Efficient MMSE Estimation and Uncertainty Processing for Multienvironment Robust Speech Recognition

José A. González*, Antonio M. Peinado, *Senior Member, IEEE*, Angel M. Gómez, and José L. Carmona

***Abstract*—This paper presents a feature compensation framework based on minimum mean square error (MMSE) estimation and stereo training data for robust speech recognition. In our proposal, we model the clean and noisy feature spaces in order to obtain clean feature estimates. However, unlike other well-known MMSE compensation methods such as SPLICE or MEMLIN, which model those spaces with Gaussian mixture models (GMMs), in our case every feature space is characterized by a set of prototype vectors which can be alternatively considered as a vector quantization (VQ) codebook. The discrete nature of this feature space characterization introduces two significative advantages. First, it allows the implementation of a very efficient MMSE estimator in terms of accuracy and computational cost. On the other hand, time correlations can be exploited by means of hidden Markov modeling (HMM). In addition, a novel subregion-based modeling is applied in order to accurately represent the transformation between the clean and noise domains. In order to deal with unknown environments, a multiple-model approach is also explored. Since this approach has been shown quite sensitive to incorrect environment classification, we adapt two uncertainty processing techniques, soft-data decoding and exponential weighting, to our estimation framework. As a result, environment miss-classifications are concealed, allowing a better performance under unknown environments. The experimental results on noisy digit recognition show a relative improvement of 87.93% in word accuracy regarding the baseline when clean acoustic models are used, while a 4.54% is achieved with multi-style trained models.**

***Index Terms*—Robust speech recognition, feature vector compensation, minimum mean square error estimation, stereo-data.**

I. Introduction

The performance of automatic speech recognition (ASR) systems seriously degrades as the mismatch between operating and training conditions increases. There are several sources of mismatch, such as additive background noise, channel distortion, various speakers, different accents, etc. [1]–[3]. In order to compensate this mismatch, a substantial effort has been made and many techniques have been developed to achieve real-world applications based on speech recognition technologies.

This work has been supported by an FPU grant from the Spanish Ministry of Science and Innovation and by project MEC-FEDER TEC2007-66600.

The authors are with the Department of Teoría de la Señal Telemática y Comunicaciones, Universidad de Granada, Spain (e-mail: joseangl@ugr.es; amp@ugr.es; amgg@ugr.es; maqueda@ugr.es). Tel. +34-958240845. Fax: +34-958242740.

The techniques for noise robust speech recognition can be roughly grouped into two major classes: model-domain and feature-domain approaches [1]. Model-domain techniques try to adapt the pre-trained acoustic models in order to better match the environmental testing conditions. The maximum a posteriori (MAP) estimation [4], maximum likelihood linear regression (MLLR) [5], and parallel model combination (PMC) [6] are examples of techniques belonging to this category. On the other hand, feature-domain methods focus on modifying or enhancing the input test data to be closer to the clean training condition or to be less sensitive to the variability introduced by the acoustic noise. In this category we can find solutions based on noise reduction, such as spectral subtraction [7], Wiener filtering [8], and Ephraim-Malah filtering [9]. Other feature-domain solutions compensate the extracted features in order to reduce the effect of the environmental noise. Well-known methods included in this last subcategory are cepstral mean normalization (CMN) [10], histogram equalization (HEQ) [11], [12], vector Taylor series (VTS) expansion [13], [14], and more recently a noise suppression algorithm based on the minimum mean square error (MMSE) criterion [15]. Although model-domain techniques are very flexible, they involve a large computational load that increases as the acoustic model size (number of Gaussians) in the recognizer grows. In contrast, feature-domain approaches do not introduce any modification on the recognizer and, therefore, are largely independent of the acoustic model complexity. Also, they have the advantage that can be seamlessly implemented into existing systems, since only a module that preprocesses the feature vectors is needed. Furthermore, Deng et al. [16] have shown that denoising or preprocessing can achieve a superior performance over acoustic model adaptation.

