedor se evalúa por medio de pruebas de reconocimiento en las que se cuenta el número de errores producidos. El cociente entre el número de errores y el número de elementos reconocidos constituye la tasa de error y representa la probabilidad de cometer errores de reconocimiento. Sea n_w el número de palabras totales en la transcripción de referencia asociada a una frase de *test* dada. Asimismo notemos como n_i, n_b y n_s al número de errores de inserción (palabras adicionales insertadas en la frase reconocida), borrado (palabras presentes en la transcripción de referencia pero no en la obtenida por el sistema de reconocimiento) y sustitución (palabras reconocidas como otras palabras en la frase reconocida), respectivamente, cometidos por el reconocedor al transcribir dicha frase de test. Estos errores se contabilizan usando un alineamiento basado en un algoritmo de programación dinámica entre las transcripción de referencia y la obtenida por el sistema de reconocimiento. En base a estas variables,

se define la tasa de error de palabra (WER, *Word Error Rate*) como

$$\mathrm{WER} = \frac{n_i + n_b + n_s}{n_w} \tag{6.1}$$

Alternativamente, podemos expresar el rendimiento del reconocedor en términos de la tasa de acierto o precisión del reconocimiento de palabra (WAcc, *Word Accuracy*),

$$\mathrm{WAcc} = 1 - \mathrm{WER} \tag{6.2}$$

Tanto el WER como el WAcc se suelen representar en porcentajes, pudiéndose obtener tasas de error superiores al 100 % o, equivalentemente, WAccs negativos, debido a las inserciones.

6.2.1.2. Medidas de confianza

A la hora de comparar dos sistemas de reconocimiento, A y B, en función de sus tasas de reconocimiento de palabras, nos debemos preguntar si son o no estadísticamente significativas las diferencias entre ambos sistemas. Dicho con otras palabras, si la tasa de acierto obtenida para el sistema A es mayor que la del sistema B, nos preguntamos si A es mejor que B. Para responder a esta pregunta, definimos un intervalo de confianza alrededor de cada tasa de reconocimiento de forma que podamos afirmar cómo de fiables son las conclusiones o hipótesis que establecemos. Para calcular la amplitud de estos intervalos, en esta tesis usaremos el *test z* [117]. Suponiendo una confianza $1-\alpha$ respecto a nuestra hipótesis (p.ej. el 95 %, $\alpha = 0,05$), la amplitud Δ del intervalo de confianza que con probabilidad $1-\alpha$ contendrá la tasa de palabras obtenida es

$$\Delta = z_{1-\frac{\alpha}{2}} \sqrt{\frac{\mathrm{WAcc}(1-\mathrm{WAcc})}{n}} \tag{6.3}$$

donde n es el número total de ensayos (número de palabras) y $z_{1-\alpha/2}$ es el cuantil $1-\alpha/2$ de la distribución normal estándar. Por ejemplo para $\alpha = 0,05$, es decir, si deseamos obtener un nivel de confianza del 95 %, $z_{1-\alpha/2} = 1,96$.

En la figura 6.3 se muestra la amplitud Δ de los intervalos de confianza al 95 % obtenidos para las bases de datos Aurora2 y Aurora4. Para calcular estos intervalos hemos de tener en cuenta que el número de palabras en las transcripciones de referencia de los conjuntos de evaluación con voz limpia es de $n = 13159$ para Aurora2 y de $n = 10653$ para Aurora4. Atendiendo a estos resultados podemos decir que, en el caso de Aurora2, mejoras en el incremento del WAcc superiores a 0,78, 0,68 y 0,51 % serán significativas para tasas de reconocimiento en torno a 70 %, 80 % y 90 %, respectivamente. Para Aurora4, como se aprecia, las mejoras deben ser levemente mayores

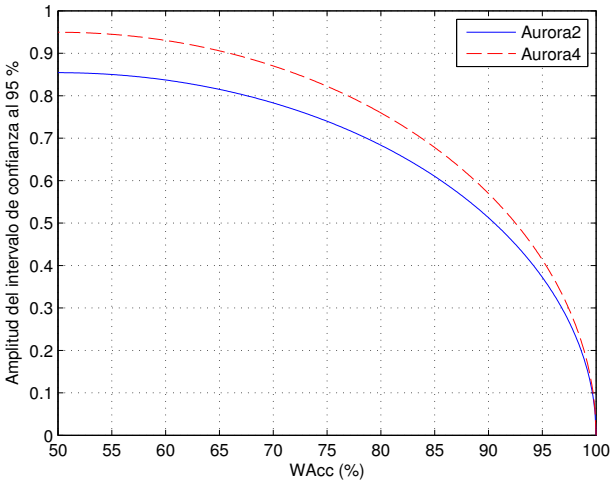

Figura 6.3: Amplitud de los intervalos de confianza al 95 % en función de la tasa de reconocimiento de palabras, WAcc (%), para las pruebas de reconocimiento sobre los conjuntos de voz limpia de las bases de datos Aurora2 y Aurora4.

para ser consideradas significativas. En el resto de este capítulo capítulo, con el fin de no sobrecargar los gráficos y tablas de resultados obtenidos, se omitirán los intervalos de confianza asociados con las medidas obtenidas, recomendándose la consulta de la figure 6.3 para más detalles.

6.2.2. Resultados de referencia

En este apartado presentamos los resultados de reconocimiento en términos de tasa de palabras correctamente reconocidas, WAcc (%), obtenidos para las bases de datos Aurora2 y Aurora4 cuando se emplean modelos acústicos (HMMs) entrenados con voz sin distorsionar y distintas representaciones de la voz. En primer lugar, evaluaremos la robustez al ruido del *front-end* básico de la ETSI, ETSI FE [1], en combinación con varias de las técnicas de normalización de características estudiadas en la sección 2.2.3.1: (i) coeficientes MFCC sin compensar, (ii) MFCCs normalizados en media (CMN) y (iii) MFCCs normalizados por la técnica de ecualización de histogramas, HEQ, propuesta en [238]. Asimismo, también presentaremos los resultados de reconocimiento obtenidos por el *front-end* avanzado de la ETSI, ETSI AFE [2], en estas bases de datos. Este método de extracción de características, como se comentó anteriormente, incluye ciertos módulos de procesamiento especialmente diseñados para incrementar la robustez del sistema de reconocimiento frente al ruido.

En la tabla 6.1 se presentan los resultados de referencia obtenidos para Aurora2

	SET A				SET B				SET C		
SNR	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station	Subway	Street	**Media**
Limpio	99,11	99,09	98,96	99,23	99,11	99,09	98,96	99,23	99,32	99,00	**99,11**
20 dB	96,90	91,81	96,57	96,39	92,08	96,19	93,86	94,23	93,92	95,41	**94,74**
15 dB	92,08	75,54	84,76	89,36	77,62	87,67	80,73	81,86	87,32	89,75	**84,67**
10 dB	73,87	48,76	57,35	68,44	53,98	64,54	54,58	54,92	72,61	74,43	**62,35**
5 dB	45,23	22,16	23,38	35,11	26,68	34,13	27,38	25,08	46,33	49,03	**33,45**
0 dB	20,94	9,52	8,65	11,05	10,29	15,75	12,47	9,56	19,71	23,73	**14,17**
-5 dB	10,72	5,80	6,53	6,60	6,51	8,40	7,28	7,53	10,10	11,06	**8,05**
Media	65,80	49,56	54,14	60,07	52,13	59,66	53,80	53,13	63,98	66,47	**57,87**

Tabla 6.1: Resultados de reconocimiento, WAcc (%), obtenidos para los tres conjuntos de evaluación de Aurora2 usando modelos entrenados con voz limpia y características extraídas por el ETSI FE.

cuando se emplea la parametrización básica extraída por el ETSI FE (MFCCs sin compensar). Los resultados de reconocimiento se han detallado para cada uno de los conjuntos de evaluación de esta base de datos (*sets* A, B y C), cada tipo de ruido y cada nivel de SNR ("Limpio" indica reconocimiento con voz sin distorsionar). Asimismo, se incluyen las medias parciales calculadas promediando las tasas de reconocimiento de 0 a 20 dB por tipo de ruido (última fila de la tabla) y las medias globales por SNR (última columna en negrita de la tabla). Por último, en la celda de la esquina inferior derecha, se presenta la media global de reconocimiento obtenida promediando las medias parciales de 0 dB a 20 dB, como es costumbre hacerlo en esta base de datos.

Como era de esperar, el rendimiento del sistema de reconocimiento se ve mermado cuando éste trabaja sobre voz distorsionada, siendo esta degradación más evidente a bajas SNRs (5 dB, 0 dB y -5 dB). Incluso en una tarea tan simple como la que define Aurora2, el rendimiento a estas SNRs decae a niveles en los que el sistema de reconocimiento es ineficaz.

Las tablas 6.2 y 6.3 muestran los resultados de reconocimiento obtenidos cuando las características de voz se normalizan mediante CMN y HEQ, respectivamente. Como podemos apreciar, la mejora del rendimiento obtenida tras aplicar estas técnicas básicas es notable, especialmente a niveles de SNR intermedios (p.ej. a 5 dB la mejora que se produce por usar CMN es de aproximadamente 10 puntos y de 43 por el uso de HEQ). También vemos que la mejora introducida por HEQ es mucho mayor que la proporciona CMN. La razón de este comportamiento se estudió en la sección 2.2.3.1: HEQ normaliza todos los momentos de la distribución de voz ruidosa, mientras que CMN solo normaliza la media de esta distribución.

Los resultados de reconocimiento obtenidos por el estándar ETSI AFE se presentan

	SET A				SET B				SET C		
SNR	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station	Subway	Street	**Media**
Limpio	99,14	99,09	98,99	99,17	99,14	99,09	98,99	99,17	99,17	99,12	**99,11**
20 dB	96,22	97,64	97,70	96,42	98,10	97,01	98,03	98,06	96,53	97,16	**97,29**
15 dB	90,67	93,83	92,01	90,34	95,18	92,17	94,72	93,83	90,82	91,93	**92,55**
10 dB	71,23	79,47	68,74	69,27	83,60	74,24	84,58	78,28	71,91	74,24	**75,56**
5 dB	38,19	47,22	32,84	34,80	54,44	42,26	52,94	44,80	38,07	42,68	**42,82**
0 dB	21,40	23,34	19,95	18,45	26,19	22,52	27,97	23,14	21,89	22,07	**22,69**
-5 dB	13,82	12,48	12,38	10,18	13,23	12,15	15,39	13,88	13,79	11,96	**12,92**
Media	63,54	68,30	62,25	61,86	71,50	65,64	71,65	67,62	63,84	65,62	**66,18**

Tabla 6.2: Resultados de reconocimiento, WAcc (%), obtenidos para los tres conjuntos de evaluación de Aurora2 usando modelos entrenados con voz limpia y características extraídas por el ETSI FE normalizadas en media (CMN).

	SET A				SET B				SET C		
SNR	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station	Subway	Street	**Media**
Limpio	98,96	99,06	98,96	99,23	98,96	99,06	98,96	99,23	98,89	98,94	**99,03**
20 dB	96,99	98,22	98,18	97,25	97,70	97,82	98,30	97,87	97,18	97,70	**97,72**
15 dB	93,98	96,13	96,63	94,79	96,13	95,83	97,08	96,33	94,14	95,77	**95,68**
10 dB	88,27	91,93	92,87	87,10	92,08	91,02	92,69	91,98	88,39	90,66	**90,70**
5 dB	74,24	74,64	79,09	74,42	77,96	76,69	80,76	79,57	75,25	76,81	**76,94**
0 dB	46,91	42,08	47,45	47,95	48,45	47,40	53,77	48,50	48,42	46,46	**47,74**
-5 dB	18,64	14,15	13,84	21,44	19,37	18,32	20,64	16,08	18,76	18,56	**17,98**
Media	80,08	80,60	82,84	80,30	82,46	81,75	84,52	82,85	80,68	81,48	**81,76**

Tabla 6.3: Resultados de reconocimiento, WAcc (%), obtenidos para los tres conjuntos de evaluación de Aurora2 usando modelos entrenados con voz limpia y características extraídas por el ETSI FE ecualizadas (HEQ).

en la tabla 6.4. A la vista de estos resultados debemos destacar la gran capacidad que posee este estándar para mitigar la degradación producida por el ruido en las características de voz. Así, el sistema que emplea las características extraídas por el ETSI AFE proporciona mejoras relativas medias del 50,37 %, 31,49 % y 6,43 % respecto a los sistemas basados en características ETSI FE sin normalizar, normalizadas en media y ecualizadas, respectivamente. Al igual que antes, las mejoras alcanzadas son especialmente notables a SNRs intermedias. También es de destacar el excepcional rendimiento alcanzado por el ETSI AFE en el ruido tipo *car* (motor de un automóvil) del *set* A. Al ser este ruido relativamente estacionario, su densidad de potencia espectral puede ser estimada fácilmente por los módulos oportunos del *front-end* avanzado y, por consiguiente, la señal a la salida del doble filtrado de Wiener resultará muy parecida a la señal de voz original.

	SET A				SET B				SET C		
SNR	Subway	Babble	Car	Exhibition	Restaurant	Street	Airport	Station	Subway	Street	**Media**
Limpio	99,32	99,06	99,16	99,35	99,32	99,06	99,16	99,35	99,29	99,21	**99,23**
20 dB	98,13	98,07	98,78	98,21	98,07	97,82	98,60	98,61	97,42	97,82	**98,15**
15 dB	96,28	96,86	97,85	96,70	95,55	96,58	97,61	97,04	95,61	96,52	**96,66**
10 dB	92,69	92,02	95,97	93,52	91,77	93,29	94,33	94,79	91,65	92,26	**93,23**
5 dB	85,94	81,77	90,16	85,50	79,18	84,61	86,88	86,52	82,84	82,29	**84,57**
0 dB	66,50	53,14	70,80	65,87	54,87	64,03	64,27	67,42	59,10	58,95	**62,50**
-5 dB	35,12	21,64	33,46	34,99	22,38	32,26	31,40	34,00	29,44	28,60	**30,33**
Media	87,91	84,37	90,71	87,96	83,89	87,27	88,34	88,88	85,32	85,57	**87,02**

Tabla 6.4: Resultados de reconocimiento, WAcc (%), obtenidos para los tres conjuntos de evaluación de Aurora2 usando modelos entrenados con voz limpia y características extraídas por el ETSI AFE.

Para terminar este apartado, en la tabla 6.5 mostramos los resultados de reconocimiento de palabras para la base de datos Aurora4. En la tabla se ha notado por FE, CMN y HEQ a los sistemas de reconocimiento que emplean las características extraídas por el estándar ETSI FE sin normalizar, normalizadas en media y ecualizadas, respectivamente. El sistema AFE, por otro lado, usa los coeficientes MFCC extraídos por el estándar ETSI AFE. Además de la media global de reconocimiento sobre los catorce conjuntos de evaluación incluidos en Aurora4, en la tabla también se muestra la mejora relativa (M.R.) en tanto por ciento de cada sistema con respecto al sistema de referencia FE.

Las conclusiones podemos extraer a partir de los resultados plasmados en la tabla 6.5 son similares a las extraídas para Aurora2. En primer lugar, constatamos la fragilidad de la parametrización ETSI FE frente al ruido, sólo que en este caso el efecto del ruido sobre el rendimiento del sistema es mucho mayor, al ser más compleja la tarea de reconocimiento de Aurora4 que la de Aurora2. Incluso aplicando CMN o HEQ las tasas de reconocimiento obtenidas no superan el 50 % en algunos casos (p.ej. en los conjuntos T-07 y T-14). Otro aspecto que se pone de relevancia en Aurora4 es el efecto de la distorsión producida por las diferentes respuestas en frecuencia de los micrófonos empleados en la grabación de las frases de *test*. Esta distorsión puede interpretarse como un ruido convolutivo y hace que, en el caso del sistema FE, la tasa de reconocimiento caiga desde el 82,26 % obtenido para el conjunto T-01, hasta el 62,92 % de palabras reconocidas en T-08. De nuevo el sistema que mejor rendimiento ofrece es el que emplea el estándar ETSI AFE como método de extracción de características. Para este sistema, exceptuando la condición limpia del conjunto T-01, el conjunto de evaluación donde se logran los mejores resultados es T-02, que vuelve a coincidir con

	T-01	T-02	T-03	T-04	T-05	T-06	T-07	T-08	T-09	T-10	T-11	T-12	T-13	T-14	Media	M.R.
FE	87,26	54,96	34,17	39,34	34,15	38,11	31,31	62,92	40,31	25,74	28,88	23,61	29,33	22,17	39,45	–
CMN	87,69	75,30	53,24	53,15	46,80	56,36	45,38	77,04	64,24	45,30	42,07	36,15	47,43	36,67	54,77	38,85
HEQ	87,75	75,83	60,71	60,13	60,34	62,58	58,19	78,24	66,67	51,62	50,36	47,58	52,42	47,30	61,41	55,67
AFE	88,25	81,41	69,14	64,80	67,44	66,34	68,78	80,57	74,76	61,89	56,47	58,75	60,13	59,87	68,47	73,58

Tabla 6.5: Resultados de reconocimiento, WAcc (%), obtenidos para la base de datos Aurora4 usando modelos acústicos entrenados con voz limpia y distintas representaciones: (1) características sin compensar extraídas por el ETSI FE (FE), (2) características extraídas por el ETSI FE compensadas en media (CMN), (3) características ETSI FE ecualizadas (HEQ) y (4) características extraídas por el ETSI AFE (AFE).

frases contaminadas con ruido estacionario del motor de un coche.

Con esto concluimos la exposición de los resultados de referencia a los que nos remitiremos en los siguientes apartados para comparar las técnicas propuestas. Pasamos a continuación a la evaluación experimental de las técnicas propuestas en este trabajo.

6.2.3. Técnicas de compensación basadas en datos estéreo

En este apartado procedemos a presentar los resultados de reconocimiento de voz obtenidos por las técnicas de compensación descritas en el capítulo 3. Como comentamos anteriormente, dada la reducida variabilidad de los ruidos incluidos tanto en Aurora2 como en Aurora4, en este apartado emplearemos la versión ampliada de Aurora2 descrita en la sección 6.1.1.2. Las técnicas propuestas serán evaluadas por medio de reconocedores (HMMs) entrenados con voz limpia. Los detalles de estos modelos (topología, número de estados, gaussianas por estado, etc.) pueden encontrarse en la sección 6.1.3.1. En cuanto a la representación de la voz empleada, de partida tomaremos la que proporciona el estándar ETSI FE (MFCCs junto con sus derivadas discretas), aunque más tarde también se realizarán pruebas con el *front-end* avanzado (ETSI AFE).

Como se estudió en el capítulo 3, las técnicas de compensación propuestas en ese capítulo emplean diccionarios VQ para modelar las características de voz. El entrenamiento de estos diccionarios se lleva a cabo usando el algoritmo de las k-medias [178] sobre cada uno de los 55 conjuntos distintos de ficheros de entrenamiento disponibles, resultando en un total de 55 diccionarios VQ. Salvo que se indique explícitamente lo contrario, los diccionarios contendrán 256 celdas VQ. Inicialmente estos diccionarios modelarán únicamente las características estáticas de la voz (MFCCs). No obstante, en la sección 6.2.3.2 se estudiará el efecto que tiene sobre la compensación el modelado conjunto de las características estáticas y dinámicas. A partir de los diccionarios VQ anteriores y de las grabaciones estéreo disponibles, se procederá a estimar los pa-

rámetros de las transformaciones oportunas que cada técnica de compensación aplica (p.ej. vectores de corrección, subregiones, etc.). En el cálculo de estas transformaciones siempre estarán involucrados dos diccionarios VQ, uno que modela las características de voz limpia y otro las características degradadas con un cierto ruido acústico, ambos con el mismo número de celdas VQ ($M_x = M_y = 256$).

Además de las técnicas descritas en el capítulo 3, en los siguientes apartados también evaluaremos el rendimiento mostrado por dos de las técnicas de compensación basadas en datos estéreo más conocidas en la literatura: SPLICE [67, 68] y MEMLIN [43, 44]. Al contrario que nuestra propuesta, SPLICE y MEMLIN modelan los espacios de voz mediante GMMs. De forma similar al entrenamiento de los diccionarios VQ, estos GMMs se entrenarán usando los 55 conjuntos de entrenamiento disponibles, pero empleando en este caso el algoritmo EM [66]. Para hacer justa la comparación entre las distintas técnicas de compensación, los GMMs empleados tendrán 256 gaussianas con covarianzas diagonales.

En todos los casos, sean las técnicas de compensación que aquí proponemos o SPLICE/MEMLIN, la compensación de las características estáticas se realizará trama a trama. A partir de las características compensadas se calcularán entonces los parámetros dinámicos de la voz y, finalmente, el reconocedor empleará los vectores resultantes para descodificar la señal de voz.

La evaluación experimental de las técnicas de compensación propuestas se organizará de la siguiente forma. En el siguiente apartado evaluaremos dichas técnicas bajo la situación ideal en la que se conoce a priori el tipo de degradación que contamina la señal de voz observada. El objetivo de estos experimentos es el estudio del rendimiento máximo que podemos esperar de cada técnica de compensación cuando no se producen errores en la identificación del entorno acústico que contamina la voz. Seguidamente, en la sección 6.2.3.2 mostraremos las mejoras obtenidas mediante la explotación de las correlaciones temporales de la voz por medio del uso de modelos ocultos de Markov. La sección 6.2.3.3 abarcará el estudio de situaciones realistas en la que se desconoce la identidad del ruido que contamina la señal de voz. Esta situación se abordará mediante la adopción del esquema basado en múltiples modelos estudiado en la sección 3.2.4 y el tratamiento de la incertidumbre de la estimación en el reconocedor.

6.2.3.1. Experimentos oráculo

La tabla 6.6 muestra los resultados de reconocimiento en términos de precisión de palabras, WAcc (%), obtenidos por diferentes sistemas de reconocimiento en la base de datos Aurora2 ampliada. Para cada sistema se presenta la media parcial para cada valor de SNR, así como la media global (Media) para las SNRs en el intervalo $[0, 20]$

dB. Los sistemas que han aparecen recogidos en la tabla son los siguientes. El sistema de referencia ETSI FE (FE) emplea la representación estándar calculada por el *front-end* básico de la ETSI y modelos de voz entrenados con voz limpia. Por otro lado, el sistema ETSI AFE (AFE) emplea la parametrización más robusta al ruido extraída por el *front-end* avanzado. El sistema identificado como "Ideal" denota la condición de reconocimiento ideal en presencia de ruido acústico, esto es, cada frase de *test* se descodifica usando unos modelos acústicos entrenados con voz degradada en las mismas condiciones que la frase que se reconoce.

El resto de sistemas que aparecen en la tabla emplean diferentes técnicas de compensación sobre la parametrización ETSI FE (coeficientes MFCC estáticos) y modelos entrenados con voz limpia. En todos los casos las técnicas se han aplicado asumiendo que se dispone de información oráculo sobre el tipo de degradación que afecta la voz, esto es, se conoce el tipo de ruido y nivel de SNR. Al disponer de esta información oráculo podemos evaluar la cota superior del rendimiento de cada técnica, ya que únicamente se emplearán los modelos y transformaciones específicamente entrenadas para combatir esa degradación.

Las técnicas de compensación que se han evaluado son las siguientes. En primer lugar, en la tabla aparecen SPLICE y MEMLIN, así como dos versiones discretas HD (*Hard Decision*) de las mismas. Estas dos versiones se desarrollaron a fin de simular el comportamiento de los diccionarios VQ mediante los GMMs que estas técnicas emplean. Así, las versiones HD de SPLICE y MEMLIN únicamente usan la gaussiana más probable en cada instante de tiempo para estimar las características de voz. El resto de técnicas que aparecen en la tabla se derivan del estimador VQMMSE propuesto en el capítulo 3, a saber, Q-VQMMSE (sec. 3.2.1), S-VQMMSE (sec. 3.2.2), J-VQMMSE (sec. 3.2.3) y W-VQMMSE (sec. 3.2.3). Para estas dos últimas técnicas se muestra, además, distintas configuraciones obtenidas en función del tipo de matrices de covarianza empleadas en la estimación: matrices identidad (iJ-VQMMSE y iW-VQMMSE, que en este caso degeneran en técnicas idénticas), matrices diagonales (dJ-VQMMSE y dW-VQMMSE) y matrices no diagonales o completas (fJ-VQMMSE y fW-VQMMSE).

Como cabría esperar, el ruido degrada seriamente el rendimiento de los sistemas de reconocimiento de nuestro estudio. Incluso el sistema Ideal en el que se reconoce con modelos acústicos entrenados con voz ruidosa, el rendimiento a bajas SNRs (0 y -5 dB) es simplemente inaceptable para una tarea tan simple como la que define Aurora2. Esta degradación se debe a la pérdida de información provocada por el ruido, haciendo que parte de la información original sea irrecuperable. De nuevo sorprende la capacidad del estándar ETSI AFE (AFE en la tabla) para reducir la discrepancia entre los modelos acústicos entrenados con voz limpia y las frases degradadas que se

Técnica	Limpio	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Media
FE	99,02	90,79	75,53	50,70	25,86	11,27	6,18	50,83
AFE	99,22	98,24	96,95	93,68	84,37	62,46	29,53	87,14
Ideal	99,02	98,66	98,29	97,02	92,16	75,78	34,88	92,38
SPLICE	99,02	98,09	95,87	88,88	70,62	39,04	15,99	78,50
MEMLIN	99,02	98,36	97,01	92,43	78,26	47,03	18,76	82,62
HD-SPLICE	99,02	97,95	95,28	87,52	67,97	37,21	15,74	77,19
HD-MEMLIN	99,02	98,30	96,74	91,42	75,79	44,53	18,11	81,36
Q-VQMMSE	96,19	93,72	90,21	81,24	61,82	31,33	14,39	71,66
S-VQMMSE	99,02	97,93	96,28	90,57	74,70	43,02	18,57	80,50
iJ-VQMMSE	99,02	98,23	96,79	91,60	76,82	46,60	20,02	82,01
dJ-VQMMSE	99,02	97,01	95,22	89,44	73,77	41,90	17,93	79,47
fJ-VQMMSE	99,02	97,37	95,76	89,64	74,10	42,38	18,09	79,85
iW-VQMMSE	99,02	98,23	96,79	91,60	76,82	46,60	20,02	82,01
dW-VQMMSE	99,02	98,33	97,06	92,43	78,70	48,88	20,26	83,08
fW-VQMMSE	99,02	98,37	97,15	92,88	79,61	50,04	20,89	83,61

Tabla 6.6: Tasa de palabras reconocidas (WAcc) obtenidas por diferentes técnicas de compensación en la base de datos Aurora2 ampliada. En el caso de las técnicas basadas en grabaciones estéreo, los resultados han sido obtenidos usando información oráculo sobre la identidad del ambiente acústico que degrada la señal observada.

reconocen, obteniendo una mejora relativa del 71,43 % respecto al sistema basado en el ETSI FE (FE en la tabla). Debemos notar que el ETSI AFE podría considerarse como una técnica de compensación basada en modelos de distorsión, mientras que el resto de técnicas de compensación se basan en datos estéreo. En el siguiente apartado veremos que ambos enfoques no son mutuamente excluyentes: las técnicas basadas en datos estéreo pueden trabajar con características de voz extraídas por el ETSI AFE a fin de reducir el error residual que este estándar no es capaz de eliminar.

El rendimiento alcanzado por las distintas técnicas de compensación basadas en datos estéreo supera con creces el rendimiento del sistema de referencia (ETSI FE). En el caso de las técnicas basadas en el estimador VQMMSE observamos diferentes comportamientos. En primer lugar, constatamos que existe una gran diferencia entre el estimador Q-VQMMSE y el resto de técnicas. Esta diferencia se debe a que Q-VQMMSE es la única técnica en la que el vector de características observado se cuantifica explícitamente, mientras que en el resto de técnicas el proceso de cuantificación sólo afecta al cómputo de las probabilidades del estimador MMSE. Así, el error que se incurre al cuantificar la observación explica la diferencia absoluta de 8,84 % que existe entre las técnicas Q-VQMMSE y S-VQMMSE.

Por otro lado, la tabla 6.6 también refleja las bondades de modelar mejor las transformaciones que se producen en las características de voz entre los espacios limpio y distorsionado a causa del ruido. Por ejemplo, tanto S-VQMMSE como las técnicas iJ-VQMMSE y iW-VQMMSE aplican una corrección aditiva de la forma $\hat{\boldsymbol{x}} = \boldsymbol{y} - \boldsymbol{r}$ al vector observado $\boldsymbol{y}$ para compensar la degradación producida por el ruido. No obstante, aunque la forma de la transformación que se aplica sea la misma, los resultados obtenidos son diferentes: 80,50 % para S-VQMMSE y 82,01 % para las otras dos técnicas. Esta diferencia de rendimiento se debe a que los vectores de corrección $\boldsymbol{r}$ se estiman de forma más precisa en las técnicas iJ-VQMMSE y iW-VQMMSE. Un paso más en la idea de representar mejor la degradación debida al ruido consiste en modelarla mediante transformaciones afines de la forma $\hat{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}$. Esto es justamente lo que se hace en las versiones de las técnicas J-VQMMSE y W-VQMMSE que emplean matrices de covarianza diagonales y no diagonales. Al menos en la técnica W-VQMMSE, cuanto más compleja sea la transformación aplicada sobre los vectores de voz ruidosa mejores son los resultados de reconocimiento obtenidos. Así vemos que los mejores resultados de reconocimiento (83,61 %) se obtienen cuando W-VQMMSE emplea matrices de covarianza completas (fW-VQMMSE) en el modelado de las subregiones que esta técnica estima. No obstante, la pequeña diferencia en rendimiento entre el uso de matrices de covarianza diagonales (83,08 %) y matrices completas (83,61 %), puede que no esté justificada por el sobrecoste computacional que supone el uso de estas últimas.

Un resultado sorprendente que revela la tabla es la caída en el porcentaje de palabras reconocidas que se produce en la técnica J-VQMMSE cuando se usan matrices de covarianza distinta a la identidad en el modelado de las subregiones. Recordemos que tanto J-VQMMSE como W-VQMMSE se basan en un modelado de subregiones de los espacios de características limpio y distorsionado. La diferencia entre ambas técnicas radica en la forma en la que se deducen las transformaciones entre las subregiones de ambos espacios. La transformación que J-VQMMSE aplica asume que los vectores de características de cada par de subregiones, limpia y distorsionada, pueden modelarse conjuntamente como una distribución normal multivariante (ver ecuación (3.41)). Por otro lado, la técnica W-VQMMSE se deriva de la transformación de blanqueo de la ecuación (3.42), requiriendo únicamente que ambas subregiones por separado sean gaussianas. Esta diferencia entre las suposiciones sobre las que se fundamentan ambas técnicas, J-VQMMSEy W-VQMMSE, explica los resultados obtenidos. En particular, el modelado independiente de los espacios de características limpio y distorsionado mediante diccionarios VQ hace que la suposición en la que se basa J-VQMMSE (distribución conjunta tipo gaussiano) deje de ser cierta. Por tanto, para un correcto funcionamiento de J-VQMMSE se requeriría un modelado conjunto de ambos espacios

(ruidoso y limpio).

La tabla 6.6 también muestra los resultados de reconocimiento obtenidos por las técnicas SPLICE y MEMLIN en la base de datos Aurora2 ampliada. La comparativa entre estas dos técnicas revela la superioridad de MEMLIN sobre SPLICE, justificándose esta superioridad por el modelado más complejo que MEMLIN lleva a cabo [44]. Por otro lado tenemos que, de entre las técnicas de compensación que proponemos en este trabajo, la más parecida a MEMLIN es iW-VQMMSE. Comparando las tasas de reconocimiento obtenidas por ambas métodos observamos que iW-VQMMSE es ligeramente inferior a MEMLIN, estando justificada esta diferencia por la decisión ruda que se aplica en los cuantificadores VQ (la observación sólo pertenece a una celda VQ), frente a la decisión suave de MEMLIN. No obstante, debemos remarcar que la técnica iW-VQMMSE es computacionalmente más eficiente que MEMLIN (ver tabla 3.1).

Para incrementar la eficiencia de MEMLIN y SPLICE, podemos intentar emular el comportamiento de los diccionarios VQ eligiendo, en cada instante de tiempo, únicamente la gaussiana del GMM ruidoso que proporciona la mayor probabilidad a posteriori. Esto es justamente lo que hacen las versiones HD de SPLICE y MEMLIN. En este caso se puede apreciar que el rendimiento de ambas técnicas decae por debajo de iW-VQMMSE. Esto parece indiciar que los parámetros estimados por las técnicas VQMMSE para cada entorno (vectores de corrección o subregiones) compensan en parte la degradación introducida por la cuantificación VQ. Por tanto, a pesar de que a priori pueda parecer que el modelado VQ es menos potente que el basado en GMMs, los resultados de la tabla demuestran que un diseño cuidadoso del estimador puede hacer que ambos enfoques sean prácticamente equivalentes.

Para concluir este apartado, en la figura 6.4 se muestra el rendimiento de las técnicas estudiadas en función del número de componentes (gaussianas o celdas VQ) de los modelos de voz empleados. De las dos técnicas que usan un modelado basado en subregiones, J-VQMMSE y W-VQMMSE, sólo presentamos los resultados correspondientes a W-VQMMSE por ser la que mejores resultados proporciona. Como cabría esperar, existe una correlación positiva entre la complejidad del modelo acústico empleado (medida en número de componentes) y la tasa de reconocimiento obtenida: a mayor número de componentes, mejor quedan representados los distintos espacios de características y, por consiguiente, más fina será la compensación de la voz. De nuevo observamos que la técnica Q-VQMMSE se aleja del comportamiento global del resto de técnicas debido a la cuantificación explícita de los vectores de voz que esta técnica lleva a cabo.

Un caso particular que merece la pena considerar es aquél en el que la compensación de las características de voz se realiza con modelos de voz (diccionarios VQ o GMMs)

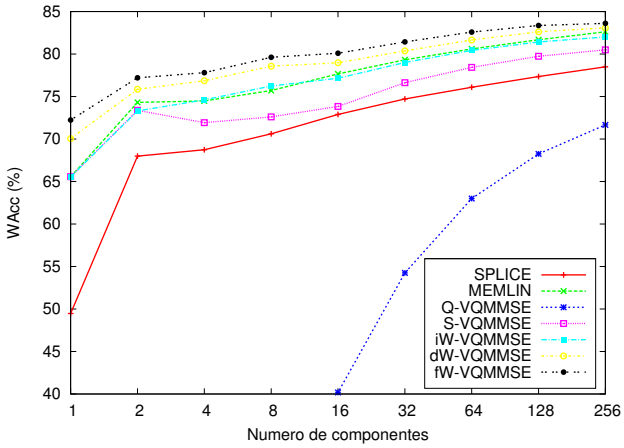

Figura 6.4: Tasas de reconocimiento de palabras obtenidas por diferentes técnicas de compensación en función del número de componentes (gaussianas o celdas VQ) del modelo de voz.

de una única componente. En este caso, las técnicas de compensación que aparecen en la gráfica funcionan de cierta forma como técnicas de normalización dependientes del tipo de ruido y nivel de SNR considerados, lo que justifica que los resultados obtenidos no sean excesivamente bajos. A modo de ejemplo, las técnicas S-VQMMSE y iW-VQMMSE pueden verse como técnicas de normalización de la media (CMN) donde la corrección que se aplica a los vectores observados es constante para cada tipo de ruido y nivel de SNR. Por otra parte, el funcionamiento de la técnica Q-VQMMSE en este caso trivial cae hasta niveles del 7,67 % de palabras reconocidas (no mostrado en la gráfica). La justificación de este comportamiento es simple: esta técnica reemplaza los vectores observados por una combinación lineal de los centroides del diccionario VQ que modela el espacio limpio. Al contar este diccionario únicamente con un centroide, se da la situación de que todos los vectores de la elocución ruidosa se sustituyen por la media del espacio limpio. Finalmente, en la gráfica también observamos que para este caso trivial el modelo basado en subregiones y el basado en celdas VQ convergen.

6.2.3.2. Modelado temporal de la voz

Continuamos con la evaluación de las técnicas de compensación basadas en VQ cuando se realiza un modelado conjunto de la distribución en frecuencia y la evolución temporal de las características de voz. Al igual que en el apartado anterior, aquí volvemos a asumir que se conoce por adelantado el tipo de ruido que contamina cada frase de *test* y su nivel de SNR.

Técnica		Limpio	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Media
FE		99,02	90,79	75,53	50,70	25,86	11,27	6,18	50,83
AFE		99,22	98,24	96,95	93,68	84,37	62,46	29,53	87,14
Ideal		99,02	98,66	98,29	97,02	92,16	75,78	34,88	92,38
FE	SPLICE	99,02	98,40	97,56	94,48	82,24	46,42	18,63	83,82
	MEMLIN	99,02	98,52	97,85	95,58	87,11	59,28	24,52	87,67
	Q-VQMMSE	95,46	93,21	91,05	84,70	65,85	33,42	13,97	73,65
	S-VQMMSE	99,02	98,17	97,07	93,31	80,81	49,37	20,14	83,75
	iW-VQMMSE	99,02	98,38	97,46	94,62	84,75	56,18	23,73	86,28
	dW-VQMMSE	99,02	98,39	97,59	94,91	85,73	58,36	24,00	87,00
	fW-VQMMSE	99,02	98,35	97,56	94,93	86,07	59,64	24,61	87,31
AFE	Q-VQMMSE	95,60	93,56	91,28	85,25	70,23	39,20	12,84	75,90
	S-VQMMSE	99,22	98,32	97,39	94,71	86,30	63,07	27,46	87,96
	iW-VQMMSE	99,22	98,61	97,93	95,89	89,19	69,46	32,62	90,22
	dW-VQMMSE	99,22	98,70	98,05	96,19	89,93	71,47	34,94	90,87
	fW-VQMMSE	99,22	98,65	97,99	96,10	89,92	72,29	36,57	90,99

Tabla 6.7: Resultados de reconocimiento obtenidos por las técnicas de compensación basadas en datos estéreo al considerar un modelado conjunto de las características estáticas y dinámicas de la voz. También se muestra el efecto de compensar las características extraídas por diferentes *front-ends*.

Tal y como se mencionó en el capítulo 5, una de las estrategias más simples que podemos contemplar para que los modelos de voz representen también la evolución temporal de la voz es por medio de las características dinámicas. Así, en lugar de modelar únicamente las características estáticas, podemos modificar el entrenamiento para que los modelos reflejen la distribución conjunta de las características estáticas y dinámicas. Por su parte, las técnicas de compensación trabajarán conjuntamente sobre los parámetros estáticos y dinámicos a la vez. Además de evaluar el efecto de las características dinámicas sobre la estimación MMSE, las técnicas de compensación se han probado tanto con la parametrización ETSI FE como con la extraída por el ETSI AFE. Esto nos permite comprobar si es factible utilizar las técnicas propuestas como postprocesamiento de sistemas robustos ya existentes (en nuestro caso el ETSI AFE).

En la tabla 6.7 aparecen reflejados los resultados de reconocimiento obtenidos en este caso. En primer lugar, observamos que efectivamente las técnicas de compensación basadas en datos estéreo pueden dirigirse a reducir el error residual que no es capaz de eliminar el ETSI AFE. Aunque el sistema que aquí planteamos no es realista al emplear información oráculo sobre el tipo de ruido, sí que comprobamos que el margen de mejora del que disponemos es amplio: desde el 87,14 % alcanzado por el sistema

ETSI AFE hasta el 90,99 % obtenido por la técnica fW-VQMMSE trabajando sobre parámetros extraídos por este estándar. Vemos, además, que este último valor se acerca al rendimiento ofrecido por el sistema Ideal, superándolo incluso en ciertas ocasiones (p.ej. a -5 dB).

Comparando las tablas 6.6 y 6.7 comprobamos que la inclusión de las características dinámicas en los modelos de voz (diccionarios VQ o GMMs con 256 componentes) resulta en las siguientes mejoras relativas: 6,78 % para SPLICE, 6,11 % para MEMLIN, 4,04 % en S-VQMMSE, 5,21 % en iW-VQMMSE, 4,72 % en dW-VQMMSE y, por último, 4,43 % para fW-VQMMSE, siendo todas las mejoras positivas y rondando el 5 %. En el caso de la técnica Q-VQMMSE, el modelado conjunto resulta en una degradación leve del 0,35 %. De nuevo esta degradación viene producida por la cuantificación que se lleva a cabo en esta técnica: al modelar un espacio de dimensión mayor con el mismo número de centroides, el error de cuantificación crece. La tabla también refleja que el modelado con componentes dinámicas beneficia más a la técnica MEMLIN. No obstante, fW-VQMMSE obtiene resultados similares con un coste computacional menor.

Siguiendo con el estudio sobre la aplicación de modelos de voz más sofisticados en el estimador VQMMSE, otra de las propuestas que hicimos en el capítulo 5 fue el uso modelos ocultos de Markov. Para calcular estos HMMs, basta añadir una matriz con las probabilidades de transición entre las distintas celdas del diccionario (cada celda equivale a un estado del HMM) a los diccionarios VQ usados previamente para el espacio limpio (256 celdas y modelando las características estáticas y dinámicas). Para reflejar explícitamente que ahora se emplean HMMs en lugar de diccionarios VQ, notaremos a las técnicas resultantes mediante el sufijo HMMSE en lugar de VQMMSE.

La tabla 6.8 muestra de forma resumida la tasa de reconocimiento promedio obtenida por cada sistema (considerando ambos estándares de parametrización ETSI FE y ETSI AFE) y la mejora relativa (M.R.) alcanzada por el sistema en cuestión por el uso de HMMs en comparación con los resultados de la tabla 6.7. Otra vez vemos que la aplicación de mejores modelos de voz en el contexto del estimador MMSE implica ganancias positivas en las tasas de palabras correctamente reconocidas. En particular, observamos que el modelado HMM permite paliar parcialmente los problemas de cuantificación que se han venido comentando para la técnica Q-VQMMSE. Para el resto de técnicas la ganancia obtenida por la aplicación de los HMMs es menor que en el caso de Q-VQMMSE, pero significativo en cualquier caso.

	FE		AFE	
Técnica	Media	M.R.	Media	M.R.
Q-HMMSE	83,74	7,83	84,65	11,53
S-HMMSE	87,37	5,92	90,12	2,46
iW-HMMSE	88,19	4,34	90,69	0,52
dW-HMMSE	88,50	4,00	91,07	0,22
fW-HMMSE	88,66	4,16	91,18	0,21

Tabla 6.8: Tasas promedias de reconocimiento de palabras y mejoras relativas (M.R.) alcanzadas por las técnicas de compensación basadas en datos estéreo al aplicar HMMs durante la estimación MMSE .

6.2.3.3. Pruebas en ambientes desconocidos

La experimentación que hemos venido realizando hasta ahora de las distintas técnicas de compensación nos ha permitido acotar el rendimiento esperado de las misma en situaciones ideales, pero no así en situaciones realistas. En este sentido cabe decir que uno de los errores que pueden limitar el rendimiento de estas técnicas es la mala identificación del tipo de ruido y nivel de SNR de la elocución observada. A fin de evaluar el efecto de estos errores en el proceso de compensación, en este apartado recurriremos a varios experimentos de reconocimiento en los que las distintas técnicas se enfrentan tanto a frases contaminadas con ruidos y SNRs conocidas, como a frases contaminadas con ruidos y SNRs diferentes de las empleadas durante el entrenamiento. En definitiva, usaremos para este cometido los dos conjuntos de evaluación, Set A y Set B, definidos dentro de la base de datos Aurora2 ampliada (ver sección 6.1.1.2 para más detalles).

Dado que en estos experimentos desconocemos la identidad de la degradación que contamina la señal de voz, a la hora de calcular las estimas oportunas de voz usaremos el enfoque basado en múltiples modelos de voz descrito en la sección 3.2.4. De forma resumida, en este enfoque se calculan una serie de estimas parciales de voz $\hat{\boldsymbol{x}}^{(e)}$ $(e = 1, \ldots, E)$, una por cada entornos acústico e considerado en el entrenamiento del sistema ($E = 55$ entornos en nuestro caso). Estas estimas finalmente se combinan de acuerdo a sus respectivas probabilidades a posteriori $P(e|\boldsymbol{y})$ calculadas mediante GMMs entrenados para cada condición ruidosa (estos GMMs son los mismos que los empleados por SPLICE y MEMLIN). Además del esquema de modelos múltiples, en estos experimentos evaluaremos las dos técnicas de tratamiento de la incertidumbre descritas en la sección 5.2: la descodificación *soft-data* (SD) y el algoritmo ponderado de Viterbi (WVA). En el caso de la técnica WVA, de los enfoques descritos en la sección 5.2.2 para estimar el factor de pesado exponencial ρ_t al que van elevadas las probabilidades de observación del reconocedor, aquí únicamente consideraremos el cálculo

basado en la entropía de la distribución a posteriori.

Si antes hemos visto cómo las técnicas de compensación propuestas pueden reducir el error residual del ETSI AFE y, de esta forma, mejorar el rendimiento global del sistema de reconocimiento, otra estrategia que también da excelentes resultados es el uso de modelos acústicos multiestilo (multicondición). Así, además de evaluar el rendimiento de los distintos métodos propuestos en base a modelos acústicos entrenados con voz limpia, en este apartado también evaluaremos su rendimiento usando modelos multicondición. Para entrenar estos modelos se han seleccionado 37 de los 55 conjuntos de entrenamiento disponibles en la base de datos Aurora2 ampliada: el conjunto con las elocuciones sin distorsionar, más otros 36 conjuntos resultantes de considerar los 9 tipos de ruido del entrenamiento a 4 niveles de SNR (20, 15, 10 y 5 dB). Estos datos son procesados posteriormente por cada técnica de compensación por separado para obtener una especie de conjuntos de voz pseudolimpios, que son los que finalmente se emplean para entrenar los HMMs del reconocedor.

La tabla 6.9 muestra los resultados de reconocimiento obtenidos por las diferentes técnicas de compensación en los dos conjuntos de evaluación de la base de datos Aurora2 ampliada. La tabla también muestra la media de los resultados obtenidos para ambos conjuntos, así como la mejora relativa (en tanto por ciento) respecto al sistema de referencia ETSI FE correspondiente (usando modelos limpios o multiestilo entrenados con este estándar). De las técnicas derivadas del estimador VQMMSE que hemos ido evaluando hasta ahora, en esta tabla sólo presentamos los resultados correspondientes a la técnica dW-HMMSE, ya que ésta supone un compromiso entre precisión del estimador y eficiencia computacional del mismo. Asimismo hemos de remarcar que los resultados presentados para SPLICE y MEMLIN no son comparables a los de la técnica dW-HMMSE, puesto que las dos primeras técnicas no modelan la evolución temporal de la voz. En este sentido debemos decir que un mejor modelado en estas técnicas implicaría mejores resultados de reconocimiento, tal y como ocurre en nuestras propuestas.

A grandes rasgos los resultados de la tabla 6.9 nos demuestran el buen resultado alcanzado por el esquema de múltiples modelos para ruidos conocidos (Set A), no siendo así para aquellas condiciones acústicas no consideradas en la fase de entrenamiento (Set B). En el Set B observamos una pérdida del rendimiento generalizada para todas las técnicas de compensación. Esta pérdida, no obstante, se mitiga parcialmente mediante el uso de modelos multicondición, que como vemos incrementan el rendimiento base de todas las técnicas respecto a reconocer con modelos entrenados con voz limpia.

Otra opción complementaria para reducir la discrepancia generada al trabajar bajo condiciones no vistas en el entrenamiento, consiste en explotar la incertidumbre residual

Técnica		Set A	Set B	Media	M.R.
HMMs limpios	FE	50,83	40,28	45,56	–
	SPLICE	83,07	73,90	78,49	72,28
	MEMLIN	86,77	75,52	81,15	78,12
	dW-HMMSE	89,10	77,69	83,40	83,06
	dW-HMMSE+SD	89,36	79,72	84,54	85,56
	dW-HMMSE+WVA	89,83	81,41	**85,62**	87,93
HMMs multiestilo	FE	88,73	79,78	84,26	–
	SPLICE	88,87	80,92	84,90	0,77
	MEMLIN	89,78	80,92	85,35	1,30
	dW-HMMSE	90,86	82,73	86,80	3,01
	dW-HMMSE+SD	90,67	82,99	86,83	3,06
	dW-HMMSE+WVA	90,82	85,34	**88,08**	4,54

Tabla 6.9: Tasas de reconocimiento obtenidas por las técnicas de compensación basadas en datos estéreo en combinación con el esquema de realce basado en múltiples modelos acústicos. En la tabla se incluye los resultados obtenidos por las distintas técnicas para modelos entrenados con voz limpia y modelos multicondición. Asimismo se presentan los resultados obtenidos por dos técnicas de explotación de la incertidumbre: *soft-data* (SD) y el algoritmo ponderado de Viterbi (WVA).

del proceso de estimación MMSE. A este respecto también constatamos los ambos enfoques SD y WVA incrementan la robustez frente al ruido acústico de los distintos sistema de reconocimiento: para el Set B las mejoras relativas obtenidas por el enfoque SD son de 2,61 % en caso de usar modelos limpios y 0,31 % para modelos multiestilo, mientras que la técnica WVA consigue unas mejoras del 4,79 % si se usan modelos limpios y del 3,15 % si los modelos multicondición. En el Set A observamos que la mejora conseguida es menor, puesto que la incertidumbre para los ruidos conocidos es también menor.

Comparando las técnicas SD y WVA comprobamos que en todos los casos el rendimiento de los sistemas basados en WVA es superior al de los basados en SD. Dos son las razones que justifican este comportamiento. En primer lugar, SD asume que la distribución a posteriori de la voz limpia $p(\boldsymbol{x}|\boldsymbol{y})$ (la PDF de evidencia que mencionábamos en el capítulo 5) es gaussiana, cosa que por supuesto no tiene por qué ser cierta en la gran mayoría de situaciones [48, 122]. Por otro lado, en la técnica WVA se ajustan experimentalmente los parámetros de la función sigmoide empleada para calcular el factor de pesado que afecta a las probabilidades de observación del reconocedor. En el caso de la técnica SD, las varianzas calculadas por el estimador no se someten a este ajuste, por lo que pueden estar sesgadas con respecto a su valor real.

Retomando el análisis de los resultados de la tabla 6.9, se observa que las técnicas SPLICE y MEMLIN sufren una pequeña degradación en el Set A respecto a los resultados oráculo mostrados en la tabla 6.7: 0,75 % para SPLICE y 0,10 % para MEMLIN. Si bien este error puede ser atribuido a las pequeñas discrepancias que se producen en el esquema de múltiples modelos durante la identificación del ruido y SNR que contamina la señal de voz, observamos que en el caso de la técnica dW-HMMSE la aplicación de este esquema repercute en una mejora relativa del 0,68 % con respecto a los resultados de la tabla 6.8. La explicación de esta mejora la podemos atribuir a una mejor identificación de la SNR instantánea: en ciertas ocasiones la SNR instantánea trama a trama se desvía de la global y, por consiguiente, puede que resulte más provechoso emplear modelos entrenados para valores de SNR ligeramente distintos. En el esquema de múltiples modelos esto se consigue empleando modelos entrenados con el mismo tipo de ruido pero a una SNR diferente, o interpolando las correcciones obtenidas para dos o más modelos acústicos.

6.2.4. Técnicas de reconstrucción espectral

En la sección anterior hemos evaluado las técnicas de compensación propuestas en el capítulo 3 de esta tesis. Una de las primeras conclusiones que extraemos de los resultados obtenidos es la limitación que supone el requerimiento de datos estéreo en estas técnicas. Asimismo, en la tabla 6.9 observamos que el rendimiento de las mismas se degrada cuando las condiciones de evaluación son distintas de las consideradas durante la fase de entrenamiento. Para paliar en parte estas deficiencias, en el capítulo 4 propusimos dos técnicas de reconstrucción espectral que eluden el uso de grabaciones estéreo al emplear información a priori sobre el efecto del ruido (aditivo) sobre el espectro de la voz. En concreto, la técnica de imputación TGI descrita en la sección 4.2 emplea máscaras de segregación binarias que identifican los elementos del espectro dominados por la energía de la voz de los dominados por el ruido. Por otro lado, la técnica MMSR planteada en la sección 4.3 emplea descripciones probabilísticas sobre la distribución del ruido, esto es, modelos a priori del ruido.

Esta sección la dedicamos a la evaluación experimental de ambas técnicas de reconstrucción. Al igual que antes, la evaluación experimental se llevará a cabo mediante una serie de experimentos de reconocimiento automático de voz sobre bases de datos que contienen grabaciones de voz limpia y grabaciones con ruido añadido de forma artificial. En particular, las bases de datos que emplearemos en esta sección son Aurora2 (la base de datos estándar, no la versión ampliada) y Aurora4. Los modelos acústicos y modelos de lenguaje empleados durante los experimentos de reconocimiento son los que se detallan en las secciones 6.1.3.1 y 6.1.3.2. En cuanto a la parametrización de la voz

usada, emplearemos la extraída por el estándar básico de la ETSI (FE). Al contrario que las técnicas de compensación basadas en datos estéreo, TGI y MMSR trabajan en el dominio del banco de filtros Mel (log-Mel). Por tanto, después de obtener las estimas oportunas de voz en este dominio, se empleará la DCT para calcular los parámetros MFCC correspondientes. A partir de estos parámetros se computarán las características dinámicas de la voz y, por último, se aplicará CMN para incrementar la robustez de los parámetros extraídos frente al ruido convolutivo.

Las técnicas TGI y MMSR las evaluaremos usando modelos de voz (GMMs) entrenados con las características log-Mel (con $D = 23$ filtros Mel) extraídas de los conjuntos de entrenamiento limpios de las bases de datos Aurora2 y Aurora4. En principio los GMMs de voz contarán con $M = 256$ componentes, aunque en algunos experimentos este número se alterará explícitamente para evaluar la influencia de la complejidad del modelo en la precisión de la reconstrucción espectral. A fin de representar fielmente las correlaciones que presentan las características log-Mel, las gaussianas de estos modelos contarán con matrices de covarianza no diagonales (completas).

La evaluación experimental de las técnicas TGI y MMSR se llevará a cabo usando tanto experimentos oráculo como realistas. Los primeros estarán orientados a evaluar el rendimiento de cada técnica en situaciones ideales donde no se producen errores en la estimación de la máscara de segregación, caso de TGI, o en la estimación del modelo de ruido, caso de MMSR. Debe notar que en este caso ambas técnicas son equivalentes (ver sección 4.3.1.1), por lo que sólo incluiremos los resultados de TGI usando máscaras de segregación oráculo. Para calcular estas máscaras, en primer lugar estimaremos el ruido real que contamina cada elocución de *test*. En Aurora2 y Aurora4 esto se realiza fácilmente sin más que sustraer a cada señal de voz ruidosa su correspondiente versión limpia. A partir de este ruido, se calculará el nivel de SNR real para cada elemento del espectrograma. Por último, dicha SNR se umbralizará usando un valor de 7 dB para obtener la máscara binaria que finalmente emplearemos en los experimentos de reconocimiento oráculo.

En los experimentos realistas, por contra, las técnicas TGI y MMSR se evaluarán usando máscaras de segregación y modelos de ruido estimados a partir de la propia señal observada. En el caso de TGI, las máscaras binarias se calcularán usando el mismo procedimiento que acabamos de describir para las máscaras oráculo salvo por dos detalles: (i) el umbral de SNR se optimizará empíricamente para cada técnica usando conjuntos de validación y (ii) se usarán estimaciones del ruido que contamina la señal de voz en lugar del ruido real. Estas estimaciones se obtendrán usando el procedimiento que describimos a continuación:

1. Seleccionar los N primeros y N últimos vectores de características de la elocución,

siendo $N = 20$ en Aurora2 y $N = 35$ en Aurora4.

2. Calcular las dos medias parciales correspondientes a los N primeros y N últimos vectores de la frase:

$$\boldsymbol{\mu}_n^1 = \frac{1}{N}\sum_{t=1}^{N}\boldsymbol{y}_t, \quad \boldsymbol{\mu}_n^2 = \frac{1}{N}\sum_{t=T-N+1}^{N}\boldsymbol{y}_t \tag{6.4}$$

3. Estimación de la potencia del ruido:

   a) Para las N primeras y N últimas tramas de la elocución la potencia del ruido coincide con la propia observación: $\hat{\boldsymbol{n}}_t = \boldsymbol{y}_t$ ($t \in \{1, \ldots, N, N-T+1, \ldots, T\}$).

   b) En el resto de instantes de tiempo ($t \in [N+1, N-T]$) la potencia se estima interpolando linealmente los valores de $\boldsymbol{\mu}_n^1$ y $\boldsymbol{\mu}_n^2$.

4. Si en algún momento la potencia estimada $\hat{\boldsymbol{n}}_t$ supera a la de la propia observación $\boldsymbol{y}_t$, acotamos el valor estimado al valor observado.

Para los experimentos realistas de la técnica MMSR, inicialmente emplearemos estas mismas estimas de ruido junto con una matriz de covarianza fija e invariante al tiempo obtenida también a partir de las N primeras y N últimas tramas. Posteriormente evaluaremos la técnica MMSR en combinación con el algoritmo iterativo de ajuste del GMM de ruido descrito en la sección 4.3.3.

6.2.4.1. Reconstrucción basada en máscaras de segregación

En esta sección evaluaremos el rendimiento de la técnica de imputación TGI propuesta en la sección 4.2. Junto con nuestra propuesta, mostraremos también los resultados experimentales obtenidos por la técnica CBR estudiada en la sección 2.2.5.4. En el siguiente punto evaluaremos ambas técnicas, TGI y CBR, en el contexto de las bases de datos Aurora2 y Aurora4 usando tanto máscaras de segregación oráculo como estimadas.

Resultados iniciales. La tabla 6.10 muestra los resultados de reconocimiento (WAcc) obtenidos por las mencionadas técnicas de imputación en la base de datos Aurora2. Los resultados se han detallado para los casos en los que las técnicas emplean máscaras binarias oráculo (Oráculo) o máscaras binarias estimadas (Real). Asimismo, con un fin meramente comparativo, se incluyen los resultados de referencia de las tablas 6.2 y 6.4 para los estándares ETSI FE+CMN (FE) y ETSI AFE (AFE).

Técnica		Limpio	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Media	M.R.
FE		99,11	97,29	92,55	75,56	42,82	22,69	12,92	66,18	–
AFE		99,23	98,15	96,66	93,23	84,57	62,50	30,33	87,02	31,49
Oráculo	CBR	99,11	98,90	98,52	97,41	94,78	88,05	71,48	95,53	44,35
	TGI	99,11	99,01	98,75	97,99	96,11	90,90	77,34	96,55	45,89
Real	CBR	98,88	97,58	95,28	89,58	77,19	52,77	23,63	82,48	24,63
	TGI	98,91	97,44	95,03	89,12	77,23	53,97	24,57	82,56	24,74

Tabla 6.10: Rendimiento de las técnicas de imputación en la base de datos Aurora2 usando máscaras oráculo y máscaras estimadas.

Como podemos apreciar, en el caso oráculo ambas técnicas de imputación sobrepasan con creces los resultados obtenidos por los sistemas FE y AFE, siendo esta mejora especialmente notable a SNRs intermedias y bajas. En concreto, si comparamos los resultados del sistema TGI con los del FE vemos que la diferencia entre emplear un sistema u otro equivale a una diferencia en SNR de 15 dB (p.ej. el resultado del FE a 10 dB es 75,56 %, siendo este resultado comparable al 77,34 % que la técnica TGI alcanza a -5 dB). En el caso de la comparativa con el *front-end* avanzado esta diferencia es menor, de unos 7,5 dB. Aunque estos resultados hay que interpretarlos con cuidado ya que son oráculo, sí que permiten justificar la validez del modelo de enmascaramiento de la voz en el que se basan las distintas técnicas de reconstrucción que proponemos en este trabajo.

En la situación más realista en la que las máscaras de segregación son estimadas a partir de la propia señal observada, vemos que el rendimiento de ambas técnicas de imputación sufre caídas del 16,95 % (TGI) y 15,82 % (CBR) con respecto a la reconstrucción con máscaras oráculo. Es más, en este caso dichas técnicas se sitúan por detrás del estándar ETSI AFE. La caída en rendimiento que se produce en estas técnicas se debe principalmente a dos causas: la sencillez del algoritmo de estimación de ruido que estamos empleando, el cual no es capaz de estimar correctamente las características de los ruidos poco estacionarios, y el uso de máscaras binarias de segregación. Como ya mencionamos, un error en la estimación de la potencia del ruido puede magnificarse cuando se usan máscaras binarias por las siguientes razones. En la evaluación experimental de la técnica MMSR veremos que parte de estos problemas se pueden suavizar usando máscaras continuas o modelos estadísticos del ruido.

La tabla 6.11 presenta los resultados de reconocimiento para la base de datos Aurora4. En esta base de datos podemos constatar un comportamiento diferenciado en las técnicas de imputación para los conjuntos de evaluación cuyas frases se contaminan únicamente con ruido aditivo (T-02 al T-07) frente a los que también contienen ruido

	T-01	T-02	T-03	T-04	T-05	T-06	T-07	T-08	T-09	T-10	T-11	T-12	T-13	T-14	Media	M.R.
FE	87,69	75,30	53,24	53,15	46,80	56,36	45,38	77,04	64,24	45,30	42,07	36,15	47,43	36,67	54,77	–
AFE	88,25	81,41	69,14	64,80	67,44	66,34	68,78	80,57	74,76	61,89	56,47	58,75	60,13	59,87	68,47	25,01
Oráculo																
CBR	87,69	86,59	82,98	83,82	81,95	85,65	81,30	79,17	77,15	71,87	71,44	68,48	73,81	69,81	78,69	43,67
TGI	87,69	87,07	84,81	85,02	83,22	86,10	83,11	79,97	78,82	73,87	73,40	71,74	75,32	72,41	80,18	46,39
Real																
CBR	87,07	81,32	59,61	53,15	59,01	58,10	61,72	79,36	74,54	53,93	47,36	50,12	51,62	55,80	62,34	13,81
TGI	87,43	81,39	60,00	55,86	58,29	60,36	59,44	79,26	74,63	55,45	47,38	48,38	54,04	54,94	62,63	14,35

Tabla 6.11: Rendimiento de las técnicas de imputación en la base de datos Aurora4 usando máscaras oráculo y máscaras estimadas.

convolutivo (T-08 al T-14). Este último se debe a las diferencias entre las respuestas en frecuencia del micrófono empleado para grabar las frases de entrenamiento y las de los micrófonos empleados para grabar los conjuntos T-08 al T-14. Así, constatamos que mientras en los primeros conjuntos el rendimiento de TGI y CBR usando máscaras oráculo no se aleja demasiado del rendimiento obtenido usando voz limpia (T-01), la tasa de palabras correctamente reconocidas decae en la segunda mitad de los conjuntos de evaluación. Los resultados de los sistemas ETSI FE y ETSI AFE nos demuestran que este problema no es exclusivo de las técnicas de imputación estudiadas, indicando por consiguiente una limitación compartida por las distintas técnicas a la hora de combatir los discrepancias que se generan por el uso de diferentes micrófonos.

Otra vez comprobamos que el uso de máscaras estimadas acarrea una pérdida importante del rendimiento del reconocedor. En esta situación la única condición de *test* para la que las técnicas TGI y CBR muestran un rendimiento aceptable es en el conjunto T-02, que contiene frases contaminadas con el ruido tipo *car*. De nuevo justificamos estos resultados por las limitaciones del estimador de ruido empleado y de las máscaras binarias. No obstante, el margen de ganancia entre los resultados oráculo y los obtenidos usando máscaras reales es amplio, por lo que es de esperar que mejoras tanto en el estimador de ruido, como en las representaciones de la fiabilidad de los elementos del espectro, conduzcan a mejoras significativas en la precisión del reconocimiento.

Para concluir este punto, en la figura 6.5 mostramos dos gráficas con los resultados de reconocimiento oráculo obtenidos por TGI y CBR para cada base de datos en función del número de gaussianas del modelo de voz. Como cabría esperar, TGI supera a CBR en todas las condiciones mostradas. Así, el rendimiento en Aurora2 de TGI con un modelo de una sola gaussiana (92,75 %) es equivalente al acanzado por CBR con modelos de 256 gaussianas (92,04 %). En Aurora4 tenemos que TGI con un modelo de 16 gaussianas obtiene una tasa de reconocimiento de palabras del 79,14 %, tasa

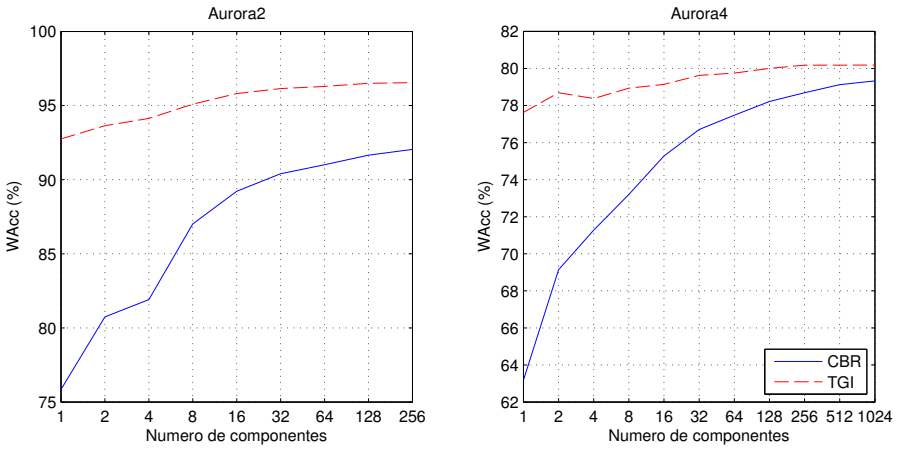

Figura 6.5: Resultados de reconocimiento oráculo obtenidos por las técnicas de imputación TGI y CBR en función del número de gaussianas del modelo de voz.

que coincide con el resultado de CBR con un modelo de 128 gaussianas. Dos son las diferencias clave entre ambas técnicas que justifican estos resultados. En primer lugar, TGI explota mejor que CBR las correlaciones entre las distintas características de voz a través del uso de modelos con matrices de covarianza no diagonales. Por otro lado, las restricciones que impone el modelo de enmascaramiento a las características estimadas se integran completamente en las distribuciones truncadas que TGI emplea, mientras que CBR debe postprocesar las estimaciones ya calculadas para satisfacer estas restricciones.

Modelado temporal de la voz. En este punto evaluaremos las propuestas realizadas en la sección 5.1 sobre del modelado temporal de la voz en el contexto de la reconstrucción de espectros. De esta forma, someteremos a estudio tanto el modelado de las correlaciones temporales de orden corto mediante segmentos de voz (PATCH) descrito en la sección 5.1.1, como el modelado mediante modelos ocultos de Markov (HMM) expuesto en el apartado 5.1.2. Dado que de las técnicas de imputación evaluadas en el punto anterior TGI es la que mejores resultados ofrece, sólo consideraremos la extensión temporal para esta técnica y no para CBR.

Para encontrar el valor óptimo de los parámetros τ y δ de la ventana deslizante usada por la técnica PATCH (ver ecuación (5.1)), se emplearon conjuntos de validación independientes para cada base de datos. En ambos casos, Aurora2 y Aurora4, el valor óptimo de los parámetros resultó ser $\tau = 5$ y $\delta = 1$ para un 95 % de la varianza explicada por PCA. Sorprendentemente, el valor hallado para la longitud de la ventana, $\tau = 5$, coincide aproximadamente con la duración promedio (65 ms) de un fonema en inglés

Técnica		Limpio	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Media	M.R.
FE		99,11	97,29	92,55	75,56	42,82	22,69	12,92	66,18	–
AFE		99,23	98,15	96,66	93,23	84,57	62,50	30,33	87,02	31,49
Oráculo	TGI	99,11	99,01	98,75	97,99	96,11	90,90	77,34	96,55	45,89
	TF-MFR	99,10	99,01	98,84	98,29	97,00	93,67	84,65	97,36	47,11
	PATCH	99,11	99,02	98,81	98,12	96,70	92,84	83,39	97,10	46,71
	HMM	99,11	99,01	98,75	98,22	96,77	92,80	85,99	97,11	46,73
Real	TGI	98,91	97,44	95,03	89,12	77,23	53,97	24,57	82,56	24,74
	TF-MFR	98,94	97,76	95,59	90,28	78,42	55,35	25,96	83,48	26,14
	PATCH	98,93	97,73	95,49	90,15	77,92	54,03	25,38	83,06	25,51
	HMM	98,92	97,49	95,12	89,80	78,96	58,01	29,93	83,88	26,74

Tabla 6.12: Resultados de reconocimiento en Aurora2 de las extensiones para la explotación de las correlaciones temporales de la voz en la técnica de imputación TGI.

[60].

A fin de comparar nuestra propuesta con otras técnicas de la literatura, también evaluaremos la técnica TF-MFR (*Time-Frequency Missing-Feature Reconstruction*, reconstrucción de características perdidas en tiempo-frecuencia) propuesta en [161]. Esta técnica es similar a nuestra aproximación PATCH, salvo que en TF-MFR se disponen de dos conjuntos de modelos diferentes: un GMM similar al que emplea TGI que describe la distribución en frecuencia de la voz (típicamente con 256 gaussianas), y un conjunto de D GMMs de menor tamaño (típicamente con 8 gaussianas cada uno) que modelan la evolución temporal de cada una de las D características log-Mel por separado. Para entrenar este segundo conjunto de GMMs se suelen escoger segmentos cortos de voz conteniendo 8 ó 10 características. Ambos conjuntos de GMMs se aplican entonces para obtener las estimas oportunas de los parámetros perdidos del espectro. Finalmente, las estimas en frecuencia y en tiempo se combinan usando una regla heurística que tiene en cuenta la proporción de elementos fiables en cada estimación. Los parámetros que hemos empleado en la implementación de TF-MFR son los propuestos en [161].

La tabla 6.12 muestra los resultados de reconocimiento promedios obtenidos por las aproximaciones consideradas en la base de datos Aurora2. Podemos constatar que en todos los casos el modelado temporal de la voz redunda en una reconstrucción más precisa de las características log-Mel. Esta mejora cobra especial relevancia a bajas SNRs, donde la proporción de elementos perdidos en el espectro es alta. Así, por ejemplo, a -5 dB la tasa de reconocimiento obtenida por TGI usando máscaras oráculo es del 77,34 %, mientras que los enfoques que consideran información temporal

Técnica		Test 01-07	Test 08-14	Media	M.R.
FE		59,70	49,84	54,77	–
AFE		72,31	64,63	68,47	25,01
Oráculo	TGI	85,29	75,08	80,18	46,39
	TF-MFR	85,94	76,70	81,32	48,47
	PATCH	85,63	77,00	81,31	48,45
	HMM	85,85	78,04	81,94	49,60
Real	TGI	65,71	58,96	62,34	13,81
	TF-MFR	69,21	61,36	65,29	19,19
	PATCH	69,60	62,43	66,02	20,53
	HMM	69,81	63,48	66,65	21,68

Tabla 6.13: Resultados de reconocimiento en Aurora4 de las extensiones para la explotación de las correlaciones temporales de la voz en la técnica de imputación TGI.

obtienen tasas del 84,65 % (TF-MFR), 83,39 % (PATCH) y 85,99 % (HMM).

Los resultados obtenidos para Aurora4 se resumen en la tabla 6.13. A fin de que la presentación de los resultados sea más clara, en la tabla presentamos únicamente los resultados promedios para los conjuntos de evaluación que incluyen ruido aditivo (Test 01-07) y los que consideran además distintos micrófonos (Test 08-14).

De los tres enfoques evaluados, observamos que el que mejores resultados ofrece en ambas bases de datos es el modelado usando HMMs. El mejor comportamiento de este enfoque se justifica en base a que éste considera la elocución al completo para obtener las estimas oportunas de los elementos perdidos del espectro, mientras que las otras dos técnicas (PATH y TF-MFR) sólo consideran un número limitado de elementos vecinos alrededor del estimado. Debido a esto, es de esperar que el enfoque HMM sea más robusto a los posibles errores en la estimación de las máscaras de segregación.

De entre los otros dos enfoques, PATCH demuestra un mejor comportamiento que TF-MFR en las pruebas realistas de Aurora4, no siendo significa esta diferencia en Aurora2. De forma análoga al caso de la aproximación HMM, esto puede ser atribuido al mejor modelado de las correlaciones temporales de orden corto por parte de la técnica PATCH. Por otro lado, la estimación de PATCH se guía únicamente por la estadística de la señal de voz, mientras que TF-MFR usa reglas heurísticas para combinar las diferentes estimaciones que computa.

A pesar de la mejora que supone el uso de los enfoques HMM, PATCH y TF-MFR en combinación con la técnica TGI, en las tablas anteriores constatamos que la imputación aún se encuentra lejos del resultado obtenido por el estándar ETSI AFE. En particular, si lo comparamos con el enfoque HMM, que es el que mejores resultados obtiene en las

Técnica		Limpio	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Media	M.R.
FE		99,11	97,29	92,55	75,56	42,82	22,69	12,92	66,18	–
AFE		99,23	98,15	96,66	93,23	84,57	62,50	30,33	87,02	31,49
Oráculo	WVA	99,13	96,75	94,11	85,62	66,96	42,49	22,39	77,19	16,63
	TGI	99,11	99,01	98,75	97,99	96,11	90,90	77,34	96,55	45,89
	TGI+WVA	99,13	98,93	98,78	98,26	97,02	93,10	82,71	97,22	45,89
Real	TGI	98,91	97,44	95,03	89,12	77,23	53,97	24,57	82,56	24,74
	TGI+WVA	99,14	97,68	95,50	90,36	78,10	51,48	20,59	82,62	24,84

Tabla 6.14: Resultados de reconocimiento con incertidumbre para Aurora2 usando el algoritmo ponderado de Viterbi, WVA, y la técnica de imputación TGI. Se incluyen dos conjuntos de pruebas: Oráculo, donde las máscaras de segregación y/o los pesos que WVA emplea se calculan usando información oráculo, y Real, donde estos valores se estiman para cada elocución.

pruebas realistas, se observan diferencias relativas (a favor de ETSI AFE) de 3,74 % para Aurora2 y 2,73 % en Aurora4. Esto no hace sino confirmar las limitaciones que venimos comentando sobre la imputación de espectros en base a máscaras binarias.

Reconocimiento con incertidumbre. Para concluir la evaluación experimental de la técnica TGI, en este punto analizaremos su combinación con las técnicas de reconocimiento con incertidumbre propuestas en la sección 5.2. De las dos técnicas descritas en esa sección, a saber, técnica *soft-data* y algoritmo ponderado de Viterbi (WVA), en este punto sólo presentaremos los resultados para WVA, ya que como se muestra en la tabla 6.9 y también han comprobado diversos autores (ver p.ej. [48, 122]) WVA permite obtener mejores resultados que *soft-data*. A la hora de calcular los factores de pesado exponencial ρ_t que requiere WVA, nos basaremos en el error cuadrático medio (MSE) de cada reconstrucción (ver sección 5.2.2). Dado que la fiabilidad de la estimación no cambia al expresarla en otro domino (p.ej. en el cepstrum), podremos emplear los pesos ρ_t para ponderar las probabilidades de observación del modelo acústico durante la etapa de reconocimiento.

En la tabla 6.14 se presentan los resultados de reconocimiento para Aurora2 usando el paradigma de reconocimiento con incertidumbre. De nuevo se vuelven a considerar diferentes tipos de experimentos oráculo y realistas. En el caso de los experimentos oráculo, además de emplear máscaras de segregación oráculo usaremos incertidumbres (factores de pesado ρ_t) oráculo. En el caso de la técnica TGI+WVA que aparece en la tabla, estos factores oráculo se calculan a partir del error cuadrático $\|\boldsymbol{x}_t - \hat{\boldsymbol{x}}_t\|^2$, donde $\boldsymbol{x}_t$ es el vector con las características limpias de voz y $\hat{\boldsymbol{x}}_t$ es la reconstrucción obtenida

Técnica		Test 01-07	Test 08-14	Media	M.R.
FE		59,70	49,84	54,77	–
AFE		72,31	64,63	68,47	25,01
Oráculo	WVA	61,27	51,93	56,60	3,33
	TGI	85,29	75,08	80,18	46,39
	TGI+WVA	85,40	76,73	81,06	48,00
Real	TGI	65,71	58,96	62,34	13,81
	TGI+WVA	68,20	60,59	64,39	17,57

Tabla 6.15: Resultados de reconocimiento con incertidumbre para Aurora4 usando el algoritmo ponderado de Viterbi, WVA, y la técnica de imputación TGI. Se incluyen dos conjuntos de pruebas: Oráculo, donde las máscaras de segregación y/o los pesos que WVA emplea se calculan usando información oráculo, y Real, donde estos valores se estiman para cada elocución.

usando máscaras oráculo, ambos vectores expresados en el dominio de los MFCCs. Para el caso de la técnica WVA en solitario, el factor se calcula de la misma forma, pero usando el vector observado $\boldsymbol{y}_t$ en lugar de $\hat{\boldsymbol{x}}_t$. Finalmente, al error cuadrático anterior se le vuelve a aplicar una compresión sigmoidal para que quede en el intervalo $[0, 1]$. En los experimentos realistas, tanto las máscaras como los factores de pesado se estiman a partir de la propia observación.

Como puede apreciarse, el mero hecho de reconocer con la propia elocución ruidosa pero usando incertidumbres oráculo (WVA) supone una mejora relativa del 16,63 % con respecto al sistema de partida ETSI FE, demostrando esto el potencial de la técnica WVA. Por otro parte, el esquema combinado TGI+WVA proporciona los mejores resultados de reconocimiento en la condición oráculo con una gran diferencia sobre AFE. Como cabría esperar, las mejoras producidas son especialmente notables a SNRs bajas. Para el caso de los experimentos realistas la diferencia entre los sistemas TGI y TGI+WVA se pierde, no siendo significativas las diferencias obtenidas entre ambos. Observando los resultados de la tabla 6.14 vemos que el problema reside en los dos niveles de SNR más bajos, donde el sistema TGI+WVA se comporta significativamente peor que TGI. Creemos que el problema reside en los errores que contiene la máscara de segregación a estas SNRs, los cuales provocan, por un lado, una estimación pobre del espectro original de la voz y, por otro, un cómputo desacertado de los pesos ρ_t.

La tabla 6.15 recoge los resultados de reconocimiento para Aurora4. A diferencia de Aurora2, el sistema TGI+WVA ahora se comporta significativamente mejor que TGI en las pruebas realistas, fruto de la mayor complejidad de Aurora4. También hay que destacar que en Aurora4 los niveles de SNR de las elocuciones de *test* son más

altos que en Aurora2, permitiendo así estimaciones más precisas de la fiabilidad de las reconstrucciones.

6.2.4.2. Reconstrucción basada en modelos de ruido

Para concluir con la evaluación de las técnicas de compensación propuestas, en este apartado presentaremos los resultados de reconocimiento obtenidos por la técnica de reconstrucción espectral MMSR propuesta en la sección 4.3.1. Como ya comentamos en dicho capítulo, tanto MMSR como TGI se derivan del modelo de enmascaramiento descrito en la sección 4.1, pero mientras MMSR emplea estimas o modelos de ruido para calcular los valores estimados de voz, TGI emplea máscaras binarias para este fin.

Junto con los resultados de MMSR, en esta sección presentaremos también las tasas de reconocimiento alcanzadas por otras técnicas de compensación similares: (i) imputación TGI usando máscaras oráculo (Oráculo), (ii) imputación TGI usando máscaras binarias estimadas (TGI), (iii) imputación usando máscaras de segregación continuas (S-TGI, *Soft* TGI) y (iv) compensación mediante la técnica VTS con desarrollos de orden 0 (VTS-0) y orden 1 (VTS-1). Para la reconstrucción con máscaras continuas, S-TGI, emplearemos el algoritmo de reconstrucción propuesto en [225] usando las máscaras continuas derivadas del modelo de enmascaramiento (sección 4.3.2). De entre la multitud de versiones que podemos encontrar en la literatura para la técnica VTS, aquí emplearemos la propuesta en [239] por su simplicidad y los buenos resultados que proporciona. En todos los casos se emplearán GMMs con 256 componentes y matrices de covarianza diagonales como modelos a priori para las características log-Mel. Por otro lado, el espectro del ruido se estimará usando el procedimiento descrito más arriba a partir del comienzo y final de cada elocución. Por último, las técnicas MMSR y VTS-1 requieren, para cada estima de ruido, una matriz de covarianza con el error esperado de dicha estimación. Aquí emplearemos una matriz fija para todas las estimas de ruido que se calculará también a partir de los vectores iniciales y finales de la frase.

Resultados iniciales. En la tabla 6.16 se muestran las tasas de reconocimiento obtenidas por las distintas técnicas de compensación en Aurora2. Debe notarse que existe cierta discrepancia que existe entre los resultados de la técnica TGI que aparecen en la tabla y los presentados en los apartados anteriores usando máscaras estimadas (p.ej. en la tabla 6.14). A fin de evaluar todas las técnicas de compensación bajo un marco unificado, en este apartado se ha decidido emplear modelos de voz con covarianzas diagonales, mientras que anteriormente se han empleado modelos con matrices completas. Así, mientras que todas las técnicas evaluadas pueden trabajar con modelos de covarianzas diagonales, MMSR no es capaz de explotar las covarianzas completas,

Técnica	Limpio	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Media	M.R.
FE	99,11	97,29	92,55	75,56	42,82	22,69	12,92	66,18	–
AFE	99,23	98,15	96,66	93,23	84,57	62,50	30,33	87,02	31,49
Oráculo	99,11	99,01	98,74	97,84	95,72	89,64	73,79	96,19	45,34
TGI	98,88	97,45	95,32	90,01	78,47	54,99	25,55	83,25	25,79
S-TGI	98,91	97,91	96,32	91,74	79,77	55,30	26,20	84,21	27,24
VTS-0	98,86	98,02	96,62	92,12	78,55	49,36	21,38	82,93	25,31
VTS-1	98,90	98,16	97,12	94,17	86,47	65,66	33,42	88,32	33,44
MMSR	98,91	98,08	96,69	92,77	82,18	58,76	27,21	85,70	29,49

Tabla 6.16: Evaluación de la técnica de reconstrucción espectral MMSR en Aurora2 y comparativa con otras técnicas de compensación similares.

ya que asume independencia entre las distintas características de voz. Esto no supone una gran limitación ya que, como puede apreciarse para TGI, para modelos de voz con un número alto de componentes la diferencia entre ambos tipos de matrices es insignificante.

Como cabría esperar, la reconstrucción TGI con máscaras binarias oráculo (Oráculo), equivalente a MMSR con ruido oráculo, logra los mejores resultados de reconocimiento de entre las técnicas evaluadas. Este valor puede considerarse como la cota superior en el rendimiento de las técnicas derivadas del modelo de enmascaramiento de la voz, a saber, TGI, S-TGI y MMSR. De estas tres últimas técnicas la que mejor rendimiento ofrece es MMSR, seguida por S-TGI y, por último, TGI. Esto confirma la hipótesis que se planteó durante la evaluación de TGI justificando sus pobres resultados en situaciones realistas: a la luz de los resultados de la tabla 6.16 podemos aventurarnos a decir que el uso de máscaras binarias magnifica los errores que se producen durante la estimación del ruido. En el caso de las técnicas S-TGI y MMSR estos errores no afectan de forma tan directa, ya que se manejan descripciones probabilísticas de la fiabilidad y/o del ruido, en lugar de descripciones binarias.

En relación al estándar ETSI AFE, comprobamos que nuestra propuesta MMSR aún no está a la altura del rendimiento alcanzado por éste. No obstante, el excelente rendimiento obtenido por el sistema Oráculo y la mejora relativa lograda con MMSR en relación a TGI, hacen muy prometedora la aproximación propuesta. De esta forma, esperamos que mejoras en la estimación del ruido y la combinación de MMSR con otras técnicas orientadas a incrementar la robustez del reconocedor, salven la pequeña diferencia que separa MMSR del estándar avanzado.

Después del sistema oráculo, la técnica que mejores resultados proporciona es VTS-1, llegando a superar incluso al estándar ETSI AFE. En la tabla vemos que VTS-1

también sobrepasa los resultados de nuestra propuesta MMSR, siendo la diferencia más notable a SNRs bajas. Para explicar esta diferencia, consideraremos los dos errores que pueden afectar a ambas técnicas: el error de aproximación del modelo de distorsión de la voz y los errores que se producen al estimar el ruido. Con respecto al primer error, en la sección 2.2.2.2 se estudió que VTS emplea un desarrollo en serie para aproximar el modelo de distorsión mediante un polinomio de cierto grado. El error de aproximación dependerá, por tanto, del orden del polinomio elegido. En el caso de la técnica MMSR, el error que comete la aproximación *log-max* en la que se basa es desconocido a priori (aunque acotado) y dependerá de la relación entre las energías de la voz y el ruido (ver la figura 4.1). No obstante, estos errores de aproximación que afectan tanto a VTS como a MMSR pueden considerarse despreciables en comparación con el efecto de los errores de estimación del ruido en la compensación.

En relación a los errores en la estimación del ruido, los resultados de la tabla 6.16 parecen indicar que la aproximación VTS es más robusta que MMSR frente a ellos. Si observamos la tabla, constatamos que la mayor diferencia entre ambas técnicas se produce en los niveles bajos de SNR, niveles para los que la varianza del ruido es mayor y, por tanto, también lo es el error que se produce al intentar estimarlo. Para probar la hipótesis de que MMSR es más frágil que VTS ante estos errores, se diseñó una simulación sintética consistente en la estimación de datos unidimensionales distorsionados artificialmente con ruido aditivo de distinto grado. En dicha simulación supusimos que los valores de voz limpia x se modelan mediante una distribución normal $x \sim \mathcal{N}(\mu_x = 5, \sigma_x = 2)$, mientras que el ruido se modela mediante una segunda distribución $n \sim \mathcal{N}(\mu_n, \sigma_n = 1)$ con valores de media en el intervalo $\mu_n \in [-5, 10]$ a fin de evaluar distintos niveles de SNR. Estas distribuciones se emplearon entonces para generar secuencias de $T = 100000$ números pseudoaleatorios para cada distribución, que se combinaron posteriormente de acuerdo al modelo de distorsión de los parámetros de voz en el dominio log-Mel, a saber, $y_t = \log(e^{x_t} + e^{n_t})$. Finalmente, aplicamos cada técnica de compensación por separado a las secuencias de datos ruidosos $(y_1, \ldots, y_T)$ para estimar los valores de voz originales $(\hat{x}_1, \ldots, \hat{x}_T)$. A fin de evaluar el efecto de los errores en la estimación del ruido, durante el cómputo de $\hat{x}_t$ las técnicas VTS y MMSR emplearon versiones ruidosas $\hat{n}_t$ de los valores reales de ruido n_t. Estas valores ruidosos se obtuvieron añadiendo ruido blanco gaussiano a distintos niveles de potencia (de -10 dB a 50 dB) a los valores originales n_t.

La figura 6.6 muestra las curvas obtenidas para VTS y MMSR con los errores cuadráticos medios (MSE) de cada reconstrucción en función del error cometido en las estimaciones $\hat{n}_t$ (medido en términos de la SNR de $\hat{n}_t$ con respecto a los valores originales n_t). Estas gráficas confirman la hipótesis que habíamos planteado antes:

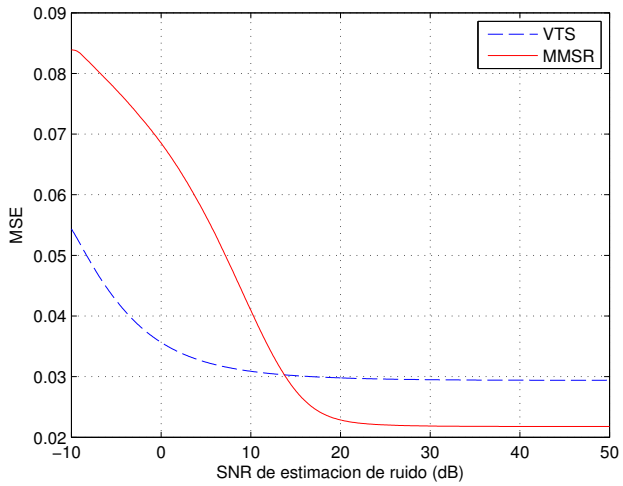

Figura 6.6: Evaluación de la robustez de las técnicas VTS y MMSR frente a errores en la estimación del ruido aditivo.

MMSR es más frágil que VTS a los errores que se pueden producir durante la estimación del ruido. Así, para niveles de SNR altos, el error cometido por MMSR es más pequeño que el de VTS, puesto que también lo será el error cometido al estimar la potencia del ruido. En cambio para niveles de SNR bajos, la varianza del ruido es mayor y también el error que se comete al estimarlo, por lo que MMSR sufrirá una pérdida de rendimiento mayor que VTS. Este pequeño experimento nos confirma el potencial de la técnica MMSR que ya nos indicaba la prueba oráculo. Esta potencialidad queda supeditada a poder obtener unas buenas estimaciones del ruido.

Continuamos con la evaluación de las técnicas propuestas, analizando ahora los resultados mostrados en la tabla 6.17 para Aurora4. De nuevo comprobamos que el sistema que mejor se comporta es el basado en VTS-1. No obstante, y a diferencia de Aurora2, esta vez nuestra propuesta MMSR se sitúa a la par que el estándar ETSI AFE (aunque MMSR es ahora mejor, la diferencia entre ambos sistemas es poco significativa). Esta mejora en el rendimiento de MMSR la podemos atribuir a que en Aurora4 la SNR promedio de las frases de *test* es mayor que en Aurora2, ya que para calcular el resultado de reconocimiento promedio también consideramos la condición limpia. Además, la diferencia entre VTS-1 y MMSR también se ha reducido.

La comparativa entre las tres técnicas basadas en el modelo de enmascaramiento, MMSR, TGI y S-TGI, revela el mismo orden de rendimiento que en el caso de Aurora2, sólo que ahora las diferencias son mayores: MMSR aventaja a TGI en un 6,77 % y a S-TGI en un 1,20 %. Como veíamos en la sección 4.3.4, aunque ambas técnicas MMSR y S-TGI son bastante parecidas entre sí, MMSE se basa en suposiciones más plausibles

Técnica	Test 01-07	Test 08-14	Media	M.R.
FE	59,70	49,84	54,77	–
AFE	72,31	64,63	68,47	25,01
Oráculo	84,97	74,54	79,75	45,61
TGI	65,16	59,01	62,09	13,36
S-TGI	71,36	63,95	67,66	23,52
VTS-0	66,84	62,43	64,63	18,00
VTS-1	73,24	66,70	69,97	27,75
MMSR	72,43	65,30	68,86	25,72

Tabla 6.17: Evaluación de la técnica de reconstrucción espectral MMSR en Aurora4 y comparativa con otras técnicas de compensación similares.

que S-TGI, lo que podría justificar la pequeña diferencia que existe entre los resultados de ambas técnicas.

Comparativa entre distintas máscaras de segregación. Habiendo evaluado el comportamiento de la técnica MMSR en distintas condiciones ruidosas, en este punto analizaremos otra de las cuestiones planteadas en el capítulo 4: el cómputo de máscaras de segregación continuas a partir del modelo de enmascaramiento de la voz. Como mencionamos en la sección 4.3.2, parte de los cálculos realizados por la técnicas MMSR pueden interpretarse como una máscara de segregación continua. En este punto evaluaremos el rendimiento de estas máscaras en comparación con las siguientes máscaras alternativas: máscaras binarias oráculo, máscaras binarias estimadas y máscaras continuas calculadas a partir de estimas locales de SNR tras aplicar un compresión sigmoidal [31]. Una vez calculadas, las máscaras serán empleadas por la técnica de reconstrucción oportuna (TGI en las máscaras binarias y S-TGI en las continuas) para obtener las estimas finales de voz usadas por el reconocedor.

El rendimiento de la reconstrucción espectral para las frases de Aurora2 usando las distintas máscaras de segregación se resume en la figura 6.7. En ésta apreciamos que el uso de máscaras oráculo supone una ganancia promedio de 10 dB respecto a reconocer con las máscaras estimadas (binarias o continuas). De entre las máscaras estimadas, la que mejor se comporta es la continua obtenida aplicando una compresión sigmoidal a las estimas de SNR. No obstante, la diferencia entre los tres tipos de máscaras no es elevada: el rendimiento promedio de las máscaras binarias es 83,25 %, para las continuas derivadas del modelo de enmascaramiento es 84,21 % y para las continuas sigmoidales es del 84,80 % de palabras reconocidas.

Los resultados obtenidos para Aurora4 se presentan en la figura 6.8. De nuevo la

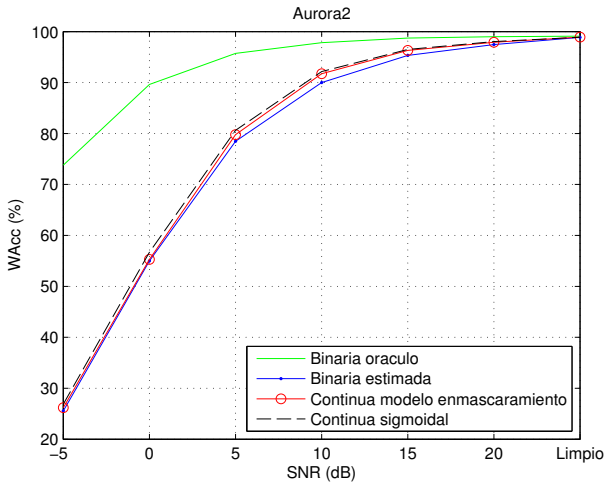

Figura 6.7: Evaluación del proceso de reconstrucción espectral con distintos tipos de máscaras de segregación en Aurora2.

máscaras que mejor se comportan de entre las estimadas son las sigmoidales, seguidas de cerca por las derivadas de la técnica MMSR (modelo de enmascaramiento) y, por último, las binarias. Para esta base de datos observamos que las diferencias entre las distintas máscaras es más significativa que en Aurora2: 62,09 % de WAcc obtenido usando máscaras binarias, 67,66 % con las continuas basadas en el modelo de enmascaramiento y 69,26 % en el caso de las continuas sigmoidales. Este último resultado, 69,26 %, es destacable, ya que se aproxima al rendimiento alcanzado por VTS-1 en esta base de datos (69,97 % según la tabla 6.17).

Para justificar el buen rendimiento alcanzado por las máscaras sigmoidales en comparación con las derivadas de la técnica MMSR, debemos considerar lo siguiente. En el cómputo de las máscaras sigmoidales interviene una función sigmoide cuyos parámetros deben ajustarse experimentalmente para cada base de datos. El ajuste de estos parámetros se realiza usando conjuntos de validación que, aunque diferentes de los conjuntos de *test*, contienen condiciones acústicas similares. Este ajuste permite que se compense en parte las deficiencias del estimador de ruido empleado. En cambio, las máscaras derivadas de la técnica MMSR no sufren ningún ajuste extra, sino que se derivan siguiendo un marco estadístico estricto en el que intervienen los modelos de voz y ruido. En este sentido cabe decir que un postprocesamiento de estas máscaras podría reportar las mismas ventajas vistas en el caso sigmoidal.

Estimación EM del modelo de ruido. Para concluir con la evaluación de las técnicas derivadas del modelo de enmascaramiento, en este punto analizaremos el ren-

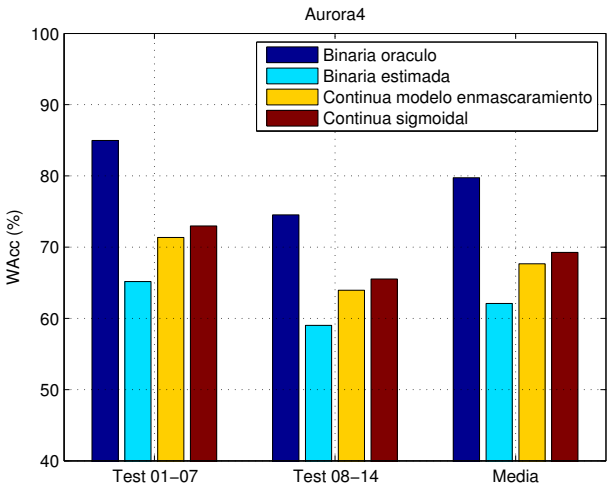

Figura 6.8: Evaluación del proceso de reconstrucción espectral con distintos tipos de máscaras de segregación en Aurora4.

dimiento del algoritmo iterativo no supervisado para la estimación del modelo de ruido propuesto en la sección 4.3.3. Si recordamos, este algoritmo ajustaba un GMM con M_n componentes (gaussianas con covarianzas diagonales) a cada frase de *test* para modelar el ruido aditivo presente en dicha frase. A fin de evaluar el rendimiento de dicho algoritmo, en este punto presentaremos las tasas de reconocimiento de palabras (WAcc) obtenidas por la técnica MMSR usando tanto los modelos de ruido estimados por este algoritmo, como las estimaciones simples de ruido (Simple) que hemos venido empleando hasta ahora. Además de la comparativa entre estas dos estrategias de estimación de ruido, en los resultados experimentales mediremos el rendimiento de la reconstrucción MMSR en función del número de componentes del GMM de ruido.

La tabla 6.18 muestra la comparativa entre los distintos métodos de estimación del modelo de ruido en la base de datos Aurora2. En primer lugar observamos que, como cabría esperar, el proceso de reconstrucción espectral se beneficia del uso de mejores modelos de ruido, aventajando en la mayoría de situaciones los GMMs a las estimaciones simples de ruido. Las únicas situaciones en las que la reconstrucción con GMMs de ruido sufre una merma con respecto a emplear estimaciones simples son, bien cuando el GMM cuenta con una sola componente (1 gauss), bien cuando el GMM cuenta con un número elevado de componentes (8 gauss). Los modelos de ruido con una única componente no son capaces de modelar con propiedad las características de los ruidos no estacionarios, es por ello que sufren una degradación con respecto a las estimas simples. Por otra parte, debemos considerar que la degradación que se produce al considerar GMMs con un gran número de componentes se debe a la falta

Modelo de ruido		Limpio	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Media	M.R.
Simple		98,91	98,08	96,69	92,77	82,18	58,76	27,21	85,70	–
GMM	1 gauss	99,11	98,20	96,72	92,53	81,73	57,35	25,00	85,31	-0,46
	2 gauss	99,10	98,16	96,83	93,26	83,32	59,34	26,89	86,18	0,57
	4 gauss	99,04	98,14	96,70	93,36	83,31	59,88	27,85	**86,28**	0,68
	6 gauss	99,08	98,24	96,69	93,23	82,90	58,62	27,35	85,94	0,28
	8 gauss	99,06	98,19	96,82	93,11	82,50	57,37	26,40	85,60	-0,11

Tabla 6.18: Resultados WAcc (%) obtenidos en Aurora2 por la técnica de reconstrucción espectral MMSR empleando estimaciones de ruido (Simple) o, de forma alternativa, GMMs de ruido estimados por el algoritmo EM propuesto.

de datos para entrenarlos de forma robusta. En este sentido creemos que el algoritmo EM propuesto se beneficiaría de un método automático para estimar el número de componentes del modelo en función de aspectos tales como la duración de la elocución, la estacionariedad del ruido, etc. Este trabajo queda fuera del objetivo de esta tesis y se plantea, por tanto, como trabajo futuro.

En la tabla 6.19 se muestran los resultados correspondientes a Aurora4. De nuevo la reconstrucción con GMMs de 4 gaussianas produce los mejores resultados de reconocimiento. Por contra, la reconstrucción usando GMMs con muy pocas componentes o, de forma alternativa, con un excesivo número de ellas, supone una degradación del rendimiento con respecto a emplear estimaciones simples del ruido. Un caso particular de los GMMs de ruido que merece la pena comentar por su parecido con las estimas simples de ruido son los que cuentan con 2 componentes. Recordemos que las estimaciones simples se calculan interpolando linealmente las medias parciales obtenidas a partir de los segmentos de ruido presentes al comienzo y final de cada frase. Así vemos que ambas estrategias (las estimas simples de ruido y los modelos de 2 componentes) estiman de forma explícita los parámetros de dos gaussianas para representar el ruido aditivo de cada frase. Aparte de esto, otro punto en común entre ambas estrategias es que estas gaussianas se estiman inicialmente a partir del ruido inicial y final de cada frase, si bien el algoritmo iterativo que proponemos refina posteriormente estos parámetros teniendo en cuenta la elocución al completo. A raíz de estos comentarios podemos evidenciar la similitud entre los resultados de reconocimiento alcanzados por ambos enfoques en ambas bases de datos.

Aunque la mejora relativa que hemos conseguido en ambas bases de datos por el uso de GMMs de ruido en lugar de ruido estimado no sea demasiado elevada, debe notarse que el algoritmo EM propuesto es susceptible de varias mejoras que elevarían su rendimiento. Además de la selección dinámica y automática del número de compo-

Modelo de ruido		Test 01-07	Test 08-14	Media	M.R.
Simple		72,43	65,30	68,86	–
GMM	1 gauss	71,60	65,37	68,49	-0,55
	2 gauss	72,37	65,62	68,99	0,19
	4 gauss	72,73	65,81	**69,27**	0,59
	6 gauss	72,43	65,46	68,95	0,13
	8 gauss	72,04	64,94	68,49	-0,54
	10 gauss	71,79	64,85	68,32	-0,79

Tabla 6.19: Resultados WAcc (%) obtenidos en Aurora4 por la técnica de reconstrucción espectral MMSR empleando estimaciones de ruido (Simple) o, de forma alternativa, GMMs de ruido estimados por el algoritmo EM propuesto.

nentes mencionado anteriormente, podemos, por ejemplo, extender este algoritmo para que también estime de forma no supervisada las componentes del ruido convolutivo. Asimismo, en lugar de emplear GMMs como modelos a priori de voz, el algoritmo también podría funcionar, introduciendo en éste los cambios oportunos, con los HMMs del reconocedor, incorporando con ello restricciones temporales adicionales en pos de un mejor ajuste del modelo de ruido. Esto también permitiría acotar la optimización de los parámetros del modelo de ruido si se dispone de las transcripciones (más o menos fidedignas) de cada elocución. Estas cuestiones serán tratadas más detenidamente en el capítulo 7.

6.3. Resumen

A lo largo de este capítulo hemos evaluado las técnicas de compensación propuestas en esta tesis usando un conjunto extensivo de experimentos de reconocimiento automático de voz en condiciones ruidosas. Los experimentos se han basado en dos tareas de reconocimiento: una de dígitos conectados (Aurora2) y otra de gran vocabulario (Aurora4). Ambas bases de datos contienen elocuciones en inglés de distinta duración y con ruidos incorporados de forma artificial a distintas SNRs.

La primera parte del capítulo ha estado dedicada a evaluar el rendimiento de las técnicas de compensación derivadas del estimador VQMMSE propuesto en el capítulo 3. A este respecto, los resultados obtenidos han demostrado la viabilidad de la compensación basada en datos estéreo siempre que la variabilidad acústica de los ruidos a los que se enfrenta el sistema de reconocimiento esté acotada. Los experimentos también han demostrado la importancia que cobra en este estimador el modelar con exactitud las transformaciones que el ruido produce en las características de voz. De entre las

técnicas de compensación derivadas de VQMMSE, los resultados experimentales nos indican que la que posee una capacidad mayor para reducir la degradación producida por el ruido es W-VQMMSE. Comparada con otras técnicas de compensación similares como SPLICE y MEMLIN, nuestra propuesta W-VQMMSE ofrece un rendimiento comparable a éstas, pero con una carga computacional mucho menor.

La segunda mitad del capítulo se ha orientado a medir el rendimiento de las técnicas de reconstrucción espectral basadas en el modelo de enmascaramiento de la voz, a saber, TGI y MMSR. Empezando por TGI, hemos demostrado que el conocimiento del patrón de enmascaramiento de la señal observada puede producir mejoras relativas de hasta el 45,89 % en Aurora2 y del 46,39 % en el caso de Aurora4. Estas mejoras, por supuesto, se tratan de resultados oráculo, pero nos sirven para constatar el amplio margen de ganancia del que disponemos. En la práctica, no obstante, las mejoras relativas que se obtienen son menores. Un análisis de los resultados alcanzados por TGI usando máscaras estimadas nos ha desvelado la fragilidad esta técnica a los errores producidos en la estimación del ruido.

Los resultados obtenidos por TGI sugieren una filosofía alternativa a la hora de abordar el problema de reconstrucción de espectros incompletos de voz. Basada en el mismo principio, la técnica MMSR difiere principalmente de TGI en que no requiere de una segmentación a priori del espectro observado en componentes fiables y no fiables, sino que trabaja con descripciones probabilísticas sobre la distribución del ruido. Esto le otorga a MMSR una mayor flexibilidad y robustez a errores en comparación con la técnica TGI. Así, el reconocimiento con los espectros de voz estimados por MMSR acarrea unas mejoras relativas (en tasas de palabras reconocidas) del 2,94 % en Aurora2 y del 10,90 % en Aurora4 frente a TGI. Lo que es más, nuestra propuesta se sitúa a la par que otras técnicas del estado del arte como el estándar ETSI AFE, siendo nuestra técnica conceptualmente más simple y más susceptible de mejora. En comparación con VTS, MMSR ha demostrado ser más sensible a los errores en la estimación del ruido. Esto ha conducido a remarcar la importancia que cobra en estas técnicas el disponer de estimas de ruido precisas.

Además de producir mejoras significativas en la reconstrucción de la voz ruidosa, la filosofía de emplear modelos de ruido en el proceso de reconstrucción también permite, por un lado, estimar máscaras de segregación continuas sin ningún esfuerzo añadido y, por otro, estimar iterativamente dichos modelos a partir de la propia frase a reconocer. Los resultados experimentales han demostrado que las máscaras continuas obtenidas de esta forma alcanzan mejoras relativas significativas respecto a las binarias empleadas por TGI: 1,15 % en Aurora2 y 11,55 % en Aurora4. La comparativa con otro tipo de máscaras continuas, no obstante, desvela que nuestras máscaras sufren algunas

carencias que se podrían mitigar con un postprocesamiento diseñado específicamente para corregirlas. En cuanto a la estimación iterativa de los GMMs de ruidos, los experimentos de reconocimiento han mostrado que estos mejoran levemente los resultados obtenidos usando estimaciones de ruido simples.

Para concluir, durante la experimentación de las distintas técnicas de compensación también se han evaluado otros aspectos accesorios a éstas, pero que han demostrado ser de gran importancia para su implantación en sistemas reales: nos referimos al modelado de las correlaciones temporales de la voz y a la explotación de la incertidumbre de la estimación en el reconocedor. Ambos aspectos han probado ser útiles de cara a robustecer los sistemas de reconocimiento frente al ruido.

CAPÍTULO 7

Conclusiones

Este trabajo surge de la necesidad de incrementar la robustez de los sistemas de reconocimiento automático del habla en condiciones de ruido. Para alcanzar esta meta, se han investigado diversos métodos orientados a mitigar el efecto del ruido sobre la representación de la voz. En el siguiente apartado se resumen las conclusiones más importantes extraídas a lo largo de esta tesis.

7.1. Conclusiones

- Se ha analizado el efecto del ruido (aditivo y convolutivo) en la representación de la voz usada por la mayoría de los sistemas de reconocimiento actuales. Desde el punto de vista estadístico, el ruido genera una discrepancia entre las distribuciones de probabilidad correspondientes a los datos de entrenamiento y a los de evaluación. Esta discrepancia implica que los modelos acústicos del reconocedor no modelan correctamente la voz a reconocer y, por tanto, la precisión del sistema de reconocimiento disminuye.

- Se ha llevado a cabo una revisión de las distintas estrategias propuestas para incrementar la robustez del reconocimiento de voz en condiciones de ruido.

- A pesar de que la estrategia óptima de reconocimiento con ruido conlleve el uso de modelos acústicos entrenados bajo la misma condición de evaluación, se ha discutido que esta estrategia es impracticable en la mayoría de situaciones. Por otra parte, la adaptación en tiempo real de los modelos para ajustarse a ruidos altamente no estacionarios, también puede ser inviable en ciertos sistemas por

el alto coste computacional ello que conllevaría. Por tanto, se ha argumentado que este problema requiere una solución más dinámica y eficiente que las dos anteriores. La aproximación de compensación de las características de voz es la estrategia adoptada en este trabajo.

- Como primer intento se ha propuesto un conjunto de técnicas de compensación que asumen la existencia de grabaciones estéreo con canales simultáneos de voz limpia y ruidosa. A partir de estas grabaciones, las técnicas propuestas derivan las transformaciones oportunas que se aplican posteriormente a la voz ruidosa para compensarla. Cuando estas técnicas son evaluadas en las mismas condiciones de ruido que las consideradas en el entrenamiento, los resultados de reconocimiento se sitúan ligeramente por debajo que los obtenidos usando modelos entrenados en las condiciones de evaluación. En presencia de ruidos desconocidos, por contra, el error de reconocimiento obtenido por estas técnicas es mucho mayor.

- Para contrarrestar esta merma que se produce frente a los ruidos desconocidos, se ha demostrado que las técnicas propuestas pueden usarse en combinación con otras técnicas de reconocimiento robusto con resultados satisfactorios. Por ejemplo, estas técnicas se pueden emplear como postprocesamiento de las características extraídas por un *front-end* robusto (p.ej. el ETSI AFE), o en combinación con métodos de modelado robusto como el entrenamiento multicondición. En ambos casos se obtiene una mejora significativa.

- Tomando como base el paradigma de datos perdidos, se ha propuesto una técnica de imputación denominada TGI para la estimación de las características ruidosas del espectro. En aquellas condiciones en las que sea posible conocer el patrón de enmascaramiento de la señal observada, los resultados de reconocimiento obtenidos por esta técnica son comparables a los obtenidos usando voz limpia. Sin embargo, el rendimiento de esta técnica disminuye cuando las máscaras de segregación son estimadas.

- Se ha demostrado que la principal fragilidad de TGI reside en las máscaras binarias que emplea. Cuando estas máscaras se calculan a partir de ciertas estimaciones de la potencia del ruido, los posibles errores de estimación del ruido se magnifican por la decisión binaria tomada. En estos casos se ha comprobado que decisiones más flexibles, como las que conllevan las máscaras continuas, produce estimadores más robustos

- Como respuesta a la fragilidad del uso de máscaras binarias en las técnicas MD, se ha derivado una técnica de reconstrucción alternativa denominada MMSR.

Esta técnica combina la simplicidad del modelo de enmascaramiento en el que se basan las técnicas MD con el uso de modelos probabilísticos del ruido, cuya robustez a los errores de estimación es mayor. Además de proporcionar mejores resultados de reconocimiento que TGI, la derivación de MMSR auna en uno sólo dos problemas tratados por separado por las técnicas MD: (i) estimación de las máscaras de segregación y (ii) estimación de las características ruidosas del espectro.

- Asimismo, se ha comprobado que parte de los cálculos que MMSR lleva a cabo pueden interpretarse alternativamente como una máscara de segregación continua. Frente a otras máscaras continuas, el método propuesto presenta varias ventajas. En primer lugar, el marco bayesiano del que se derivan las máscaras permite la explotación de otro tipo de información que pudiera ser útil de cara a segregar mejor el espectro (p.ej. la frecuencia fundamental del locutor). Por otro lado, el método propuesto no requiere el ajuste experimental de ninguno de sus parámetros.

- El algoritmo EM propuesto para la estimación del modelo de ruido ha mostrado ser eficiente en la caracterización de ruidos en las bases de datos evaluadas. En comparación con otras técnicas de estimación de ruido, el algoritmo permite una representación más fiel de las características de los ruidos no estacionarios.

- Se han analizado varias propuestas para la explotación de la redundancia temporal de la voz en las técnicas de compensación propuestas. Por un lado, se han investigado distintas estrategias para modelar las correlaciones temporales de tiempo corto. Por otro lado, se han derivado las expresiones oportunas para adaptar las técnicas propuestas a un modelado basado en HMMs. En todos los casos los resultados de reconocimiento obtenidos avalan al modelado HMM como el método más eficiente de los estudiados, ya que permite considerar toda la información recogida en la elocución durante el proceso de compensación.

- Finalmente, se han estudiado dos estrategias alternativas para el cómputo y explotación de medidas de incertidumbre extraídas del proceso de compensación: propagación de la varianza de la estimación MMSE (técnica *soft-data*) y algoritmo ponderado de Viterbi (WVA). Entre estas dos estrategias, experimentalmente se ha demostrado que el algoritmo WVA produce mejores resultados de reconocimiento que la técnica *soft-data*. Esta diferencia se ha justificado por la falta de precisión de algunas de las suposiciones en las que se fundamenta la aproximación *soft-data*.

7.2. Contribuciones

Las principales contribuciones de este trabajo pueden resumirse en:

- Desarrollo de un conjunto de técnicas de compensación de características basadas en estimación MMSE, modelado VQ de los espacios de características y uso de datos estéreo para derivar las transformaciones que se aplican a la voz ruidosa [122, 123, 124].
- Propuesta de un algoritmo de reconstrucción espectral para la estimación de las características distorsionadas en los espectrogramas de voz, supuesto que se conoce a priori la segregación del espectro observado en características fiables y distorsionadas [126, 127].
- Desarrollo de una técnica de compensación basada en un modelo analítico de distorsión/enmascaramiento de las características espectrales de voz [121, 125].
- Propuesta de un algoritmo para la estimación del grado de fiabilidad (enmascaramiento) de las características espectrales de voz en presencia de ruido aditivo [121, 125].
- Derivación de un método iterativo basado en el algoritmo EM para el ajuste de GMMs al ruido aditivo presente en señales de voz ruidosas.
- Estudio y aplicación de diferentes alternativas para la explotación de la redundancia temporal de la voz dentro de las técnicas de compensación [122, 127].
- Derivación de distintas medidas de incertidumbre sobre las estimas de voz realizadas y su tratamiento por parte del reconocedor de voz [122, 126].

7.3. Trabajo futuro

Una parte de las técnicas propuestas en esta tesis se han derivado del modelo de enmascaramiento de la voz expuesto en el capítulo 4. Aunque los resultados experimentales obtenidos han confirmado la alta precisión de este modelo, podemos vislumbrar varias líneas de investigación orientadas a incrementar la precisión y robustez del mismo. En este sentido, sería interesante investigar el modelado del error residual cometido por este modelo y su posterior explotación en las distintas técnicas derivadas del mismo. A este respecto, en la figura 4.1 se presentaban distintos histogramas calculados sobre Aurora2 con el error del modelo de enmascaramiento en función de la SNR de

la señal. En base a la información recogida en esta gráfica, parece razonable que la distribución de dicho error se puede aproximar mediante una PDF cuyos parámetros dependen de la SNR de la señal observada. Esta PDF sería posteriormente empleada por las distintas técnicas de reconstrucción propuestas para realizar una estimación más precisa de las características ruidosas.

Además del punto anterior, otra cuestión a tratar en el futuro es establecer los lazos de unión que comparten el modelo de enmascaramiento estudiado y el enmascaramiento que se produce en el oído humano. Creemos que el modelo de enmascaramiento es una simplificación de este último y, desde este punto de vista, se podrían aprovechar los avances realizados en el campo de la audición humana. Aspectos clave que deberían ser analizados incluyen el proceso de enmascaramiento temporal y el uso de distintos umbrales de enmascaramiento en función de la frecuencia.

En relación a las técnicas de compensación propuestas, un aspecto que podría mejorar la precisión de la estimación MMSE sería el uso de modelos de voz dependientes del locutor, frente a los independientes del locutor empleados hasta ahora. Si bien es cierto que la identidad del locutor suele ser desconocida, sí que sería posible estimar un conjunto de parámetros que modelasen las características propias de cada locutor. Por dar unos ejemplos, se podrían estimar la longitud del tracto vocal o una matriz de transformación que maximizase la verosimilitud de los datos observados, tal y como se hace en MLLR. Estas parámetros dependientes del locutor se emplearían posteriormente para adaptar los modelos de voz al locutor y, de esta forma, obtener unas estimaciones más precisas.

Otra cuestión que habría que investigar es el uso de matrices de covarianza completas (no diagonales) en los modelos de fuente (voz o ruido) empleados por la técnica MMSR. El uso de este tipo de matrices permitiría explotar mejor las correlaciones entre las distintas características de cara a la estimación de aquellas que se encuentren perdidas. Sin embargo, este uso se rechazó en su momento porque conduce a la evaluación de todas las segregaciones voz/ruido posibles para cada vector observado, lo que resulta impracticable en la mayoría de casos. No obstante, la gran mayoría de estos casos pueden descartarse en la práctica por ser muy poco probables. Así, se podría diseñar una estrategia de poda que emplease información de bajo nivel como la relación armónica entre frecuencias, *onset*/*offset* común entre frecuencias, posición espacial de la que proviene cada fuente (en caso de disponer de varios micrófonos), etc., para descartar las segregaciones poco probables y evaluar únicamente las posibles.

El algoritmo EM para la estimación del modelo de ruido también es susceptible de diversas mejoras. Quizás la más inmediata sea la extensión del mismo para que también estime las características del ruido convolutivo. Aparte de esto, tenemos la cuestión de

cómo estimar automáticamente el número de componentes del GMM de ruido. Como se ha comentado varias veces, esta cuestión es importante, ya que permitiría un balanceo entre precisión del modelo de ruido obtenido y número de datos requeridos para estimar sus parámetros robustamente.

Por último, de las dos técnicas estudiadas para el tratamiento de la incertidumbre de las estimas, la que mejor resultados ha obtenido es el algoritmo ponderado de Viterbi. En los experimentos que se han presentado, este algoritmo se ha evaluado usando un único peso por vector de características. En este sentido también sería interesante explorar las posibilidades que ofrece el uso de un peso por característica del vector, de manera que durante el reconocimiento se tenga un mayor control sobre el cómputo de las probabilidades de observación.

APÉNDICE A

Eficiencia de las técnicas de compensación basadas en datos estéreo

En este apéndice simplificaremos las expresiones asociadas con varias técnicas de estimación MMSE que emplean datos estéreo. En concreto, nuestro estudio se centrará en las técnicas de estimación VQMMSE estudiadas en la sección 3.2, así como SPLICE [67] y MEMLIN [44]. Estos dos últimos métodos son dos de los representantes mejor conocidos en la literatura de estimaciones MMSE basadas en grabaciones estéreo y que usan GMMs para el modelado de los espacios de características. A fin de ofrecer una expresión computacionalmente simplificada de cada estimador, tendremos en cuenta que algunos de los términos que aparecen en las mismas pueden ser agrupados y precalculados durante la fase de entrenamiento, con el consiguiente ahorro computacional. Basándonos en las expresiones simplificadas obtenidas, posteriormente se mostrará la complejidad asintótica en tiempo de cada técnica. Esta se expresará mediante la notación $O(\boldsymbol{f}(\boldsymbol{y}))$, ya que nos proporciona una cota superior para el tiempo requerido por el estimador $\hat{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{y})$ y, de esta forma, podemos hacernos una idea de la complejidad de cada estimador [18].

A la hora de derivar la complejidad asintótica de cada técnica, obviaremos el tiempo relativo al cómputo de las probabilidades a posteriori de cada región del espacio de características. Esto se debe a que todas las técnicas de compensación mencionadas anteriormente calculan, para cada vector observado $\boldsymbol{y}$, la probabilidad de observación del mismo para cada una de las regiones $k_y = 1, \ldots, M_y$ del modelo. En el caso de

las técnicas SPLICE y MEMLIN, estas probabilidades involucran la evaluación de M_y funciones de densidad normales multivariantes $p(\boldsymbol{y}|k_y)$. Por su parte, las técnicas derivadas del estimador VQMMSE emplean la distancia de Mahalanobis $\mathrm{d}(\boldsymbol{y}, k_y)$ definida en la ecuación (3.16) para este propósito. El coste computacional de ambas expresiones es muy similar, ya que tanto $p(\boldsymbol{y}|k_y)$ como $\mathrm{d}(\boldsymbol{y}, k_y)$ incluyen formas cuadráticas que son las que conllevan la mayor parte del cómputo.

A.1. SPLICE

- Expresión simplificada:

$$\hat{\boldsymbol{x}} = \boldsymbol{y} - \sum_{k_y=1}^{M_y} P(k_y|\boldsymbol{y})\boldsymbol{r}_{k_y}, \tag{A.1}$$

  donde $\boldsymbol{r}_{k_y}$ es el vector de corrección asociado con la componente k_y-ésima del GMM. Este vector se calcula en la fase de entrenamiento usando grabaciones estéreo (ver sección 2.2.3.4).

- Complejidad asintótica: $O(M_y D)$ siendo M_y el número de componentes del GMM que modela el espacio de características distorsionadas y Dla dimensión del vector $\boldsymbol{y}$.

A.2. MEMLIN

- Expresión simplificada:

$$\begin{aligned}\hat{\boldsymbol{x}} &= \boldsymbol{y} - \sum_{k_y}^{M_y} \underbrace{\sum_{k_x}^{M_x} \boldsymbol{r}_{k_x k_y} P(k_x|k_y)}_{\boldsymbol{r}_{k_y}} P(k_y|\boldsymbol{y}) \\ &= \boldsymbol{y} - \sum_{k_y}^{M_y} P(k_y|\boldsymbol{y})\boldsymbol{r}_{k_y}.\end{aligned} \tag{A.2}$$

  $\boldsymbol{r}_{k_x k_y}$ es el vector de corrección asociado al par de gaussianas (k_x, k_y). De nuevo este vector se obtiene durante la fase de entrenamiento a partir de datos estéreo (ver ecuación (2.134)).

- Complejidad asintótica: $O(M_y D)$.

A.3. Q-VQMMSE

- Expresión simplificada:

$$\hat{\boldsymbol{x}} = \underbrace{\sum_{k_x=1}^{M_x} P(k_x|k_y^*)\boldsymbol{\mu}_x^{(k_x)}}_{\boldsymbol{\mu}_x^{(k_y^*)}} = \boldsymbol{\mu}_x^{(k_y^*)}, \tag{A.3}$$

  siendo $\boldsymbol{\mu}_x^{(k_x)}$ el centroide de la celda k_x-ésima del espacio limpio y k_y^* la celda del espacio distorsionado a la que $\boldsymbol{y}$ pertenece.

- Complejidad asintótica: $O(1)$.

A.4. S-VQMMSE

- Expresión simplificada:

$$\hat{\boldsymbol{x}} = \boldsymbol{y} - \underbrace{\sum_{k_x=1}^{M_x} P(k_x|k_y^*)\left(\boldsymbol{\mu}_x^{(k_x)} - \boldsymbol{\mu}_y^{(k_y^*)}\right)}_{\boldsymbol{r}^{(k_y^*)}} = \boldsymbol{y} - \boldsymbol{r}^{(k_y^*)}. \tag{A.4}$$

- Complejidad asintótica: $O(D)$.

A.5. J-VQMMSE

- Expresión simplificada:

$$\begin{aligned}\hat{\boldsymbol{x}} &= \sum_{k_x=1}^{M_x} P(k_x|k_y^*)\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}, k_x, k_y^*] = \sum_{k_x=1}^{M_x} P(k_x|k_y^*)\left(\boldsymbol{A}^{(k_x,k_y^*)}\boldsymbol{y} + \boldsymbol{b}^{(k_x,k_y^*)}\right) \\ &= \underbrace{\left(\sum_{k_x=1}^{M_x} P(k_x|k_y^*)\boldsymbol{A}^{(k_x,k_y^*)}\right)}_{\boldsymbol{A}^{(k_y^*)}}\boldsymbol{y} + \underbrace{\sum_{k_x=1}^{M_x} P(k_x|k_y^*)\boldsymbol{b}^{(k_x,k_y^*)}}_{\boldsymbol{b}^{(k_y^*)}} \\ &= \boldsymbol{A}^{(k_y^*)}\boldsymbol{y} + \boldsymbol{b}^{(k_y^*)},\end{aligned} \tag{A.5}$$

  donde

$$\boldsymbol{A}^{(k_x,k_y)} = \boldsymbol{\Sigma}_{xy}^{(k_x,k_y)}{\boldsymbol{\Sigma}_y^{(k_x,k_y)}}^{-1}, \tag{A.6}$$

$$\boldsymbol{b}^{(k_x,k_y)} = \boldsymbol{\mu}_x^{(k_x,k_y)} - \boldsymbol{\Sigma}_{xy}^{(k_x,k_y)}{\boldsymbol{\Sigma}_y^{(k_x,k_y)}}^{-1}\boldsymbol{\mu}_y^{(k_x,k_y)}. \tag{A.7}$$

En la ecuaciones anteriores $\boldsymbol{\mu}_x^{(k_x,k_y)}$ y $\boldsymbol{\mu}_y^{(k_x,k_y)}$ son las medias de las subregiones correspondientes de los espacios de características limpias y distorsionadas, respectivamente. De forma análoga, $\boldsymbol{\Sigma}_y^{(k_x,k_y)}$ en la matriz de covarianza de la subregión del espacio distorsionado, mientras que $\boldsymbol{\Sigma}_{xy}^{(k_x,k_y)}$ es la matriz de correlación cruzada. Todos estos términos se pueden calcular fácilmente durante la fase de entrenamiento usando grabaciones estéreo (ver sección 3.2.3).

- Complejidad asintótica: la complejidad de la técnica J-VQMMSE dependerá de la forma de la matrices $\boldsymbol{A}^{(k_x,k_y)}$ que aparecen en la ecuación (A.5). Consideraremos los siguientes tres casos:

  - Matriz identidad: la expresión simplificada es,

$$\hat{\boldsymbol{x}} = \boldsymbol{y} - \underbrace{\left(\boldsymbol{\mu}_y^{(k_x,k_y^*)} - \boldsymbol{\mu}_x^{(k_x,k_y^*)}\right)}_{\boldsymbol{r}^{(k_x,k_y^*)}}, \tag{A.8}$$

    y la complejidad es $O(D)$.

  - Matriz diagonal: definimos

$$\boldsymbol{a}^{(k_x,k_y)} = \operatorname{diag}\left(\boldsymbol{A}^{(k_x,k_y)}\right), \tag{A.9}$$

$$\boldsymbol{b}^{(k_x,k_y)} = \boldsymbol{\mu}_x^{(k_x,k_y)} - \boldsymbol{a}^{(k_x,k_y)} \circ \boldsymbol{\mu}_y^{(k_x,k_y)}, \tag{A.10}$$

    donde $\operatorname{diag}(\boldsymbol{A})$ es el vector con los elementos de la diagonal principal de $\boldsymbol{A}$ y el operador $\circ$ denota el producto vectorial por componentes. Entonces la expresión simplificada viene dada por

$$\hat{\boldsymbol{x}} = \boldsymbol{a}^{(k_y^*)} \circ \boldsymbol{y} + \boldsymbol{b}^{(k_y^*)}, \tag{A.11}$$

    y la complejidad asintótica es $O(D)$.

  - Matriz completa (no diagonal): la forma simplificada es la que aparece en (A.5) y la complejidad es $O(D^2)$.

A.6. W-VQMMSE

- Expresión simplificada:

$$\begin{aligned}\hat{\boldsymbol{x}} &= \sum_{k_x=1}^{M_x} P(k_x|k_y^*)\left(\boldsymbol{D}^{(k_x,k_y^*)}\boldsymbol{y}+\boldsymbol{e}^{(k_x,k_y^*)}\right)\\ &= \underbrace{\left(\sum_{k_x=1}^{M_x} P(k_x|k_y^*)\boldsymbol{D}^{(k_x,k_y^*)}\right)}_{\boldsymbol{D}^{(k_y^*)}}\boldsymbol{y} + \underbrace{\sum_{k_x=1}^{M_x} P(k_x|k_y^*)\boldsymbol{e}^{(k_x,k_y^*)}}_{\boldsymbol{e}^{(k_y^*)}}\\ &= \boldsymbol{D}^{(k_y^*)}\boldsymbol{y}+\boldsymbol{e}^{(k_y^*)},\end{aligned} \tag{A.12}$$

  donde

$$\boldsymbol{D}^{(k_x,k_y)} = \left(\boldsymbol{\Sigma}_x^{(k_x,k_y)}\right)^{1/2}\left(\boldsymbol{\Sigma}_y^{(k_x,k_y)}\right)^{-1/2}, \tag{A.13}$$

$$\boldsymbol{e}^{(k_x,k_y)} = \boldsymbol{\mu}_x^{(k_x,k_y)} - \left(\boldsymbol{\Sigma}_x^{(k_x,k_y)}\right)^{1/2}\left(\boldsymbol{\Sigma}_y^{(k_x,k_y)}\right)^{-1/2}\boldsymbol{\mu}_y^{(k_x,k_y)}. \tag{A.14}$$

- Complejidad asintótica: al igual que la técnica J-VQMMSE, la complejidad de esta técnica también dependerá de la forma de la matrices $\boldsymbol{D}^{(k_x,k_y)}$ que aparecen en la ecuación (A.12). De nuevo consideramos los siguientes casos:

  - Matriz identidad: la expresión simplificada coincide con (A.8) y la complejidad es $O(D)$.
  - Matriz diagonal: definimos

$$\boldsymbol{d}^{(k_x,k_y)} = \boldsymbol{\sigma}_x^{(k_x,k_y)} \circ \boldsymbol{\sigma}_y^{(k_x,k_y)^{-1}}, \tag{A.15}$$

$$\boldsymbol{e}^{(k_x,k_y)} = \boldsymbol{\mu}_x^{(k_x,k_y)} - \boldsymbol{d}^{(k_x,k_y)} \circ \boldsymbol{\mu}_y^{(k_x,k_y)}, \tag{A.16}$$

    entonces la expresión simplificada viene dada por,

$$\hat{\boldsymbol{x}} = \boldsymbol{d}^{(k_y^*)} \circ \boldsymbol{y} + \boldsymbol{e}^{(k_y^*)}, \tag{A.17}$$

    y la complejidad asintótica es $O(D)$.
  - Matriz completa: la forma simplificada es la que aparece en (A.12) y la complejidad asintótica es $O(D^2)$.

APÉNDICE B

Distribución normal truncada

Este apéndice contiene una breve revisión de las propiedades de la distribución normal truncada. Una distribución normal truncada es una distribución de probabilidad de variable continua que se distribuye según una normal, pero cuyos valores están acotados superiormente, inferiormente o por ambos lados. En el capítulo 4 esta distribución de probabilidad es muy utilizada, ya que permite incorporar de forma elegante las restricciones impuestas por el modelo de enmascaramiento de la ecuación (4.4)[1] en el estimador MMSE de la voz basado en GMMs.

B.1. Definición formal

Sea Z una variable aleatoria que sigue una distribución normal (unidimensional) de parámetros μ y σ, $Z \sim \mathcal{N}(\mu, \sigma)$. Notemos además por X al subconjunto de Z que toma valores dentro del intervalo $[a, b]$. Entonces decimos que X se distribuye según una distribución normal truncada y su función de densidad de probabilidad viene dada por [149]:

$$p(x|a \leq x \leq b, \mu, \sigma) = \begin{cases} \gamma_{a,b}\, \mathcal{N}(x; \mu, \sigma) & \text{si } \ a \leq x \leq b \\ 0 & \text{en otro caso} \end{cases}, \tag{B.1}$$

donde $\gamma_{a,b}$ es un factor de normalización que hace que la PDF anterior integre a uno:

$$\gamma_{a,b} = \frac{1}{\Phi(b; \mu, \sigma) - \Phi(a; \mu, \sigma)}, \tag{B.2}$$

[1] Esta restricción se refiere a que para cierto elemento y del espectro log-Mel, la energía de la voz x en dicho elemento estará acotada superiormente por la observación, esto es, $x \leq y$.

donde Φ denota la función de distribución acumulada.

Un caso particular de la distribución normal truncada es aquél en el que la distribución se encuentra acotada superiormente (acotada por la derecha), esto es, $X \in (-\infty, b]$. Merece la pena que estudiemos con detenimiento este caso, ya que se nos presenta durante la estimación de las características de la voz en las técnicas derivadas del modelo de enmascaramiento de la ecuación (4.4). En concreto nos encontramos con que el valor de la voz a estimar, x, se encuentra acotado superiormente por la observación $b = y$ e inferiormente por $-\infty$ (asumimos que los valores de energía se expresan en el dominio log-Mel). En tal caso, la probabilidad de x viene dada por

$$p(x|x \leq y, \mu, \sigma) = \begin{cases} \dfrac{\mathcal{N}(x;\mu,\sigma)}{\Phi(y;\mu,\sigma)} & \text{si } \ x \leq y \\ 0 & \text{en otro caso} \end{cases} . \tag{B.3}$$

B.2. Momentos de la distribución

En este apartado presentamos de forma resumida las expresiones que permiten calcular los dos primeros momentos de la distribución normal truncada, así como su varianza. La derivación detallada de estas expresiones, así como las expresiones correspondientes a otros momentos de orden superior, pueden encontrarse en [74, 149].

Empezamos con el estudio de la distribución truncada general (acotada por ambos lados), que luego particularizaremos a nuestra distribución de interés (acotada a la derecha). Definimos, en primer lugar, α y β como las variables tipificadas correspondientes a los límites a y b, respectivamente,

$$\alpha = \frac{a - \mu}{\sigma}, \tag{B.4}$$

$$\beta = \frac{b - \mu}{\sigma}. \tag{B.5}$$

Asimismo, notaremos mediante $\rho(\alpha, \beta)$ al siguiente cociente de probabilidades, el cual aparece repetidas veces en las expresiones de los momentos:

$$\rho(\alpha, \beta) = \frac{\mathcal{N}(\alpha) - \mathcal{N}(\beta)}{\Phi(\beta) - \Phi(\alpha)}, \tag{B.6}$$

siendo las distribuciones anteriores de media cero y varianza unitaria, es decir, distribuciones normales estándar.

A partir las variables anteriores podemos definir los dos primeros momentos de la distribución normal truncada como [74, 149]:

$$\mathbb{E}[x|a \leq x \leq b, \mu, \sigma] = \mu + \sigma\rho(\alpha, \beta), \tag{B.7}$$

$$\mathbb{E}[x^2|a \leq x \leq b, \mu, \sigma] = \mu^2 + 2\mu\sigma\rho(\alpha, \beta) + \sigma^2 \left(1 + \frac{\alpha\mathcal{N}(\alpha) - \beta\mathcal{N}(\beta)}{\Phi(\beta) - \Phi(\alpha)}\right), \tag{B.8}$$

y la varianza se puede calcular como

$$\begin{aligned}
&\mathrm{Var}[x|a \leq x \leq b, \mu, \sigma] \\
&\quad = \mathbb{E}[x^2|a \leq x \leq b, \mu, \sigma] - \mathbb{E}[x|a \leq x \leq b, \mu, \sigma]^2 \\
&\quad = \underbrace{\left[\mu^2 + 2\mu\sigma\rho(\alpha, \beta) + \sigma^2 \left(1 + \frac{\alpha\mathcal{N}(\alpha) - \beta\mathcal{N}(\beta)}{\Phi(\beta) - \Phi(\alpha)}\right)\right]}_{\mathbb{E}[x^2|a \leq x \leq b, \mu, \sigma]} - \underbrace{\left[\mu^2 + \sigma^2\rho(\alpha, \beta)^2 + 2\mu\sigma\rho(\alpha, \beta)\right]}_{\mathbb{E}[x|a \leq x \leq b, \mu, \sigma]^2} \\
&\quad = \sigma^2 \left(1 + \frac{\alpha\mathcal{N}(\alpha) - \beta\mathcal{N}(\beta)}{\Phi(\beta) - \Phi(\alpha)} - \rho(\alpha, \beta)^2\right).
\end{aligned} \tag{B.9}$$

Estudiamos, en segundo lugar, el caso de la distribución truncada a la derecha. En este caso al valor tipificado de la observación lo notamos mediante $\overline{y}$, es decir,

$$\overline{y} = \frac{y - \mu}{\sigma}, \tag{B.10}$$

y el cociente de la ecuación (B.6) se corresponde con

$$\rho(\overline{y}) = -\frac{\mathcal{N}(\overline{y})}{\Phi(\rho)}. \tag{B.11}$$

Luego los dos primeros momentos de la distribución son[1],

$$\mathbb{E}[x|x \leq y, \mu, \sigma] = \frac{\int_{-\infty}^{y} x\mathcal{N}(x; \mu, \sigma)dx}{\Phi(\overline{y})} = \mu + \sigma\rho(\overline{y}), \tag{B.12}$$

$$\mathbb{E}[x^2|x \leq y, \mu, \sigma] = \frac{\int_{-\infty}^{y} x^2\mathcal{N}(x; \mu, \sigma)dx}{\Phi(\overline{y})} = \mu^2 + 2\mu\sigma\rho(\overline{y}) + \sigma^2 \left[1 + \overline{y}\rho(\overline{y})\right], \tag{B.13}$$

y la expresión para el cálculo de la varianza es

$$\begin{aligned}
&\mathrm{Var}[x|x \leq y, \mu, \sigma] \\
&\quad = \frac{\int_{-\infty}^{y} (x - \mu)^2\mathcal{N}(x; \mu, \sigma)dx}{\Phi(\overline{y})} \\
&\quad = \mathbb{E}[x^2|x \leq y, \mu, \sigma] - \mathbb{E}[x|x \leq y, \mu, \sigma]^2 = \sigma^2 \left[1 + \overline{y}\rho(\overline{y}) - \rho(\overline{y})^2\right].
\end{aligned} \tag{B.14}$$

[1]Notar que $\Phi(\overline{y}) = \Phi(y; \mu, \sigma)$

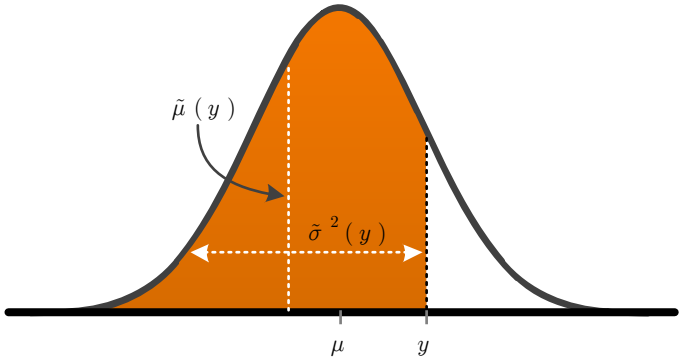

Figura B.1: Media y varianza de una distribución normal truncada.

Finalmente en la figura B.1 se muestra un esquema gráfico de la media y la varianza de la distribución truncada a la derecha. Con objeto de simplificar la figura, hemos notado por $\tilde{\mu}(y)$ a la media de la distribución (ecuación (B.12)) y por $\tilde{\sigma}^2(y)$ a su varianza (ecuación (B.14)).

APÉNDICE C

Algoritmo EM para el ajuste del modelo de ruido

A lo largo de este apéndice procederemos a derivar las ecuaciones necesarias para el cálculo de los parámetros del modelo de ruido que emplea la técnica MMSR descrita en la sección 4.3. El tipo de modelo que aquí consideramos es un GMM con M_n gaussianas:

$$p(\boldsymbol{n}|\mathcal{M}_n) = \sum_{k_n=1}^{M_n} P(k_n|\mathcal{M}_n)p(\boldsymbol{n}|k_n,\mathcal{M}_n) = \sum_{k_n=1}^{M_n} \pi_n^{(k_n)} \mathcal{N}\left(\boldsymbol{n}; \boldsymbol{\mu}_n^{(k_n)}, \boldsymbol{\Sigma}_n^{(k_n)}\right), \tag{C.1}$$

donde $\langle \pi_n^{(k_n)}, \boldsymbol{\mu}_n^{(k_n)}, \boldsymbol{\Sigma}_n^{(k_n)} \rangle$ son los parámetros asociados con cada componente del GMM, a saber, su probabilidad a priori, la media y la matriz de covarianza (diagonal).

Para calcular los parámetros de este modelo, en primer lugar es necesario definir el criterio de optimización que se va a seguir durante el ajuste del GMM. Como suele ser habitual en la literatura, el criterio que adoptamos en esta tesis es el de máxima verosimilitud:

$$\hat{\mathcal{M}}_n = \underset{\mathcal{M}_n}{\operatorname{argmax}}\, p(\boldsymbol{Y}|\mathcal{M}_n, \mathcal{M}_x), \tag{C.2}$$

siendo $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ la secuencia de datos observados y $\mathcal{M}_x$ el GMM que modela la voz.

La optimización directa de la ecuación anterior se hace inviable por la aparición de las variables ocultas k_x y k_n durante el desarrollo de la misma. Por tanto, en este trabajo se adopta una estrategia de optimización iterativa basada en el algoritmo EM [66]. Tal y como se indicó en la sección 4.3.3, la función auxiliar $\mathcal{Q}(\mathcal{M}_n, \hat{\mathcal{M}}_n)$ que este

algoritmo optimiza viene dada por

$$\mathcal{Q}(\mathcal{M}_n, \hat{\mathcal{M}}_n) = \sum_{t=1}^{T} \sum_{k_x=1}^{M_x} \sum_{k_n=1}^{M_n} \gamma_t^{(k_x,k_n)} \left[\sum_{i=1}^{D} \log p(y_{t,i}|k_x, k_n, \hat{\mathcal{M}}_n, \mathcal{M}_x) + \log \hat{\pi}_n^{(k_n)} \right], \quad \text{(C.3)}$$

donde $\mathcal{M}_n$ denota la hipótesis actual para el modelo de ruido, $\hat{\mathcal{M}}_n$ son los parámetros del modelo que se pretenden optimizar y $\gamma_t^{(k_x,k_n)} = P(k_x, k_n|\boldsymbol{y}_t, \mathcal{M}_n, \mathcal{M}_x)$ es la probabilidad a posteriori del par de gaussianas $\langle k_x, k_n \rangle$ dada la observación $\boldsymbol{y}_t$, el GMM de voz limpia y la hipótesis actual del GMM de ruido. Esta probabilidad se calcula usando la ecuación (4.44).

Para calcular los parámetros del modelo $\hat{\mathcal{M}}_n$, simplemente derivamos la función de la ecuación (C.3) respecto al parámetro deseado, igualamos a cero la expresión resultante y, por último, despejamos el parámetro en cuestión hasta obtener una expresión definitiva que nos permita actualizar su valor. Durante la derivación de estas expresiones veremos que con frecuencia aparecen términos relativos a la probabilidad acumulada de las gaussianas del modelo de ruido. Como ya hemos comentado varias veces a lo largo de esta memoria, no existe una solución analítica para el cálculo de estas CDFs si las gaussianas en cuestión cuentan con matrices de covarianza no diagonales. A fin de salvar este problema, se asumirá que los elementos de los vectores de características son independientes entre sí y, por tanto, pueden modelarse usando GMMs con matrices de covarianza diagonales. Aunque la suposición de independencia estadística no sea totalmente cierta cuando se aplica a las características de la voz en el dominio del banco de filtros, el uso de un elevado número de gaussianas con matrices de covarianza diagonal ofrece, desde el punto de vista estadístico, la misma potencia de modelado que las PDFs con covarianzas completas [27, 57].

En los siguientes apartados se procederá a derivar por separado las fórmulas de estimación de las medias, las varianzas y los pesos de las gaussianas del GMM.

C.1. Ajuste de las medias del modelo

La derivada parcial de la función auxiliar $\mathcal{Q}(\mathcal{M}_n, \hat{\mathcal{M}}_n)$ que aparece en la ecuación (C.3) con respecto a las medias del modelo $\hat{\mathcal{M}}_n$, $\hat{\mu}_{n,i}^{(k_n)}$ $(i = 1, \ldots, D; k_n = 1, \ldots, M_n)$,

viene dada por

$$\begin{aligned}
&\frac{\partial \mathcal{Q}(\mathcal{M}_n, \hat{\mathcal{M}}_n)}{\partial \hat{\mu}_{n,i}^{(k_n)}} \\
&= \sum_{t=1}^{T} \sum_{k_x=1}^{M_x} \gamma_t^{(k_x,k_n)} \frac{\partial \log p(y_{t,i}|k_x, k_n)}{\partial \hat{\mu}_{n,i}^{(k_n)}} \\
&= \sum_{t=1}^{T} \sum_{k_x=1}^{M_x} \gamma_t^{(k_x,k_n)} \underbrace{\frac{1}{p(y_{t,i}|k_x, k_n)} \left[p(y_{t,i}|k_x) \frac{\partial \Phi(y_{t,i}|k_n)}{\partial \hat{\mu}_{n,i}^{(k_n)}} + \Phi(y_{t,i}|k_x) \frac{\partial p(y_{t,i}|k_n)}{\partial \hat{\mu}_{n,i}^{(k_n)}} \right]}_{\alpha},
\end{aligned} \tag{C.4}$$

donde, por un lado, se ha omitido cualquier mención explícita a los modelos $\mathcal{M}_x$ y $\hat{\mathcal{M}}_n$ a fin de facilitar la lectura de la misma y, por otro lado, la probabilidad $p(y_{t,i}|k_x, k_n) \equiv p(y_{t,i}|k_x, k_n, \hat{\mathcal{M}}_n, \mathcal{M}_x)$ se ha sustituido por la solución obtenida en la ecuación (4.55).

Usando la siguiente identidad (ver p.ej. [216])

$$\frac{\partial p(x)}{\partial \mu} = p(x) \frac{(x-\mu)}{\sigma^2}, \tag{C.5}$$

tenemos que la derivada de $p(y_{t,i}|k_n)$ respecto a $\hat{\mu}_{n,i}^{(k_n)}$ es

$$\frac{\partial p(y_{t,i}|k_n)}{\partial \hat{\mu}_{n,i}^{(k_n)}} = p(y_{t,i}|k_n) \frac{\left(y_{t,i} - \hat{\mu}_{n,i}^{(k_n)}\right)}{\hat{\sigma}_{n,i}^{(k_n)^2}}, \tag{C.6}$$

y la derivada parcial de $\Phi(y_{t,i}|k_n)$ viene dada por

$$\begin{aligned}
\frac{\partial \Phi(y_{t,i}|k_n)}{\partial \hat{\mu}_{n,i}^{(k_n)}} &= \int_{-\infty}^{y_{t,i}} \frac{\partial p(n_i|k_n)}{\partial \hat{\mu}_{n,i}^{(k_n)}} dn_i = \int_{-\infty}^{y_{t,i}} \left[p(n_i|k_n) \frac{\left(n_i - \hat{\mu}_{n,i}^{(k_n)}\right)}{\hat{\sigma}_{n,i}^{(k_n)^2}} \right] dn_i \\
&= \frac{1}{\hat{\sigma}_{n,i}^{(k_n)^2}} \left[\int_{-\infty}^{y_{t,i}} n_i p(n_i|k_n) dn_i - \hat{\mu}_{n,i}^{(k_n)} \int_{-\infty}^{y_{t,i}} p(n_i|k_n) dn_i \right] \\
&= \frac{\Phi(y_{t,i}|k_n)}{\hat{\sigma}_{n,i}^{(k_n)^2}} \left[\tilde{\mu}_{n,t,i}^{(k_n)} - \hat{\mu}_{n,i}^{(k_n)} \right].
\end{aligned} \tag{C.7}$$

Para obtener la cuarta igualdad de la ecuación anterior se ha de considerar que la media de la gaussiana truncada $\tilde{\mu}_{n,t,i}^{(k_n)}$ se define como (ver ecuación (B.12))

$$\tilde{\mu}_{n,t,i}^{(k_n)} = \frac{1}{\Phi(y_{t,i}|k_n)} \int_{-\infty}^{y_{t,i}} n_i p(n_i|k_n) dn_i. \tag{C.8}$$

A partir de las ecuaciones (C.6) y (C.7), podemos reescribir el valor de α en la ecuación (C.4) de la siguiente forma:

$$\alpha = \frac{1}{p(y_{t,i}|k_x,k_n)}\left[\frac{p(y_{t,i}|k_x)\Phi(y_{t,i}|k_n)}{\hat{\sigma}_{n,i}^{(k_n)^2}}\left(\tilde{\mu}_{n,t,i}^{(k_n)}-\hat{\mu}_{n,i}^{(k_n)}\right)\right.$$
$$\left.+\frac{p(y_{t,i}|k_n)\Phi(y_{t,i}|k_x)}{\hat{\sigma}_{n,i}^{(k_n)^2}}\left(y_{t,i}-\hat{\mu}_{n,i}^{(k_n)}\right)\right]. \tag{C.9}$$

Simplificamos entonces la expresión usando los pesos $w_{t,i}^{(k_x,k_n)}$ y $1-w_{t,i}^{(k_x,k_n)}$ definidos en las ecuaciones (4.59) y (4.60):

$$\begin{aligned}\alpha &= \frac{1}{\hat{\sigma}_{n,i}^{(k_n)^2}}\left[w_{t,i}^{(k_x,k_n)}\left(\tilde{\mu}_{n,t,i}^{(k_n)}-\hat{\mu}_{n,i}^{(k_n)}\right)+\left(1-w_{t,i}^{(k_x,k_n)}\right)\left(y_{t,i}-\hat{\mu}_{n,i}^{(k_n)}\right)\right]\\ &= \frac{1}{\hat{\sigma}_{n,i}^{(k_n)^2}}\left[w_{t,i}^{(k_x,k_n)}\tilde{\mu}_{n,t,i}^{(k_n)}+\left(1-w_{t,i}^{(k_x,k_n)}\right)y_{t,i}-\hat{\mu}_{n,i}^{(k_n)}\right].\end{aligned} \tag{C.10}$$

Sustituyendo en la ecuación (C.4) el valor obtenido para α, nos queda que

$$\frac{\partial\mathcal{Q}(\mathcal{M}_n,\hat{\mathcal{M}}_n)}{\partial\hat{\mu}_{n,i}^{(k_n)}} = \frac{\sum_{t=1}^{T}\sum_{k_x=1}^{M_x}\gamma_t^{(k_x,k_n)}\left[w_{t,i}^{(k_x,k_n)}\tilde{\mu}_{n,t,i}^{(k_n)}+(1-w_{t,i}^{(k_x,k_n)})y_{t,i}-\hat{\mu}_{n,i}^{(k_n)}\right]}{\hat{\sigma}_{n,i}^{(k_n)^2}}. \tag{C.11}$$

Antes de continuar con la derivación de la expresión para el cálculo de las medias del GMM, definimos las siguientes variables auxiliares que nos ayudarán a simplificar el desarrollo de la expresión anterior,

$$\gamma_t^{(k_n)} = \textstyle\sum_{k_x=1}^{M_x}\gamma_t^{(k_x,k_n)}, \qquad m_{t,i}^{(k_n)} = \textstyle\sum_{k_x=1}^{M_x}\gamma_t^{(k_x,k_n)}w_{t,i}^{(k_x,k_n)}\ , \tag{C.12}$$

y también

$$\begin{array}{l|l}
a = \displaystyle\sum_{t=1}^{T}\sum_{k_x=1}^{M_x}\gamma_t^{(k_x,k_n)}w_{t,i}^{(k_x,k_n)}\tilde{\mu}_{n,t,i}^{(k_n)} & b = \displaystyle\sum_{t=1}^{T}\sum_{k_x=1}^{M_x}\gamma_t^{(k_x,k_n)}\left(1-w_{t,i}^{(k_x,k_n)}\right)y_{t,i}\\
\quad = \displaystyle\sum_{t=1}^{T} m_{t,i}^{(k_n)}\tilde{\mu}_{n,t,i}^{(k_n)} & \quad = \displaystyle\sum_{t=1}^{T}\left(\gamma_t^{(k_n)}-m_{t,i}^{(k_n)}\right)y_{t,i}\\
\hline
c = \displaystyle\sum_{t=1}^{T}\sum_{k_x=1}^{M_x}\gamma_t^{(k_x,k_n)}\hat{\mu}_{n,i}^{(k_n)} & \\
\quad = \hat{\mu}_{n,i}^{(k_n)}\displaystyle\sum_{t=1}^{T}\gamma_t^{(k_n)} &
\end{array}. \tag{C.13}$$

Usando estas variables auxiliares, la ecuación (C.11) se puede expresar como

$$\frac{\partial\mathcal{Q}(\mathcal{M}_n,\hat{\mathcal{M}}_n)}{\partial\hat{\mu}_{n,i}^{(k_n)}} = \frac{1}{\hat{\sigma}_{n,i}^{(k_n)^2}}(a+b-c). \tag{C.14}$$

Igualando a cero y despejando $\hat{\mu}_{n,i}^{(k_n)}$ se tiene que

$$\begin{aligned}\frac{\partial \mathcal{Q}(\mathcal{M}_n, \hat{\mathcal{M}}_n)}{\partial \hat{\mu}_{n,i}^{(k_n)}} &= \frac{1}{\hat{\sigma}_{n,i}^{(k_n)^2}}(a+b-c) = 0 \\ &\Rightarrow c = a + b \\ &\Rightarrow \hat{\mu}_{n,i}^{(k_n)} \sum_{t=1}^{T} \gamma_t^{(k_n)} = \sum_{t=1}^{T} m_{t,i}^{(k_n)} \tilde{\mu}_{n,t,i}^{(k_n)} + \left(\gamma_t^{(k_n)} - m_{t,i}^{(k_n)}\right) y_{t,i}. \end{aligned} \tag{C.15}$$

Para poder despejar el término $\hat{\mu}_{n,i}^{(k_n)}$ de la ecuación anterior, supondremos que los valores de las máscaras $m_{t,i}^{(k_n)}$ y las estimas parciales de ruido $\tilde{\mu}_{n,t,i}^{(k_n)}$ son independientes del modelo de ruido a estimar $\hat{\mathcal{M}}_n$ y, por tanto, se calculan usando la hipótesis actual del GMM de ruido $\mathcal{M}_n$. En este caso la fórmula final para la estimación de las medias del modelo es

$$\hat{\mu}_{n,i}^{(k_n)} = \frac{\sum_{t=1}^{T} m_{t,i}^{(k_n)} \tilde{\mu}_{n,t,i}^{(k_n)} + \left(\gamma_t^{(k_n)} - m_{t,i}^{(k_n)}\right) y_{t,i}}{\sum_{t=1}^{T} \gamma_t^{(k_n)}}. \tag{C.16}$$

Como se puede observar, la fórmula de reestimación obtenida revela que las medias del modelo se calculan como una combinación lineal de las observaciones $y_{t,i}$ y de ciertas estimas parciales $\tilde{\mu}_{n,t,i}^{(k_n)}$. Según el modelo de enmascaramiento de la ecuación (4.4), sólo una fuente sonora, voz o ruido, puede dominar en cada instante de tiempo. Bajo esta perspectiva, la observación $y_{t,i}$ puede verse como la estimación de la energía del ruido cuando éste domina. Por otro lado, cuando es la voz la que enmascara al ruido, la única información que disponemos sobre éste es el rango en el que se moverá su energía: $n_{t,i} \in (-\infty, y_{t,i}]$. Así, $\tilde{\mu}_{n,t,i}^{(k_n)}$ indica el valor estimado para $n_{t,i}$ cuando la voz enmascara al ruido. Ambas estimas de ruido, $y_{t,i}$ y $\tilde{\mu}_{n,t,i}^{(k_n)}$, van ponderadas en la ecuación (C.16) por los valores de las máscaras $m_{t,i}^{(k_n)}$ y $(\gamma_t^{(k_n)} - m_{t,i}^{(k_n)})$ que indican, desde el punto de vista de la gaussiana k_n-ésima, la probabilidad de presencia de voz y la probabilidad de enmascaramiento, respectivamente.

Finalmente, las estimas parciales de ruido calculadas para cada instante de tiempo son promediadas a fin de obtener las medias de las gaussianas. El factor de normalización usado para calcular dichas medias, como se aprecia en (C.16), concuerda con la suma de las probabilidades a posteriori de las gaussianas $\gamma_t^{(k_n)}$. En el caso de que el GMM de ruido cuente con una única gaussiana, $M_n = 1$, entonces $\gamma_t^{(k_n)} = 1$ y el factor de normalización equivale al número total de vectores observados T.

C.2. Ajuste de las varianzas del modelo

El ajuste de las matrices de covarianza del modelo de ruido a los datos observados se efectúa siguiendo un procedimiento análogo al usado para calcular las medias. Así, la derivada parcial de la ecuación (C.3) respecto a la desviación típica $\hat{\sigma}_{n,i}^{(k_n)}$ que queremos calcular es

$$\frac{\partial \mathcal{Q}(\mathcal{M}_n, \hat{\mathcal{M}}_n)}{\partial \hat{\sigma}_{n,i}^{(k_n)}} = \sum_{t=1}^{T} \sum_{k_x=1}^{M_x} \gamma_t^{(k_x,k_n)} \underbrace{\frac{1}{p(y_{t,i}|k_x,k_n)} \left[p(y_{t,i}|k_x) \frac{\partial \Phi(y_{t,i}|k_n)}{\partial \hat{\sigma}_{n,i}^{(k_n)}} + \Phi(y_{t,i}|k_x) \frac{\partial p(y_{t,i}|k_n)}{\partial \hat{\sigma}_{n,i}^{(k_n)}} \right]}_{\beta}. \tag{C.17}$$

Otra vez en el cálculo de los parámetros del modelo aparecen derivadas parciales que involucran a PDFs y CDFs. Usando la igualdad

$$\frac{\partial p(x)}{\partial \sigma} = p(x) \frac{1}{\sigma} \left[\frac{(x-\mu)^2}{\sigma^2} - 1 \right], \tag{C.18}$$

tenemos que la derivada parcial de $p(y_{t,i}|k_n)$ respecto a su desviación típica es

$$\frac{\partial p(y_{t,i}|k_n)}{\partial \hat{\sigma}_{n,i}^{(k_n)}} = p(y_{t,i}|k_n) \frac{1}{\hat{\sigma}_{n,i}^{(k_n)}} \left[\frac{\left(y_{t,i} - \hat{\mu}_{n,i}^{(k_n)}\right)^2}{\hat{\sigma}_{n,i}^{(k_n)^2}} - 1 \right], \tag{C.19}$$

y la derivada parcial de la CDF $\Phi(y_{t,i}|k_n)$ es

$$\begin{aligned}
\frac{\partial \Phi(y_{t,i}|k_n)}{\partial \hat{\sigma}_{n,i}^{(k_n)}} &= \int_{-\infty}^{y_{t,i}} \frac{\partial p(n_i|k_n)}{\partial \hat{\sigma}_{n,i}^{(k_n)}} dn_i \\
&= \frac{1}{\hat{\sigma}_{n,i}^{(k_n)}} \int_{-\infty}^{y_{t,i}} p(n_i|k_n) \left[\frac{\left(n_i - \hat{\mu}_{n,i}^{(k_n)}\right)^2}{\hat{\sigma}_{n,i}^{(k_n)^2}} - 1 \right] dn_i \\
&= \frac{1}{\hat{\sigma}_{n,i}^{(k_n)^3}} \left[\int_{-\infty}^{y_{t,i}} n_i^2 p(n_i|k_n) dn_i + \hat{\mu}_{n,i}^{(k_n)^2} \int_{-\infty}^{y_{t,i}} p(n_i|k_n) dn_i \right. \\
&\quad \left. -2\hat{\mu}_{n,i}^{(k_n)} \int_{-\infty}^{y_{t,i}} n_i p(n_i|k_n) dn_i \right] - \frac{1}{\hat{\sigma}_{n,i}^{(k_n)}} \int_{-\infty}^{y_{t,i}} p(n_i|k_n) dn_i.
\end{aligned} \tag{C.20}$$

Para simplificar esta ecuación y dejarla en una forma más compacta, retomamos aquí parte del estudio que se llevó a cabo en el apéndice B sobre la distribución normal truncada. Así, en la ecuación (B.12) veíamos que la media de esta distribución

(momento de primer orden) se define como

$$\tilde{\mu}_n \equiv \mathbb{E}[n|n \leq y] = \frac{1}{\Phi_n(y)} \int_{-\infty}^{y} n p_n(n) dn. \tag{C.21}$$

Por otra parte, el momento no centrado de segundo orden viene dado por

$$\mathbb{E}[n^2|n \leq y] = \frac{1}{\Phi_n(y)} \int_{-\infty}^{y} n^2 p_n(n) dn. \tag{C.22}$$

Sustituyendo estas definiciones en la ecuación (C.20) resulta en la siguiente expresión

$$\begin{aligned}
&\frac{\partial \Phi(y_{t,i}|k_n)}{\partial \hat{\sigma}_{n,i}^{(k_n)}} \\
&= \frac{1}{\hat{\sigma}_{n,i}^{(k_n)^3}} \left[\mathbb{E}\left[n_i^2 | n_i \leq y_{t,i}, k_n \right] \Phi(y_{t,i}|k_n) + \hat{\mu}_{n,i}^{(k_n)^2} \Phi(y_{t,i}|k_n) - 2\hat{\mu}_{n,i}^{(k_n)} \tilde{\mu}_{n,t,i}^{(k_n)} \Phi(y_{t,i}|k_n) \right] \\
&\quad - \frac{1}{\hat{\sigma}_{n,i}^{(k_n)}} \Phi(y_{t,i}|k_n) \\
&= \frac{\Phi(y_{t,i}|k_n)}{\hat{\sigma}_{n,i}^{(k_n)^3}} \underbrace{\left[\mathbb{E}\left[n_i^2 | n_i \leq y_{t,i}, k_n \right] + \hat{\mu}_{n,i}^{(k_n)^2} - 2\hat{\mu}_{n,i}^{(k_n)} \tilde{\mu}_{n,t,i}^{(k_n)} \right]}_{= \tilde{\sigma}_{n,t,i}^{(k_n)^2} + \left(\tilde{\mu}_{n,t,i}^{(k_n)} - \hat{\mu}_{n,i}^{(k_n)} \right)^2} - \frac{\Phi(y_{t,i}|k_n)}{\hat{\sigma}_{n,i}^{(k_n)}} \\
&= \frac{\Phi(y_{t,i}|k_n)}{\hat{\sigma}_{n,i}^{(k_n)^3}} \underbrace{\left[\tilde{\sigma}_{n,t,i}^{(k_n)^2} + \left(\tilde{\mu}_{n,t,i}^{(k_n)} - \hat{\mu}_{n,i}^{(k_n)} \right)^2 \right]}_{\zeta} - \frac{\Phi(y_{t,i}|k_n)}{\hat{\sigma}_{n,i}^{(k_n)}},
\end{aligned} \tag{C.23}$$

donde, para obtener la tercera igualdad de la ecuación, ha de considerarse que la varianza de la distribución truncada se define como (ver ecuaciones (B.14) y (B.12)):

$$\begin{aligned}
\tilde{\sigma}_{n,t,i}^{(k_n)^2} &= \mathrm{Var}\left[n_i | n_i \leq y_{t,i}, k_n \right] \\
&= \mathbb{E}\left[n_i^2 | n_i \leq y_{t,i}, k_n \right] - \mathbb{E}\left[n_i | n_i \leq y_{t,i}, k_n \right]^2 \\
&= \mathbb{E}\left[n_i^2 | n_i \leq y_{t,i}, k_n \right] - \underbrace{\left(\mu_{n,i}^{(k_n)} - \sigma_{n,i}^{(k_n)} \frac{\mathcal{N}(\overline{y}_{t,i})}{\Phi(\overline{y}_{t,i})} \right)^2}_{= \tilde{\mu}_{n,t,i}^{(k_n)^2}},
\end{aligned} \tag{C.24}$$

con

$$\overline{y}_{t,i} = \frac{y_{t,i} - \mu_{n,i}^{(k_n)}}{\sigma_{n,i}^{(k_n)}}. \tag{C.25}$$

Retomando el cómputo de la derivada parcial $\frac{\partial \mathcal{Q}(\mathcal{M}_n, \hat{\mathcal{M}}_n)}{\partial \hat{\sigma}_{n,i}^{(k_n)}}$ de la ecuación (C.17), tenemos que, usando las expresiones obtenidas en (C.19) y (C.23) para $\frac{\partial p(y_{t,i}|k_n)}{\partial \hat{\sigma}_{n,i}^{(k_n)}}$ y

$\frac{\partial\Phi(y_{t,i}|k_n)}{\partial\hat{\sigma}_{n,i}^{(k_n)}}$, la variable auxiliar β que aparece en dicha ecuación puede desarrollarse de la siguiente manera:

$$\begin{aligned}
\beta &= \frac{1}{p(y_{t,i}|k_x,k_n)}\left[p(y_{t,i}|k_x)\frac{\partial\Phi(y_{t,i}|k_n)}{\partial\hat{\sigma}_{n,i}^{(k_n)}} + \Phi(y_{t,i}|k_x)\frac{\partial p(y_{t,i}|k_n)}{\partial\hat{\sigma}_{n,i}^{(k_n)}}\right] \\
&= \underbrace{\frac{p(y_{t,i}|k_x)\Phi(y_{t,i}|k_n)}{p(y_{t,i}|k_x,k_n)}}_{w_{t,i}^{(k_x,k_n)}}\frac{\zeta}{\hat{\sigma}_{n,i}^{(k_n)^3}} - \underbrace{\frac{p(y_{t,i}|k_x)\Phi(y_{t,i}|k_n)}{p(y_{t,i}|k_x,k_n)}}_{w_{t,i}^{(k_x,k_n)}}\frac{1}{\hat{\sigma}_{n,i}^{(k_n)}} \\
&\quad + \underbrace{\frac{p(y_{t,i}|k_n)\Phi(y_{t,i}|k_x)}{p(y_{t,i}|k_x,k_n)}}_{1-w_{t,i}^{(k_x,k_n)}}\frac{1}{\hat{\sigma}_{n,i}^{(k_n)}}\left[\frac{\left(y_{t,i}-\hat{\mu}_{n,i}^{(k_n)}\right)^2}{\hat{\sigma}_{n,i}^{(k_n)^2}} - 1\right] \\
&= \frac{w_{t,i}^{(k_x,k_n)}}{\hat{\sigma}_{n,i}^{(k_n)^3}}\zeta - \frac{w_{t,i}^{(k_x,k_n)}}{\hat{\sigma}_{n,i}^{(k_n)}} + \frac{1-w_{t,i}^{(k_x,k_n)}}{\hat{\sigma}_{n,i}^{(k_n)^3}}\left(y_{t,i}-\hat{\mu}_{n,i}^{(k_n)}\right)^2 - \frac{1-w_{t,i}^{(k_x,k_n)}}{\hat{\sigma}_{n,i}^{(k_n)}} \\
&= \frac{1}{\hat{\sigma}_{n,i}^{(k_n)^3}}\left[w_{t,i}^{(k_x,k_n)}\zeta + \left(1-w_{t,i}^{(k_x,k_n)}\right)\left(y_{t,i}-\hat{\mu}_{n,i}^{(k_n)}\right)^2\right] - \frac{1}{\hat{\sigma}_{n,i}^{(k_n)}}.
\end{aligned} \tag{C.26}$$

Sustituyendo la expresión obtenida para β en la ecuación (C.17), la derivada parcial de la función auxiliar resulta en

$$\begin{aligned}
&\frac{\partial\mathcal{Q}(\mathcal{M}_n,\hat{\mathcal{M}}_n)}{\partial\hat{\sigma}_{n,i}^{(k_n)}} \\
&\quad= \sum_{t=1}^{T}\sum_{k_x=1}^{M_x}\gamma_t^{(k_x,k_n)}\left\{\frac{1}{\hat{\sigma}_{n,i}^{(k_n)^3}}\left[w_{t,i}^{(k_x,k_n)}\zeta + \left(1-w_{t,i}^{(k_x,k_n)}\right)\left(y_{t,i}-\hat{\mu}_{n,i}^{(k_n)}\right)^2\right] - \frac{1}{\hat{\sigma}_{n,i}^{(k_n)}}\right\}.
\end{aligned} \tag{C.27}$$

Para terminar de obtener la expresión que permite calcular $\hat{\sigma}_{n,i}^{(k_n)}$, igualamos a cero la ecuación anterior y despejamos el valor de $\hat{\sigma}_{n,i}^{(k_n)}$:

$$\begin{aligned}
0 &= \frac{1}{\hat{\sigma}_{n,i}^{(k_n)^3}}\sum_{t=1}^{T}\sum_{k_x=1}^{M_x}\gamma_t^{(k_x,k_n)}\left[w_{t,i}^{(k_x,k_n)}\zeta + \left(1-w_{t,i}^{(k_x,k_n)}\right)\left(y_{t,i}-\hat{\mu}_{n,i}^{(k_n)}\right)^2\right] \\
&\quad - \frac{1}{\hat{\sigma}_{n,i}^{(k_n)}}\sum_{t=1}^{T}\sum_{k_x=1}^{M_x}\gamma_t^{(k_x,k_n)} \\
&= \frac{1}{\hat{\sigma}_{n,i}^{(k_n)^3}}\sum_{t=1}^{T}\left[m_{t,i}^{(k_n)}\zeta + \left(\gamma_t^{(k_n)}-m_{t,i}^{(k_n)}\right)\left(y_{t,i}-\hat{\mu}_{n,i}^{(k_n)}\right)^2\right] - \frac{1}{\hat{\sigma}_{n,i}^{(k_n)}}\sum_{t=1}^{T}\gamma_t^{(k_n)} \\
&= \sum_{t=1}^{T}\left[m_{t,i}^{(k_n)}\zeta + \left(\gamma_t^{(k_n)}-m_{t,i}^{(k_n)}\right)\left(y_{t,i}-\hat{\mu}_{n,i}^{(k_n)}\right)^2\right] - \hat{\sigma}_{n,i}^{(k_n)^2}\sum_{t=1}^{T}\gamma_t^{(k_n)},
\end{aligned} \tag{C.28}$$

donde se aplican las definiciones de las máscaras $m_{t,i}^{(k_n)}$ y las probabilidades a posteriori $\gamma_t^{(k_n)}$ que aparecen en la ecuación (C.12). Adicionalmente, para obtener la tercera expresión se han multiplicado ambos lados de la igualdad por $\hat{\sigma}_{n,i}^{(k_n)^3}$.

Finalmente, despejamos el valor de $\hat{\sigma}_{n,i}^{(k_n)^2}$ y sustituimos ζ por su valor dado en la ecuación (C.23) [1], de forma que

$$\hat{\sigma}_{n,i}^{(k_n)^2} = \frac{\sum_{t=1}^{T} m_{t,i}^{(k_n)} \left[\tilde{\sigma}_{n,t,i}^{(k_n)^2} + \left(\tilde{\mu}_{n,t,i}^{(k_n)} - \hat{\mu}_{n,i}^{(k_n)}\right)^2\right] + \left(\gamma_t^{(k_n)} - m_{t,i}^{(k_n)}\right)\left(y_{t,i} - \hat{\mu}_{n,i}^{(k_n)}\right)^2}{\sum_{t=1}^{T} \gamma_t^{(k_n)}}. \quad \text{(C.29)}$$

De nuevo la expresión obtenida para el cálculo de las varianzas abarca dos situaciones bien diferenciadas: aquella en la que el ruido enmascara a la voz y la situación inversa, esto es, aquellan donde la energía de la voz es dominante. En la primera situación, la más sencilla, la varianza se calcula de la forma usual, esto es, promediando las diferencias al cuadrado entre las observaciones y la media, $(y_{t,i} - \hat{\mu}_{n,i}^{(k_n)})^2$. La segunda situación implica una mayor complejidad, ya que se desconoce el valor exacto de la energía del ruido, disponiendo únicamente de una estimación $\tilde{\mu}_{n,t,i}^{(k_n)}$ más o menos precisa de la misma. En este caso a la diferencia entre el valor estimado y la media, $(\tilde{\mu}_{n,t,i}^{(k_n)} - \hat{\mu}_{n,i}^{(k_n)})^2$, hay que añadir el error esperado de la estimación. Este error viene modelado por la varianza $\tilde{\sigma}_{n,t,i}^{(k_n)^2}$.

C.3. Ajuste de los pesos de las componentes

Como última cuestión, en esta sección estudiaremos la estimación de los pesos $\hat{\pi}_n^{(k_n)}$ asociados con las gaussianas del GMM de ruido $\hat{\mathcal{M}}_n$. Hasta ahora hemos visto que para hallar el valor de los distintos parámetros del modelo bastaba con derivar la función auxiliar $Q(\mathcal{M}_n, \hat{\mathcal{M}}_n)$ respecto al parámetro en cuestión e igualar a cero la expresión resultante. El cálculo de los pesos de las gaussianas, no obstante, supone una dificultad añadida, ya que estos deben satisfacer las siguientes restricciones:

$$\begin{aligned} \hat{\pi}_n^{(k_n)} &\geq 0, \\ \sum_{k_n=1}^{M_n} \hat{\pi}_n^{(k_n)} &= 1. \end{aligned} \quad \text{(C.30)}$$

[1] Al igual que en la derivación de la expresión para el cálculo de las medias del modelo, aquí suponemos que $\tilde{\sigma}_{n,t,i}^{(k_n)^2}$, $\tilde{\mu}_{n,t,i}^{(k_n)}$ y $m_{t,i}^{(k_n)}$ son constantes y se calculan usando el GMM obtenido en la iteración previa.

Para asegurarnos que estas dos restricciones se satisfacen en el cálculo de $\hat{\pi}_n^{(k_n)}$, introducimos un multiplicador de Lagrange en la función auxiliar $Q(\mathcal{M}_n, \hat{\mathcal{M}}_n)$ de la ecuación (C.3), quedando ésta de la siguiente forma:

$$\begin{aligned}&\mathcal{Q}(\mathcal{M}_n, \hat{\mathcal{M}}_n)\\ &= \sum_{t=1}^{T}\sum_{k_x=1}^{M_x}\sum_{k_n=1}^{M_n}\gamma_t^{(k_x,k_n)}\left[\sum_{i=1}^{D}\log p(y_{t,i}|k_x,k_n) + \log\hat{\pi}_n^{(k_n)}\right] - \lambda\left[\sum_{k_n=1}^{M_n}\hat{\pi}_n^{(k_n)} - 1\right].\end{aligned} \tag{C.31}$$

Derivando respecto a $\hat{\pi}_n^{(k_n)}$ nos queda lo siguiente

$$\begin{aligned}\frac{\partial\mathcal{Q}(\mathcal{M}_n, \hat{\mathcal{M}}_n)}{\partial\hat{\pi}_n^{(k_n)}} &= \sum_{t=1}^{T}\sum_{k_x=1}^{M_x}\frac{\gamma_t^{(k_x,k_n)}}{\hat{\pi}_n^{(k_n)}} - \lambda = 0\\ &\Rightarrow \sum_{t=1}^{T}\sum_{k_x=1}^{M_x}\gamma_t^{(k_x,k_n)} - \lambda\hat{\pi}_n^{(k_n)} = 0.\end{aligned} \tag{C.32}$$

Sumando a ambos lados de la igualdad respecto a k_n tenemos que

$$\lambda = \sum_{t=1}^{T}\sum_{k_x=1}^{M_x}\sum_{k_n=1}^{M_n}\gamma_t^{(k_x,k_n)} = T. \tag{C.33}$$

Por lo tanto, la expresión para la estimación de las probabilidades a priori de las gaussianas es

$$\hat{\pi}_n^{(k_n)} = \frac{1}{T}\sum_{t=1}^{T}\sum_{k_x=1}^{M_x}\gamma_t^{(k_x,k_n)} = \frac{1}{T}\sum_{t=1}^{T}\gamma_t^{(k_n)}. \tag{C.34}$$

Finalizando así nuestra adaptación del algoritmo EM para la estimación del modelo de ruido.

APÉNDICE D

Conclusions

THIS work arises from the need to increase the robustness of the automatic speech recognition systems in noisy conditions. To achieve this goal, we investigated several methods aimed at mitigating the effect of noise on the speech features. The following section summarizes the most important conclusions drawn throughout this thesis.

D.1. Conclusions

- We analyzed the effect of noise (additive and convolutional) over the speech features used by most existing recognition systems. From a statistical point of view, the noise generates a mismatch between the probability distributions of the training and testing data. Thus, the acoustic models cannot properly represent the (noisy) input speech, which finally results in a recognition performance reduction.
- A review of the different strategies proposed to increase the robustness of speech recognition in noisy conditions has been carried out.
- Although the optimal strategy for recognition in noise requires acoustic models trained under the same acoustic conditions as for testing, this approach is impractical in most situations. Moreover, real-time adaptation of the models to match highly non-stationary noise, may also be unfeasible due to the high computational involved. Therefore, it has been argued that this problem requires a more dynamic and efficient than the previous two. In order to overcome these problems, the feature compensation approach is adopted in this work.

- As a first approach, we have proposed a set of compensation techniques that assume the existence of stereo recordings with simultaneous channels for clean and noisy speech. From these recordings, the proposed techniques result in suitable transformations that try to compensate the noisy speech features. When these techniques are evaluated under the same noise conditions than those considered during training, the recognition results almost achieve those obtained using models trained in testing conditions. In the presence of unknown noise, by contrast, the error recognition obtained by these techniques is greatly enhanced.

- To counteract this degradation that occurs in unknown noise, we have also shown that our proposals can be used in combination with other robust recognition techniques. For example, they can be used as post-processing of the features extracted with a robust *front-end* (e.g., ETSI AFE), or in combination with robust modeling methods such as multicondition training. In both cases, the improvement achieved is significant.

- We have proposed an MD imputation technique for the estimation of noisy spectral features called TGI. In conditions where it is possible to perfectly know the masking pattern of the observed signal, the recognition results obtained by this method are comparable to those obtained using clean speech. However, the performance of this technique is not always satisfactory when estimates masks are employed.

- We have shown that the main weakness of TGI is derived from the fact of employing binary masks. When this mask is calculated from noise power estimates, the possible errors that can occur in this estimatiion is magnified by the binary decision involved. Thus, the soft decision associated to the use of continuous masks lead us to better estimators.

- In response to the errors introduced by binary masks, we have also proposed an alternative MD reconstruction technique called MMSR. This technique combines the simplicity of the masking model common to all MD techniques with the use of probabilistic models for the noise, which results in an increased robustness against estimation errors. In addtion to better recognition results (with respect to TGI), the derivation of MMSR combines two different issues of MD into a single one: (i) mask estimation masks and (ii) noise spectrum estimation.

- We have shown that MMSR can alternatively be interpreted as a method for computing the continuous masks required by MD imputation. In comparison with other methods for mask estimation, our proposal has several advantages. First,

the Bayesian framework adopted allows the exploitation of other information that may be useful for a better spectrum segregation (e.g., the pitch frequency). Furthermore, the proposed method does not require any heuristic adjustment of experimental parameters.

- The EM algorithm proposed for estimating the noise model has proved efficient in characterizing noise in the databases evaluated. Compared with other noise estimation techniques, the algorithm allows a more accurate representation of the non-stationary noise characteristics.
- We have proposed several options for exploiting the speech temporal redundancy in the proposed compensation techniques. First, different strategies have been investigated for exploiting short-time temporal correlations. Furthermore, we have derived suitable expressions for adjusting the HMM modeling to our compensation methods. In all cases, the results obtained confirm the HMM modeling as the most efficient option, since it allows the exploitation of the whole spectral information contained in the speech utterance.
- Finally, we have studied two alternative strategies for computing and exploiting uncertainty measures obtained during the compensation process: MMSE variance propagation (*soft-data* technique) and weighted Viterbi algorithm (WVA). Between these two strategies, we have experimentally shown that the WVA algorithm yields the best results. This result is justified by the lack of precision in some of the assumptions on which the *soft-data* approach is based.

D.2. Contributions

The main contributions of this work can be summarized as:

- Development of a set of speech feature compensation techniques based on MMSE estimation, VQ-modeling of clean and noisy feature spaces and use of stereo data in order to derive the transformations applied to the noisy speech [122, 123, 124].
- A reconstruction algorithm for noisy feature estimation based on an a priori spectral segmentation of speech features into realible and distorted ones [126, 127].
- Development of a compensation technique based on an analytical model of the distortion/masking segmentation applied over the speech spectral features [121, 125].

- Proposal of an algorithm for feature reliability estimation (masking) in presence of additive noise [121, 125].
- Derivation of an iterative algorithm for GMM parameter estimation of noise models.
- Analysis of the temporal redundancy available in speech for its application in compensation techniques [122, 127].
- Derivation of several estimated feature uncertainty metrics and exploitation strategies for its use within the speech recognizer [122, 126].

D.3. Future work

Some of the techniques proposed in this thesis have been derived from the speech masking model described in chapter 4. Although the experimental results have confirmed the high accuracy of this model, it can be foreseen several research lines for increasing their accuracy and robustness. In this sense, further investigations about the modeling of the residual error committed by the model would be interesting. In this regard, figure 4.1 shown several histograms computed over Aurora2 database including the error masking pattern respect to the signal SNR. From this figure, approximating the distribution of the error by a SNR-dependent distribution seems reasonable. This pdf could be later used by different reconstruction techniques in order to obtain a more accurate estimate of the noisy features.

In addition, another issue to be addressed is to establish the relation between the studied masking model and the actual perceptual masking that occurs in the human ear. We think that the model is a simplification of the perceptual masking so it could be improved with the latest advances in this field. Key aspects that should be considered include the temporal masking process and the use of frequency-dependent masking thresholds.

Regarding to the proposed compensation techniques, one aspect which could improve the MMSE estimation accuracy is the use of speaker-dependent speech. Although the identity of the speaker is often unknown, it would be possible to estimate a set of parameters which model some speaker features. For example, the vocal tract length could be estimated as well as a transformation matrix which maximize the observed data likelihood, as it is done in MLLR. These speaker-dependent parameters can also be later employed for model speaker adaptation, and thus obtain more accurate estimates.

Another issue is the use of full covariance matrices (not diagonal) in the MMSR technique. These matrices better exploit the correlations between speech features, allowing a better estimation. However, its use was rejected since it leads to the evaluation of every possible voice/noise segmentation for each speech vector, which turns out impracticable. Nevertheless, in practice, most of these cases can be disregarded as they are unlikely. Thus, a pruning strategy employing low-level information could be designed to rule out these combinations and evaluate only the potential ones. This information would be consist, for example, of the frequency harmonic relation, *onset* / *offset* common frequency, source spatial position (if several microphones are available), etc.

The EM algorithm used for noise model estimation could also be improved. The most immediate improvement is its extension to convolutional noise. Apart from this, another issue is the automatic estimation of the number of GMM components considered for noise modeling. As mentioned before, this is a noteworthy issue, as this number balances the accuracy of the noise model and the data required for the robust estimation of its parameters.

Finally, regarding the techniques exploiting the estimate uncertainty, the best results were obtained by the weighted Viterbi algorithm. In the conducted experiments, this algorithm has been evaluated using a single weight per feature vector. However, it would be interesting to evaluate the use a weight per each feature component, since that allows a better control over the observation probability computation.

Bibliografía

[1] *ETSI ES 201 108 - Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms.* 129, 215, 220

[2] *ETSI ES 202 050 - Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms.* 129, 216, 220

[3] *ETSI ES 202 211 - Distributed speech recognition; Extended front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm.* 129

[4] *ETSI ES 202 212 - Distributed speech recognition; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm.* 129

[5] *Objective measurement of active speech level.* 213

[6] *Transmission characteristics for wideband digital loudspeaking and hands-free telephony terminals.* 215

[7] *Transmission performance characteristics of pulse code modulation channels.* 211

[8] Acero, A.: *Acoustical and environmental robustness in automatic speech recognition.* Tesis de Doctorado, Carnegie Mellon University, 1990. 15, 72

[9] Acero, A. *y cols.*: *HMM adaptation using vector Taylor series for noisy speech recognition.* En *Proc. ICSLP*, págs. 229–232, 2000. 44, 45, 186, 189

[10] Acero, A. y R. M. Stern: *Robust speech recognition by normalization of the acoustic space.* En *Proc. ICASSP*, págs. 893–896, 1991. 72

[11] Afify, M., X. Cui y Y. Gao: *Stereo-based stochastic mapping for robust speech recognition.* IEEE Trans. Audio Speech Lang. Process., 17(7):1325–1334, Septiembre 2009. 76

[12] Ahadi, S. M. y P. C. Woodland: *Combined Bayesian and predictive techniques for rapid speaker adaptation of continuous density hidden Markov models.* Comput. Speech Lang., 11(3):187–206, Julio 1997. 31

[13] Ahmed, S. y V. Tresp: *Some solutions to the missing feature problem in vision.* En Hanson, S., J. Cowan y C. Giles (editores): *Proc. Advances in Neural Information Processing Systems*, págs. 393–400. Morgan Kaufmann, San Mateo, CA, 1993. 84

[14] Akbacak, M. y J. H. L. Hansen: *Environmental sniffing: Noise knowledge estimation for robust speech systems.* IEEE Trans. Audio Speech Lang. Process., 15(2):465–477, Febrero 2007. 107

[15] Allen, J. B.: *How do humans process and recognize speech?* IEEE Trans. Speech Audio Process., 2(4):567–577, Octubre 1994. 83, 191

[16] Anastasakos, T. *y cols.*: *A compact model for speaker-adaptive training.* En *Proc. ICSLP*, págs. 1137–1140, 1996. 29, 55, 56, 57

[17] Anguita, J. y J. Hernando: *Jacobian adaptation with continuous noise estimation for real speaker verification applications.* En *Proc. IEEE Odyssey 2006: The Speaker and Language Recognition Workshop*, págs. 1–5, 2006. 108

[18] Arora, S. y B. Barak: *Computational complexity: A modern approach.* Cambridge University Press, 2009. 138, 263

[19] Arrowood, J. A. y M. A. Clements: *Using observation uncertainty in HMM decoding.* En *Proc. ICSLP*, págs. 1561–1564, 2002. 202, 203

[20] Atal, B. S.: *Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification.* J. Aco. Soc. Am., 55:1304–1312, Junio 1974. 59, 60, 144

[21] Badiezadegan, S. y R. C. Rose: *Mask estimation in non-stationary noise environments for missing feature based robust speech recognition.* En *Proc. Interspeech*, págs. 2062–2065, 2010. 103, 106

[22] Bahl, L., F. Jelinek y R. Mercer: *A Maximum Likelihood Approach to Continuous Speech Recognition.* IEEE Trans. Pattern Anal. Mach. Intell., 5(2):179–190, Marzo 1983. 8

[23] Baker, J. M. *y cols.*: *Research developments and directions in speech recognition and understanding, Part 1.* IEEE Signal Process. Mag., 26(3):75–80, Mayo 2009. 2

[24] Baker, J. M. *y cols.*: *Updated MINDS report on speech recognition and understanding, Part 2.* IEEE Signal Process. Mag., 26(4):78–85, Julio 2009. 2, 13

[25] Barker, J., M. Cooke y D. P. W. Ellis: *Decoding speech in the presence of other sound sources.* En *Proc. ICSLP*, págs. 270–273, 2000. 86, 91

[26] Barker, J., M. Cooke y D. P. W. Ellis: *Combining bottom-up and top-down constraints for robust ASR: The multisource decoder.* En *Proc. Workshop Consistent Reliable Acoustic Cues (CRAC)*, 2001. 86, 91

[27] Barker, J., M. Cooke y D. P. W. Ellis: *Decoding speech in the presence of other sources.* Speech Commun., 45(1):5–25, 2005. 86, 91, 274

[28] Barker, J., M. Cooke y P. Green: *Linking auditory scene analysis and robust ASR by missing data techniques.* En *Proc. WISP'01*, págs. 295–307, 2001. 106

[29] Barker, J., M. Cooke y P. D. Green: *Robust ASR based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise.* En *Proc. Eurospeech*, págs. 213–216, 2001. 90

[30] Barker, J. *y cols.*: *Recent advances in speech fragment decoding techniques.* En *Proc. Interspeech*, págs. 85–88, 2006. 86, 91

[31] Barker, J. *y cols.*: *Soft decisions in missing data techniques for robust automatic speech recognition.* En *Proc. ICSLP*, 2000. 90, 101, 104, 160, 184, 250

[32] Barker, J. *y cols.*: *Speech fragment decoding techniques for simultaneous speaker identification and speech recognition.* Comput. Speech Lang., 24(1):94–111, Enero 2010. 86, 91, 93

[33] Baum, L.: *An inequality and associated maximization technique in statistical estimation of probabilistic functions of Markov processes.* Inequalities, 3:1–8, 1972. 8

[34] Benesty, J., M. Mohan Sondhi y Y. Huang (editores): *Springer handbook of speech processing.* Springer-Verlag, 2008. 3, 61, 65

[35] Benítez, M. C. *y cols.*: *Including uncertainty of speech observations in robust speech recognition.* En *Proc. ICSLP*, págs. 137–140, 2004. 107, 203

[36] Bishop, C. M.: *Pattern recognition and machine learning.* Springer, 2006. 194

[37] Boll, S.: *Suppression of acoustic noise in speech using spectral subtraction.* IEEE Trans. Acoust., Speech, Signal Process., 27(2):113–120, Abril 1979. 23, 64, 119

[38] Borgström, B. J. y A. Alwan: *HMM-based estimation of unreliable spectral components for noise robust speech recognition.* En *Proc. Interspeech*, págs. 1769–1772, 2008. 191

[39] Borgström, B. J. y A. Alwan: *HMM-based reconstruction of unreliable spectrographic data for noise robust speech recognition.* IEEE Trans. Audio Speech Lang. Process., 18(6):1612–1623, 2010. 191

[40] Borgström, B. J. y A. Alwan: *A statistical approach to Mel-domain mask estimation for missing-feature ASR.* IEEE Signal Process. Lett., 17(11):941–944, Noviembre 2010. 103, 106

[41] Bregman, A.: *Auditory scene analysys.* MIT Press, 1990. 83

[42] Brown, G. J. y M. Cooke: *Computational auditory scene analysys.* Comput. Speech Lang., 8(4):297–336, Octubre 1994. 83, 92

[43] Buera, L.: *Normalización y adaptación a entornos acústicos para la robustez en sistemas de reconocimiento automático del habla.* Tesis de Doctorado, Universidad de Zaragoza, 2007. 75, 143, 225

[44] Buera, L. *y cols.*: *Cepstral vector normalization based on stereo data for robust speech recognition.* IEEE Trans. Audio Speech Lang. Process., 15(3):1098–1113, Marzo 2007. 68, 74, 75, 126, 132, 137, 138, 143, 225, 229, 263

[45] Candès, E., J. Romberg y T. Tao: *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information.* IEEE Trans. Inform. Theory, 52(2):489–509, Febrero 2006. 99

[46] Cardenal-López, A., C. García-Mateo y L. Docío-Fernandez: *Weighted Viterbi decoding strategies for distributed speech recognition over IP netw.* Speech Commun., 48(11):1422–1434, Noviembre 2006. 15

[47] Carmona, J. L.: *Reconocimiento de voz codificada sobre redes IP.* Tesis de Doctorado, Universidad de Granada, 2009. 15, 191

[48] Carmona, J. L. *y cols.*: *MMSE-based packet loss concealment for CELP-coded speech recognition.* IEEE Trans. Audio Speech Lang. Process., 18(6):1341–1353, Agosto 2010. 191, 202, 205, 235, 244

[49] Cerisara, C., S. Demange y J. P. Haton: *On noise masking for automatic missing data speech recognition: A survey and discussion*. Comput. Speech Lang., 21(3):443–457, 2007. 102

[50] Chen, S. *y cols.*: *Advances in speech transcription at IBM under the DARPA EARS program*. IEEE Trans. Audio Speech Lang. Process., 14(5):1596–1608, Septiembre 2006. 2

[51] Chesta, C., O. Siohan y C. Lee: *Maximum a posteriori linear regression for hidden Markov model adaptation*. En *Proc. Eurospeech*, págs. 211–214, 1999. 36

[52] Cho, E., J. O. Smith y B. Widrow: *Exploiting the harmonic structure for speech enhancement*. En *Proc. ICASSP*, págs. 4569–4572, 2012. 108

[53] Cohen, I.: *Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging*. IEEE Trans. Speech Audio Process., 11(5):466–475, Septiembre 2003. 108, 109

[54] Cohen, I. y B. Berdugo: *Noise estimation by minima controlled recursive averaging for robust speech enhancement*. IEEE Signal Process. Lett., 9(1):12–15, Enero 2002. 108, 109

[55] Cooke, M.: *A glimpsing model of speech perception in noise*. J. Acoust. Soc. Am., 119(3):1562–1573, Marzo 2006. 81, 83, 90

[56] Cooke, M., P. D. Green y M.D. Crawford: *Handling missing data in speech recognition*. En *Proc. ICSLP*, págs. 1555–1558, 1994. 84

[57] Cooke, M. *y cols.*: *Robust automatic speech recognition with missing and unreliable acoustic data*. Speech Commun., 34(3):267–285, Junio 2001. 22, 84, 86, 87, 89, 150, 274

[58] Cooke, M., J. R. Hershey y S. J. Rennie: *Monaural speech separation and recognition challenge*. Comput. Speech Lang., 24(1):1–15, Enero 2010. 93

[59] Cooke, M., A. Morris y P. D. Green: *Missing data techniques for robust speech recognition*. En *Proc. ICASSP*, págs. 863–866, 1997. 84, 85, 86

[60] Crystal, T. H. y A. S. House: *Segmental durations in connected-speech signals: Current results*. J. Aco, 83(4):1553–1573, Abril 1988. 242

[61] Cui, X., M. Afify y Y. Gao: *N-best based stochastic mapping on stereo HMM for noise robust speech recognition*. En *Proc. Interspeech*, págs. 1261–1264, 2008. 76

[62] Cui, X., M. Afify y B. Zhou: *Stereo-based stochastic mapping with context using probabilistic PCA for noise robust automatic speech recognition.* En *Proc. ICASSP*, págs. 4705–4708, 2012. 76

[63] Cunningham, S. y M. Cooke: *The role of evidence and counter-evidence in speech perception.* En *Proc. ICPh'99*, págs. 215–218, 1999. 90

[64] Davis, S. y P. Mermelstein: *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences.* IEEE Trans. Acoust., Speech, Signal Process., 28(4):357–366, 1980. 4, 17, 88

[65] Demange, S., C. Cerisara y J. P. Haton: *Missing data mask estimation with frequency and temporal dependencies.* Comput. Speech Lang., 23(1):25–41, Enero 2009. 91

[66] Dempster, A. P., N. M. Laird y D. B. Rubin: *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society, 39(1):1–38, 1977. 30, 69, 107, 109, 110, 140, 148, 160, 177, 178, 180, 225, 273

[67] Deng, L. *y cols.*: *High-performance robust speech recognition using stereo training data.* En *Proc. ICASSP*, págs. 301–304, 2001. 69, 225, 263

[68] Deng, L. *y cols.*: *Large-vocabulary speech recognition under adverse acoustic environments.* En *Proc. ICSLP*, págs. 806–809, 2000. 56, 69, 126, 225

[69] Deng, L., J. Droppo y A. Acero: *Log-domain speech feature enhancement using sequential MAP noise estimation and phase-sensitive model of the acoustic environment.* En *Proc. ICSLP*, págs. 1813–1816, 2002. 16, 78, 203

[70] Deng, L., J. Droppo y A. Acero: *Enhancement of log Mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise.* IEEE Trans. Speech Audio Process., 12(2):133–143, Marzo 2004. 16, 50, 51, 111

[71] Deng, L., J. Droppo y A. Acero: *Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features.* IEEE Trans. Speech Audio Process., 12(3):218–233, Mayo 2004. 16, 51, 120

[72] Deng, L., J. Droppo y A. Acero: *Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion.* IEEE Trans. Speech Audio Process., 13(3):412–421, Mayo 2005. 78, 203

[73] Dharanipragada, S. y M. Padmanabhan: *A nonlinear unsupervised adaptation technique for speech recognition*. En *Proc. ICSLP*, págs. 556–559, 2000. 62

[74] Dhrymes, Phoebus J.: *Moments of truncated (normal) distributions*, 2005. 270, 271

[75] Digalakis, V. V., D. Rtischev y L. G. Neumeyer: *Speaker adaptation using constrained estimation of Gaussian mixtures*. IEEE Trans. Speech Audio Process., 3(5):357–366, 1995. 35

[76] Donoho, D.: *Compressed sensing*. IEEE Trans. Inform. Theory, 52(4):1289–1306, Abril 2006. 99

[77] Droppo, J. y A. Acero: *Maximum mutual information SPLICE transform for seen and unseen conditions*. En *Proc. Interspeech*, págs. 989–992, 2005. 71

[78] Droppo, J., A. Acero y L. Deng: *Uncertainty decoding with SPLICE for noise robust speech recognition*. En *Proc. ICASSP*, págs. 57–60, 2002. 69, 80

[79] Droppo, J., L. Deng y A Acero: *Evaluation of SPLICE on the Aurora 2 and 3 tasks*. En *Proc. ICSLP*, págs. 29–32, 2002. 69, 126, 138

[80] Ellis, D. y M. Gómez: *Investigations into tandem acoustic modeing for the Aurora task*. En *Proc. Eurospeech*, 2001. 4

[81] Ephraim, Y. y D. Malah: *Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator*. IEEE Trans. Acoust., Speech, Signal Process., 32(6):1109–1121, Diciembre 1984. 65, 119

[82] Ephraim, Y. y D. Malah: *Speech enhancement using a minimum mean-square error log-spectral amplitude estimator*. IEEE Trans. Acoust., Speech, Signal Process., 33(2):443–445, Abril 1985. 65, 119

[83] Ephraim, Y., D. Malah y B. H. Juang: *On the application of hidden Markov models for enhancing noisy speech*. IEEE Trans. Acoust., Speech, Signal Process., 37(12):1846–1856, Diciembre 1989. 65, 119, 191

[84] Evermann, G. *y cols.*: *Development of the 2003 CU-HTK conversational telephone speech transcription system*. En *Proc. ICASSP*, 2004. 2

[85] Fastl, H. y E. Zwicker: *Psychoacoustics - Facts and models*. Springer, 2006. 82

[86] Faubel, F. y D. Klakow: *Estimating noise from noisy speech features with a Monte Carlo variant of the expectation maximization algorithm.* En *Proc. Interspeech*, págs. 2046–2049, 2010. 113

[87] Faubel, F., J. McDonough y D. Klakow: *A phase-averaged model for the relationship between noisy speech, clean speech and noise in the log-Mel domain.* En *Proc. Interspeech*, págs. 553–556, 2008. 16

[88] Faubel, F., J. McDonough y D. Klakow: *Bounded conditional mean imputation with Gaussian mixture models: A reconstruction approach to partly occluded features.* En *Proc. ICASSP*, págs. 3869–3872, 2009. 96, 97, 98, 182

[89] Faubel, F., J. McDonough y D. Klakow: *On expectation maximization based channel and noise estimation beyond the vector Taylor series expansion.* En *Proc. ICASSP*, págs. 4294–4297, 2010. 113

[90] Faubel, F. *y cols.*: *Particle filter based soft-mask estimation for missing feature reconstruction.* En *Proc. IWAENC*, Septiembre 2008. 85, 102, 146, 160, 182, 185

[91] Faubel, F. y M. Wolfel: *Overcoming the vector Taylor series approximation in speech feature enhancement - A particle filter approach.* En *Proc. ICASSP*, págs. 557–560, 2007. 66, 113

[92] Fingscheidt, T. y P. Vary: *Softbit speech decoding: A new approach to error concealment.* IEEE Trans. Speech Audio Process., 9(3):240–251, Marzo 2001. 191

[93] Fletcher, H.: *Speech and hearing in communication.* New York: Van Nostrand Co., 1953. 82, 191

[94] Frey, B. *y cols.*: *Algonquin: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition.* En *Proc. Eurospeech*, 2001. 50, 66, 111, 189

[95] Frey, B. *y cols.*: *Learning dynamic noise models from noisy speech for robust speech recognition.* En *Proc. Advances in Neural Information Processing Systems (NIPS)*, págs. 101–108, 2001. 50, 107, 189

[96] Fujimoto, M. y S. Nakamura: *Particle pilter based non-stationary noise tracking for robust speech recognition.* En *Proc. ICASSP*, págs. 257–260, 2005. 113

[97] Gales, M. J. F.: *Model-based techniques for noise robust speech recognition.* Tesis de Doctorado, University of Cambridge, 1995. 40, 42, 43, 49, 162

[98] Gales, M. J. F.: *The generation and use of regression class trees for MLLR adaptation*. Informe técnico CUED/F-INFEG/TR263, Cambridge Universiy Engineering Department, 1996. 32

[99] Gales, M. J. F.: *Maximum likelihood linear transformations for HMM-based speech recognition*. Comput. Speech Lang., 12(2):2, Abril 1998. 33, 34, 35

[100] Gales, M. J. F.: *Cluster adaptive training of hidden Markov models*. IEEE Trans. Speech Audio Process., 8(4):417–428, 2000. 56, 57

[101] Gales, M. J. F.: *Acoustic factorisation*. En *Proc. ASRU*, 2001. 56

[102] Gales, M. J. F.: *Adaptive training for robust ASR*. En *Proc. ASRU*, págs. 15–20, 2001. 28, 29, 55

[103] Gales, M. J. F. y S. J. Young: *Cepstral parameter compensation for HMM recognition in noise*. Speech Commun., 12(3):231–239, Julio 1993. 40, 162

[104] Gales, M. J. F. y S. J. Young: *Mean and variance adaptation within the MLLR framework*. Comput. Speech Lang., 10:249–264, 1996. 34

[105] Gales, M. J. F. y S. J. Young: *Robust continuous speech recognition using parallel model combination*. IEEE Trans. Speech Audio Process., 4(5):352–359, Septiembre 1996. 40, 162

[106] Garcia, L. *y cols.*: *Class-based parametric approximation to histogram equalization for ASR*. IEEE Signal Process. Lett., 19(7):415–418, Julio 2012. 62, 63

[107] Garcia, L. *y cols.*: *Parametric nonlinear feature equalization for robust speech recognition*. En *Proc. ICASSP*, págs. 529–532, 2006. 63

[108] García-Martínez, L.: *Ecualización de histogramas en el procesado robusto de voz*. Tesis de Doctorado, University of Granada, 2007. 59, 60, 62

[109] Gauvain, J. L. y C. H. Lee: *Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains*. IEEE Trans. Speech Audio Process., 2(2):291–298, Abril 1994. 23, 30, 31

[110] Geller, T.: *Talking to machines*. Commun. ACM, 55(4):14–16, Abril 2012.

[111] Gemmeke, J. F. y B. Cranen: *Sparse imputation for noise robust speech recognition using soft masks*. En *Proc. ICASSP*, págs. 4645–4648, 2009. 101

[112] Gemmeke, J. F., B. Cranen y U. Remes: *Sparse imputation for large vocabulary noise robust ASR.* Comput. Speech Lang., 25(2):462–479, Abril 2011. 101

[113] Gemmeke, J. F., U. Remes y K. J. Palömaki: *Observation uncertainty measures a for sparse imputation.* En *Proc. Interspeech*, págs. 2262–2265, 2010. 94, 101

[114] Gemmeke, J. F. *y cols.*: *Compressive sensing for missing data imputation in noise robust speech recognition.* IEEE J. Sel. Topics Signal Process., 4(2):272–287, 2010. 100, 101

[115] Genz, A.: *Numerical computation of multivariate normal probabilities.* J. Comp. Graph. Stat., 1(2):141–149, Junio 1992. 97

[116] Gerkmann, T., C. Breithaupt y R. Martin: *Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors.* IEEE Trans. Audio Speech Lang. Process., 16(5):910–919, Julio 2008. 103, 106

[117] Gillick, L. y S. J. Cox: *Some statistical issues in the comparison of speech recognition algorithms.* En *Proc. ICASSP*, págs. 532–535, 1989. 219

[118] Gómez, A. M.: *Tratamiento de la degradación debida al canal en sistemas de reconocimiento remoto.* Tesis de Doctorado, Universidad de Granada, 2006. 15, 191

[119] Gómez, A. M. *y cols.*: *Combining media-specific FEC and error concealment for robust distributed speech recognition over loss-prone packet channels.* IEEE Trans. Multimedia, 8(6):1228–1238, Diciembre 2006. 205

[120] Gong, Y.: *Speech recognition in noisy environments: A survey.* Speech Commun., 16(3):261–291, Abril 1995. 2, 13, 21, 24, 143

[121] González, J. A., A. M. Peinado y A. M. Gómez: *MMSE feature reconstruction based on an occlusion model for robust ASR.* En *Advances in speech and language technologies for Iberian languages - IberSPEECH 2012*, Communications in Computer and Information Science, págs. 217–226. Springer, 2012. 146, 159, 260, 285, 286

[122] González, J. A. *y cols.*: *Efficient MMSE estimation and uncertainty processing for multienvironment robust speech recognition.* IEEE Trans. Audio Speech Lang. Process., 19(5):1206–1220, Julio 2011. 69, 78, 143, 191, 198, 202, 203, 205, 235, 244, 260, 285, 286

[123] González, J. A. *y cols.*: *Efficient VQ-based MMSE estimation for robust speech recognition.* En *Proc. ICASSP*, págs. 4558 – 4561, Marzo 2010. 143, 260, 285

[124] González, J. A. *y cols.*: *A feature compensation approach using VQ-based MMSE estimation for robust speech recognition.* En *VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop (FALA)*, págs. 111–114, 2010. 260, 285

[125] González, J. A. *y cols.*: *Log-spectral feature reconstruction based on an occlusion model for noise robust speech recognition.* En *Proc. Interspeech*, 2012. 85, 107, 146, 159, 260, 285, 286

[126] González, J. A. *y cols.*: *Combining missing-data reconstruction and uncertainty decoding for robust speech recognition.* En *Proc. ICASSP*, págs. 4693–4696, 2012. 78, 94, 107, 159, 182, 203, 205, 260, 285, 286

[127] González, J. A. *y cols.*: *MMSE-based missing-feature reconstruction with temporal modeling for robust speech recognition.* IEEE Trans. Audio Speech Lang. Process., 21(3):624–635, Marzo 2013. 159, 191, 198, 260, 285, 286

[128] Gopinath, R. A. *y cols.*: *Robust speech recognition in noise - performance of the IBM continuous speech recognizer on the ARPA noise spoke task.* En *Proc. of the ARPA Workshop on Spoken Language System Technology*, págs. 127–130, 1995. 49

[129] Green, P. D., M. Cooke y M. D. Crawford: *Auditory scene analysis and HMM recognition of speech in noise.* En *Proc. ICASSP*, págs. 401–404, 1995. 84

[130] Griffin, D. W. y J. S. Lim: *Signal estimation from modified short-time Fourier transform.* IEEE Trans. Acoust., Speech, Signal Process., 32(2):232–243, Abril 1984. 151

[131] Haeb-Umbach, R.: *Automatic generation of phonetic regression class trees for MLLR adaptation.* IEEE Trans. Audio Speech Lang. Process., 9(3):299–302, 2001. 32, 33

[132] Hansen, J. H. L.: *Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition.* Speech Commun., 20(2):151–170, Noviembre 1996. xv, 14

[133] Harding, S., J. Barker y G. J. Brown: *Mask estimation for missing data speech recognition based on statistics of binaural interaction.* IEEE Trans. Audio Speech Lang. Process., 14(1):58–67, Enero 2006. 92, 106

[134] Hermansky, H.: *Perceptual linear predictive (PLP) analysis for speech.* J. Acoust. Soc. Am., 87(4):1738–1752, 1990. 4, 26

[135] Hermansky, H. y N. Morgan: *RASTA processing of speech.* IEEE Trans. Speech Audio Process., 2(4):578–589, 1994. 26

[136] Hermansky, H. *y cols.*: *Compensation for the effect of communication channel in auditory-like analysis of speech (RASTA-PLP).* En *Proc. EuroSpeech*, págs. 1367–1370, 1991. 26

[137] Hernando, J. y C. Nadeu: *A comparative study of parameters and distances for noisy speech recognition.* En *Proc. Eurospeech*, págs. 91–94, 1991. 24

[138] Hernando, J., C. Nadeu y E. LLeida: *On the AR modelling of the one-sided autocorrelation sequence for noisy speech recognition.* En *Proc. ICSLP*, págs. 1593–1596, Octubre 1992. 24

[139] Hilger, F. y H. Ney: *Quantile based histogram equalization for noise robust large vocabulary speech recognition.* IEEE Trans. Audio Speech Lang. Process., 14(3):845–854, Mayo 2006. 62, 63

[140] Hirsch, H. G.: *Experimental framework for the performance evaluation of speech recognition front-ends of large vocabulary task.* Informe técnico, STQ AURORA DSR Working Group, 2002. 12, 103, 127, 210

[141] Hirsch, H. G. y D. Pearce: *The Aurora experimental framework for the performance evaluations of speech recognitions systems under noise conditions.* En *Proc. ISCA ITRW ASR 2000*, Noviembre 2000. 12, 29, 68, 103, 107, 127, 145, 158, 171, 210

[142] Holmes, J. N., W. J. Holmes y P. N. Garner: *Using formant frequencies in speech recognition.* En *Proc. Eurospeech*, págs. 2083–2087, 1997. 203

[143] Hsu, C W. y L S. Lee: *Higher order cepstral moment normalization for improved robust speech recognition.* IEEE Trans. Audio Speech Lang. Process., 17(2):205–220, Febrero 2009. 62

[144] Hu, G. y D. Wang: *Separation of fricatives and affricates.* En *Proc. ICASSP*, págs. 1101–1104, 2005. 106

[145] Hu, G. y D. Wang: *Auditory segmentation based on onset and offset analysis.* IEEE Trans. Audio Speech Lang. Process., 15(2):396–405, 2007. 106

[146] Hu, Y. y Q. Huo: *Irrelevant variability normalization based HMM training using VTS approximation of an explicit model of environmental distortions.* En *Proc. Interspeech*, págs. 1042–1045, 2007. 56

[147] Huang, X., A. Acero y H. Hon: *Spoken language processing: A guide to theory, algorithm, and system development.* Prentice Hall, 2001. 3, 37

[148] Ion, V. y R. Haeb-Umbach: *A novel uncertainty decoding rule with applications to transmission error robust speech recognition.* IEEE Trans. Audio Speech Lang. Process., 16(5):1047–1060, 2008. 15, 191

[149] Johnson, N. L., S. Kotz y N. Balakrishnan: *Continuous univariate distributions*, volumen 1. Wiley, 1994. 269, 270, 271

[150] Jolliffe, I. T.: *Principal component analysis.* Springer-Verlag New York, 2002. 194

[151] Julier, S. J. y J. K. Uhlmann: *Unscented filtering and nonlinear estimation.* Proc. IEEE, 92(3):401–422, 2004. 113

[152] Junqua, J. C.: *The Lombard reflex and its role on human listeners and automatic speech recognizers.* J. Acoust. Soc. Am., 93(1):510–524, 1993. 14, 122

[153] Junqua, J. C. y Y. Anglade: *Acoustic and perceptual studies of Lombard speech: application to isolated-words automatic speech recognition.* En *Proc. ICASSP*, págs. 841–844, 1990. 14, 122

[154] Kalinli, O., M. L. Seltzer y A. Acero: *Noise adaptive training using a vector Taylor series approach for noise robust automatic speech recognition.* En *Proc. ICASSP*, págs. 3825–3828, 2009. 56, 58, 110

[155] Kalinli, O. *y cols.*: *Noise adaptive training for robust automatic speech recognition.* IEEE Trans. Audio Speech Lang. Process., 18(8):1889–1901, Noviembre 2010. 56, 57, 58

[156] Kay, S. M.: *Fundamentals of statistical signal processing: Estimation theory*, volumen 1. Prentice Hall, 1993. 120

[157] Kim, C. y R. M. Stern: *Power-normalized cepstral coefficients (PNCC) for robust speech recognition.* En *Proc. ICASSP*, págs. 4101–4104, 2012. 26, 109

[158] Kim, C. y R. M. Stern: *Power-normalized cepstral coefficients (PNCC) for robust speech recognition.* IEEE Trans. Audio Speech Lang. Process., accepted. 4, 26

[159] Kim, D. Y., C. K. Un y N. S. Kim: *Speech recognition in noisy environments using first-order vector Taylor series.* Speech Commun., 24(1):39–49, Abril 1998. 50, 111

[160] Kim, W. y J. H. L. Hansen: *Feature compensation in the cepstral domain employing model combination.* Speech Commun., 51(2):83–96, Febrero 2009, ISSN 0167-6393. 137, 138

[161] Kim, W. y J. H. L. Hansen: *Missing-feature reconstruction by leveraging temporal spectral correlation for robust speech recognition in background noise conditions.* IEEE Trans. Audio Speech Lang. Process., 18(8):2111–2120, 2010. 191, 242

[162] Kim, W. y R. Stern: *Mask classification for missing-feature reconstruction for robust speech recognition in unknown background noise.* Speech Commun., 53(1):1–11, Enero 2011. 105

[163] Koehler, J. *y cols.*: *Integrating RASTA-PLP into speech recognition.* En *Proc. ICASSP*, págs. 421–424, 1994. 26

[164] Kolossa, D. y R. Haeb-Umbach (editores): *Robust speech recognition of uncertain or missing data: Theory and applications.* Springer-Verlag Berlin, 2011. 16, 51

[165] Kristjansson, T. T.: *Speech recognition in adverse environments: A probabilistic approach.* Tesis de Doctorado, University of Waterloo, 2002. 23, 50, 66, 68, 69, 189

[166] Kristjansson, T. T. *y cols.*: *Joint estimation of noise and channel distortion in a generalized EM framework.* En *Proc. ASRU*, 2001. 50, 107, 111

[167] Leggetter, C. J.: *Improved acoustic modelling for HMMs using linear transformations.* Tesis de Doctorado, Cambridge University, 1995. 32

[168] Leggetter, C. J. y P. C. Woodland: *Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models.* Comput. Speech Lang., 9(2):171–185, Abril 1995. 23, 33

[169] Leonard, R.: *A database for speaker-independent digit recognition.* En *Proc. ICASSP*, págs. 328–331, 1984. 211

[170] Leutnant, V. y R. Haeb-Umbach: *An analytic derivation of a phase-sensitive observation model for noise robust speech recognition.* En *Proc. Interspeech*, págs. 2395–2398, 2009. 16

[171] Li, J. *y cols.*: *High-performance HMM adaptation with joint compensation of additive and convolutive distortions via vector Taylor Series.* En *Proc. ASRU*, págs. 65–70, 2007. 50, 111

[172] Li, J. *y cols.*: *HMM adaptation using a phase-sensitive acoustic distortion model for environment-robust speech recognition.* En *Proc. ICASSP*, págs. 4069–4072, 2008. 189

[173] Li, J. *y cols.*: *Unscented transform with online distortion estimation for HMM adaptation.* En *Proc. Interspeech*, págs. 1660–1663, 2012. 66

[174] Li, Y. *y cols.*: *Incremental on-line feature space MLLR adaptation for telephony speech recognition.* En *Proc. ICSLP*, 2002. 35

[175] Liao, H.: *Uncertainty decoding for noise robust speech recognition.* Tesis de Doctorado, University of Cambridge, 2007. 19, 49, 76, 80, 110, 111

[176] Liao, H. y M. J. F. Gales: *Adaptive training with joint uncertainty decoding for robust recognition of noisy data.* En *Proc. ICASSP*, volumen 4, págs. 389–392, 2007. 56, 110, 113

[177] Liao, H. y M. J. F. Gales: *Issues with uncertainty decoding for noise robust automatic speech recognition.* Speech Commun., 50(4):265–277, 2008. 76, 80, 202

[178] Linde, Y., A. Buzo y R. Gray: *An algorithm for vector quantizer design.* IEEE Trans. Commun., 28(1):84–95, Enero 1980. 126, 140, 180, 224

[179] Lippmann, R. P.: *Speech recognition by machines and humans.* Speech Commun., 22(1):1–15, Julio 1997. 2

[180] Lippmann, R. P., E. A. Martin y D. B. Paul: *Multi-style training for robust isolated-word speech recognition.* En *Proc. ICASSP*, págs. 705–708, 1987. 29

[181] Lockwood, P. y J. Boudy: *Experiments with a nonlinear spectral subtractor (NSS), Hidden Markov models and the projection, for robust speech recognition in cars.* Speech Commun., 11(2-3):215–228, Junio 1992. 65

[182] Lu, Y. y P. C. Loizou: *A geometric approach to spectral subtraction.* Speech Commun., 50(6):453–466, 2008. 108

[183] Ma, N.: *Informing multisource decoding in robust automatic speech recognition.* Tesis de Doctorado, The University of Sheffield, 2008. 24, 81, 105

[184] Ma, N. y J. Barker: *Coupling identification and reconstruction of missing features for noise-robust automatic speech recognition.* En *Proc. Interspeech*, 2012. 106, 150, 197

[185] Ma, N. *y cols.*: *Combining speech fragment decoding and adaptive noise floor modelling.* IEEE Trans. Audio Speech Lang. Process., 20(3):818–827, Marzo 2012. 108, 109

[186] Ma, N. *y cols.*: *Exploiting correlogram structure for robust speech recognition with multiple speech sources.* Speech Commun., 49(12):874–891, 2007. 92, 105

[187] Macho, D.: *Spanish SDC-Aurora database for ETSI STQ Aurora WI008 advanced DSR front-end evaluation: Description and baseline results.* Informe técnico, STQ Aurora Working Group, 2000. 127

[188] Makhoul, J.: *Linear prediction: A tutorial review.* Proc. IEEE, 63(4):561–580, 1975. 4

[189] Mansour, D. y B. H. Juang: *The short-time modified coherence representation and its application for noisy speech recognition.* IEEE Trans. Acoust., Speech, Signal Process., 37(6):795–804, Junio 1989. 24

[190] Martin, R.: *Noise power spectral density estimation based on optimal smoothing and minimum statistics.* IEEE Trans. Speech Audio Process., 9(5):504–512, Julio 2001. 108, 109

[191] Milner, B. y A. James: *Robust speech recognition over mobile and IP networks in burst-like packet loss.* IEEE Trans. Audio Speech Lang. Process., 14(1):223–231, Enero 2006. 15

[192] Moore, B. C. J. y B. R. Glasberg: *A revision of Zwicker's loudness model.* Acta Acustica, 82(2):335–345, Abril 1996. 26

[193] Moore, B. J. C.: *An introduction to the psychology of hearing.* Academic Press, 1982. 82

[194] Moore, R. K.: *Spoken language processing: Piecing together the puzzle.* Speech Commun., 49(5):418–435, Enero 2007. 2

[195] Morales-Cordovilla, J. A.: *Técnicas de reconocimiento robusto de la voz basadas en el pitch.* Tesis de Doctorado, Universidad de Granada, 2011. 25, 108, 109

[196] Morales-Cordovilla, J. A. *y cols.*: *A pitch based noise estimation technique for robust speech recognition with missing data.* En *Proc. ICASSP*, págs. 4808–4811, Mayo 2011. 108

[197] Morales-Cordovilla, J. A. *y cols.*: *Feature extraction based on pitch-synchronous averaging for robust speech recognition.* IEEE Trans. Audio Speech Lang. Process., 19(3):640–651, Marzo 2011. 25

[198] Morales-Cordovilla, J. A. *y cols.*: *On the use of asymmetric windows for robust speech recognition.* Circuits, Syst. and Signal Process., 31(2):1–10, Septiembre 2011. 25

[199] Moreno, P. J.: *Speech recognition in noisy environments.* Tesis de Doctorado, Carnegie Mellon University, 1996. 19, 23, 66, 67, 68, 69, 73, 111, 123, 146, 182, 185, 189

[200] Moreno, P. J. *y cols.*: *Multivariate-Gaussian-based cepstral normalization for robust speech recognition.* En *Proc. ICASSP*, págs. 137–140, 1995. 73

[201] Morgan, N. y H. Hermansky: *RASTA extensions: Robustness to additive and convolutional noise.* En *Proc. ESCA Workshop on Speech Processing in Adverse Conditions*, págs. 115–118, 1992. 26

[202] Morgan, N. *y cols.*: *Pushing the spectral envelope aside: Beyond the spectral envelope as the fundamental representation for speech recognition.* IEEE Signal Process. Mag., 22(5):81–88, Septiembre 2005. 2

[203] Morris, A. C., J. Barker y H. Bourlard: *From missing data to maybe useful data: Soft data modelling for noise robust ASR.* En *Proc. of Workshop on Innovation in Speech Proccesing (WISP)*, 2001. 202

[204] Morris, A. C., M. P. Cooke y P. D. Green: *Some solutions to the missing feature problem in data classification, with application to noise robust ASR.* En *Proc. ICASSP*, págs. 737–740, 1998. 84, 202

[205] Murthi, M. N. y B. D. Rao: *All-pole modeling of speech based on the minimum variance distortionless response spectrum.* IEEE Trans. Audio Speech Lang. Process., 8(3):221–239, Mayo 2000. 27

[206] Nádas, A., D. Nahamoo y M. A. Picheny: *Speech recognition using noise-adaptive prototypes.* IEEE Trans. Acoust., Speech, Signal Process., 37(10):1495–1503, Octubre 1989. 145, 146

[207] Nakatani, T. *y cols.*: *Logmax observation model with MFCC-based spectral prior for reduction of highly nonstationary ambient noise.* En *Proc. ICASSP*, págs. 4029–4032, 2012. 58, 85, 145

[208] Neumeyer, L., A. Sankar y V. Digalakis: *A comparative study of speaker adaptation techniques.* En *Proc. Eurospeech*, págs. 417–420, 1995. 34

[209] Neumeyer, L. y M. Weintraub: *Probabilistic optimum filtering for robust speech recognition.* En *Proc. ICASSP*, págs. 417–420, 1994. 72

[210] Palomäki, K. J., G. J. Brown y J. Barker: *Techniques for handling convolutional distortion with 'missing data' automatic speech recognition.* Speech Commun., 43(1–2):123–142, Junio 2004. 86

[211] Papoulis, A. y S. Unnikrishna Pillai: *Probability, random variables and stochastic processes.* McGraw Hill, 2002. 53, 77, 154

[212] Patterson, R. D. *y cols.*: *Complex sounds and auditory images.* En *Proc. 9th International Symposium on Hearing*, págs. 429–446, 1992. 26

[213] Peinado, A. M. *y cols.*: *An integrated solution for error concealment in DSR systems over wireless channels.* En *Proc. Interspeech*, págs. 1093–1096, 2006. 205

[214] Peinado, A. M. *y cols.*: *HMM-based channel error mitigation and its application to distributed speech recognition.* Speech Commun., 41(4):549–561, Noviembre 2003. 191, 198

[215] Peinado, A. M. y J. C. Segura: *Speech recognition over digital channels: Robustness and standars.* Wiley, Julio 2006. 3, 15, 21, 61, 64, 129, 202

[216] Petersen, K. B. y M. S. Pedersen: *The matrix cookbook.* Technical University of Denmark, 2008. 46, 135, 275

[217] Pohlmann, K. C.: *Principles of digital audio.* McGraw Hill, 2005. 82

[218] Povey, D.: *Discriminative training for large vocabulary speech recognition.* Tesis de Doctorado, University of Cambridge, 2003. 8

[219] Povey, D. y K. Yao: *A basis representation of constrained MLLR transforms for robust adaptation.* Comput. Speech Lang., 26(1):35–51, Enero 2012. 35

[220] Rabiner, L. R.: *A tutorial on hidden Markov models and selected applications in speech recognition.* Proc. IEEE, 77(2):257–286, Febrero 1989. 30, 111, 197, 198

[221] Rabiner, L. R. y B. H. Juang: *Fundamentals of speech recognition*. Prentice-Hall, 1993. 3

[222] Radfar, M. H. *y cols.*: *Nonlinear minimum mean square error estimator for mixture-maximisation approximation*. Electron. Lett., 42(12):724–725, Junio 2006. 146

[223] Raj, B.: *Reconstruction of incomplete spectrograms for robust speech recognition*. Tesis de Doctorado, Carnegie Mellon University, 2000. 149

[224] Raj, B., M. L. Seltzer y R. M. Stern: *Reconstruction of missing features for robust speech recognition*. Speech Commun., 48(4):275–296, 2004. 91, 94, 95, 96, 159, 182

[225] Raj, B. y R. Singh: *Reconstructing spectral vectors with uncertain spectrographic masks for robust speech recognition*. En *Proc. ASRU*, págs. 65–70, 2005. 102, 160, 185, 246

[226] Raj, B. y R. M. Stern: *Missing-feature approaches in speech recognition*. IEEE Signal Process. Mag., 22(5):101–116, 2005. 22

[227] Ramírez, J. *y cols.*: *An effective subband OSF-based VAD with noise reduction for robust speech recognition*. IEEE Trans. Speech Audio Process., 13(6):1119–1129, Noviembre 2005. 107

[228] Raut, C. K.: *Discriminative adaptive training and Bayesian inference for speech recognition*. Tesis de Doctorado, University of Cambridge, 2009. 55

[229] Rennie, S. J., J. R. Hershey y P. A. Olsen: *Single-channel multitalker speech recognition*. IEEE Signal Process. Mag., 27(6):66–80, Noviembre 2010. 39, 146, 162, 164

[230] Reynolds, D. A.: *Gaussian mixture models*. Encyclopedia of Biometric Recognition, Febrero 2008. 89

[231] Rodbro, C. A. *y cols.*: *Hidden Markov model-based packet loss concealment for voice over IP*. IEEE Trans. Audio Speech Lang. Process., 14(5):1609–1623, Septiembre 2006. 191

[232] Roweis, S. T.: *One microphone source separation*. En *Proc. Neural Information Processing Systems (NIPS)*, págs. 793–799, 2001. 39, 162

[233] Roweis, S. T.: *Factorial models and refiltering for speech separation and denoising.* En *Proc. Eurospeech*, págs. 1009–1012, 2003. 39, 85, 145, 162

[234] Saon, G., G. Zweig y M. Padmanabhan: *Linear feature space projections for speaker adaptation.* En *Proc. ICASSP*, 2001. 35

[235] Saz, O. *y cols.*: *Algoritmos de compensación de características cepstrales para reconocimiento automático del habla robusto.* En *Actas de las III Jornadas en Tecnología del Habla*, volumen 9–14, 2004. 74

[236] Scharenborg, O.: *Reaching over the gap: A review of efforts to link human and automatic speech recognition research.* Speech Commun., 49(5):336–347, Mayo 2007. 2

[237] Segura, J. C. *y cols.*: *VTS residual noise compensation.* En *Proc. ICASSP*, págs. 409–412, 2002. 107

[238] Segura, J. C. *y cols.*: *Cepstral domain segmental nonlinear feature transformations for robust speech recognition.* IEEE Signal Process. Lett., 11(5):517–520, Mayo 2004. 60, 62, 63, 64, 144, 220

[239] Segura, J. C. *y cols.*: *Model-based compensation of the additive noise for continuous speech recognition. Experiments using the Aurora II database and tasks.* En *Proc. Eurospeech*, págs. 221–224, 2001. 66, 67, 107, 182, 246

[240] Seide, Frank y Pei Zhao: *On using missing-feature theory with cepstral features – Approximations to the multivariate integral.* En *Proc. Interspeech*, págs. 2094–2097, Septiembre 2010. 97

[241] Seltzer, M. L., B. Raj y R. M. Stern: *A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition.* Speech Commun., 43(4):379–393, Septiembre 2004, ISSN 0167-6393. 105

[242] Shannon, B. J. y K. K. Paliwal: *Feature extraction from higher-lag autocorrelation coefficients for robust speech recognition.* Speech Commun., 48(1):1458–1485, 2006. 25

[243] Shi, G., Y. Shi y Q. Huo: *A study of irrelevant variability normalization based training and unsupervised online adaptation for LVCSR.* En *Proc. Interspeech*, págs. 1357–1360, 2010. 56

[244] Shinoda, K. y C. H. Lee: *Structural MAP speaker adaptation using hierarchical priors.* En *Proc. ASRU*, págs. 381–388, 1997. 31

[245] Shinoda, K. y T. Watanabe: *Speaker adaptation with autonomous control using tree structure*. En *Proc. Eurospeech*, 1995. 33

[246] Siohan, O., C. Chesta y C. H. Lee: *Joint maximum a posteriori adaptation of transformation and HMM parameters*. IEEE Trans. Audio Speech Lang. Process., 9(4):417–428, 2001. 36

[247] Sohn, J., N. S. Kim y W. Sung: *A statistical model-based voice activity detection*. IEEE Signal Process. Lett., 6(1):1–3, Enero 1999. 107

[248] Srinivasan, S. y D. Wang: *Transforming binary uncertainties for robust speech recognition*. IEEE Trans. Audio Speech Lang. Process., 15(7):2130–2140, 2007. 94

[249] Srinivasan, S. y D. Wang: *Robust speech recognition by integrating speech separation and hypothesis testing*. Speech Commun., 52(1):72–81, 2010. 104, 159

[250] Stouten, V. *y cols.*: *Robust speech recognition using model-based feature enhancement*. En *Proc. Eurospeech*, 2003. 107

[251] Stouten, V., H. Van Hamme y P. Wambacq: *Effect of phase-sensitive environment model and higher order VTS on noisy speech feature enhancement*. En *Proc. ICASSP*, volumen 1, págs. 433–436, 2005. 16

[252] Stouten, V., H. Van Hamme y P. Wambacq: *Model-based feature enhancement with uncertainty decoding for noise robust ASR*. Speech Commun., 48(11):1502–1514, Noviembre 2006, ISSN 0167-6393. 66, 67, 78, 191, 202, 203

[253] Streeter, L. A. *y cols.*: *Acoustic and perceptual indicators of emotional stress*. J. Aco. Soc. Am., 73(4):1354–1360, Abril 1983. 14

[254] Suh, Y., J. Mikyong y K. Hoirin: *Probabilistic class histogram equalization for robust speech recognition*. IEEE Signal Process. Lett., 14(4):287–290, Abril 2007. 63

[255] Suk, Y. H., S. H. Choi y H. S. Lee: *Cepstrum third-order normalisation method for noisy speech recognition*. Electronics Letters, 35(7):527–528, Enero 1999. 62

[256] Tan, Q. F., P. G. Georgiou y S. Narayanan: *Enhanced sparse imputation techniques for a robust speech recognition front-end*. IEEE Trans. Audio Speech Lang. Process., 19(8):2418–2429, Noviembre 2011. 100

[257] Tanyer, S. G. y H. Özer: *Voice activity detection in nonstationary noise.* IEEE Trans. Speech Audio Process., 8(4):478–482, Julio 2000. 107

[258] Torre, A. de la, A. M. Peinado y A. J. Rubio: *Reconocimiento automático de la voz en condiciones de ruido.* Monografías del Dpto. de Electrónica nº 47, 2001. 13, 19, 21, 59, 62, 66, 67, 107

[259] Torre, A. de la *y cols.*: *Histogram equalization of speech representation for robust speech recognition.* IEEE Trans. Speech Audio Process., 13(3):355–366, Mayo 2005. 59, 62, 144

[260] Van Dalen, R. C.: *Statistical models for noise-robust speech recognition.* Tesis de Doctorado, University of Cambridge, 2011. 53

[261] Van Hamme, H.: *Robust speech recognition using missing feature theory in the cepstral or LDA domain.* En *Proc. Eurospeech*, págs. 3089–3092, 2003. 98

[262] Van Hamme, H.: *PROSPECT features and their application to missing data techniques for robust speech recognition.* En *Proc. Interspeech*, págs. 101–104, 2004. 98

[263] Van Hamme, H.: *Robust speech recognition using cepstral domain missing data techniques and noisy masks.* En *Proc. ICASSP*, págs. 213–216, 2004. 98

[264] Van Hout, J. y A. Alwan: *A novel approach to soft-mask estimation and log-spectral enhancement for robust speech recognition.* En *Proc. ICASSP*, 2012. 108

[265] Van Segbroeck, M. y H. Van Hamme: *Advances in missing feature techniques for robust large-vocabulary continuous speech recognition.* IEEE Trans. Audio Speech Lang. Process., 19(1):123–137, Enero 2011. 86, 98, 99

[266] Varadarajan, B., D. Povey y S. M. Chu: *Quick FMLLR for speaker adaptation in speech recognition.* En *Proc. ICASSP*, 2008. 36

[267] Varga, A. P. y R. K. Moore: *Hidden Markov model decomposition of speech and noise.* En *Proc. ICASSP*, págs. 845–848, 1990. 38, 40, 85, 145, 162, 186

[268] Viikki, O. y K. Laurila: *Cepstral domain segmental feature vector normalization for noise robust speech recognition.* Speech Commun., 25(1–3):133–147, Agosto 1998. 61

[269] Viterbi, A.: *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.* IEEE Trans. Inf. Theory, 13(2):260–269, Abril 1967. 9

[270] Vizinho, A. *y cols.*: *Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR: An integrated study.* En *Proc. Eurospeech*, 1999. 84

[271] Wang, D. y G. J. Brown (editores): *Computational auditory scene analysis: Principles, algorithms, and applications.* IEEE Press/Wiley-Interscience, 2006. 92, 106, 184

[272] Wang, Y. y M. J. F. Gales: *Speaker and noise factorization for robust speech recognition.* IEEE Trans. Audio Speech Lang. Process., 20(7):2149–2158, 2012. 56

[273] Warren, R. M., J. A. Bashford y P. W. Lenz: *Intellibility of bandpass speech.* J. Ac. Soc. Am., 108(9):1264–1268, 2000. 83, 191

[274] Warren, R. M. *y cols.*: *Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits.* Percept. Psychophys., 57(2):175–182, 1995. 83, 191

[275] Watanabe, T. y K. Shinoda: *Speech recognition using tree-structured probability density function.* En *Proc. ICSLP*, 1995. 33

[276] Wölfel, M. y J. McDonough: *Minimum variance distortionless response spectral estimation: Review and refinements.* IEEE Signal Process. Mag., 22(5):117–126, Septiembre 2005. 27

[277] Xu, H. *y cols.*: *Noise condition-dependent training based on noise classification and SNR estimation.* IEEE Trans. Audio Speech Lang. Process., 15(8):2431–2443, Nov. 2007. 137, 138

[278] Yoma, N. B., F. R. McInnes y M. A. Jack: *Weighted Viterbi algorithm and state duration modelling for speech recognition in noise.* En *Proc. ICASSP*, volumen 2, págs. 709–712, Mayo 1998. 205

[279] Young, S. *y cols.*: *The HTK Book - Version 3.4.* Cambridge University Engineering Department, Diciembre 2006. 49, 216

[280] Yu, D. *y cols.*: *Use of incrementally regulated discriminative margins in MCE training for speech recognition.* En *Proc. ICSLP*, 2006. 2

[281] Yu, K.: *Adaptive training for large vocabulary continuous speech recognition.* Tesis de Doctorado, University of Cambridge, 2006. 29, 55, 56, 57

[282] Zhu, D. y Q. Huo: *Irrelevant variability normalization based HMM training using MAP estimation of feature transforms for robust speech recognition.* En *Proc. ICASSP*, págs. 4717–4720, 2008. 56