Among the feature-domain methods, we have those techniques based on Bayesian estimation of the clean features from the noisy observations. These techniques require a training stage where the statistical relationship between clean and noisy feature domains is learned by direct feature comparison. This information can be extracted from a stereo database that includes both clean and noisy features. As mentioned by Afify et al. [17], although stereo data is usually not available, it can be relatively easy to collect for certain scenarios, e.g., in-car environments. In other cases, stereo data can be artificially generated by adding noise to the existing clean training data [18].

Using the aforementioned statistical relationship between clean and noisy domains, clean feature vectors can be esti-

mated by applying different Bayesian methods, such as the MMSE estimation. One of the first approaches based on MMSE and stereo data was proposed in [19] with the SNR-dependent cepstral normalization (SDCN) and codeword-dependent cepstral normalization (CDCN). Since then, more sophisticated techniques have appeared, as the multivariate Gaussian based cepstral normalization algorithm (RATZ) [13], stereo based piecewise linear compensation for environments (SPLICE) [20], multi-environment models based linear normalization (MEMLIN) [21], front-end joint uncertainty decoding (FE-Joint) [22], and stereo-based stochastic mapping (SSM) [17], [23]. All these techniques assume a certain prior probability density function (pdf) for the features being estimated. A common solution consists of modeling this pdf by means of a Gaussian mixture model (GMM).

In this paper, we propose a unified estimation framework for feature compensation based on the MMSE criterion. Our proposal, instead of modeling the clean and noisy feature spaces with GMMs, characterizes every space with a set of prototype vectors. This set can be alternatively considered a vector quantization (VQ) codebook that partitions the feature space into a set of cells [24]. As it will be shown, the MMSE estimators derived from this VQ representation are more efficient than other similar techniques using GMMs. However, the hard decision introduced by the use of VQ (a cell is represented by a centroid instead of a probability function) could lead to a performance reduction. For this reason, we present a novel MMSE formulation that can cope with this disadvantage. In addition, we show that the recognizer performance can be significantly improved by considering that every VQ cell contains a set of overlapping subregions, which provides a more accurate mapping between the clean and noisy domains. Furthermore, the VQ representation will allow us a straightforward modeling of time correlations by means of hidden Markov models (HMM) [25]. We will show that this modeling results in a performance improvement in terms of recognition accuracy.

The compensation methods described until now assume that the characteristics of the environmental noise are available. However, as we commented above, in many acoustic environments stereo data are unavailable. In order to overcome this limitation we will present a set of feature estimation schemes based on a multiple-model framework [26], [27]. Under this approach a set of models is trained off-line for different environmental noise conditions. Then, the final clean vector estimate is obtained as a linear combination of the estimates computed for every pre-trained model. By using a sufficient number of models, performance improvements can be achieved even for test conditions not considered in the training stage. Finally, a couple of methods that exploit the uncertainty introduced by environmental noise are shown. These methods consider the uncertainty of the estimates into the decoding process, thus reducing the mismatch with the clean speech acoustic models.

This paper is organized as follows. The next section is devoted to the fundamentals of the MMSE estimation and its application to noisy feature compensation. The proposed VQ-based MMSE estimation and its corresponding experimental results are shown in Section III. HMM modeling is introduced and tested in Section IV. In Section V, we discuss the issues of unknown acoustic environments and uncertainty processing. Finally, the paper is summarized in Section VI.

II. Noise effects and review of MMSE estimation

Let us consider a speech signal corrupted by additive and convolutional noises. In this case, the distorted Mel-frequency cepstral coefficient (MFCC) feature vector can be expressed as [19],

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{x} + \boldsymbol{h} + \boldsymbol{C}\log\left(1+\exp\left(\boldsymbol{C}^{-1}(\boldsymbol{n}-\boldsymbol{x}-\boldsymbol{h})\right)\right) \\ &= \boldsymbol{x} + f(\boldsymbol{x},\boldsymbol{h},\boldsymbol{n}) \end{aligned} \quad (1)$$

where $\boldsymbol{y}$, $\boldsymbol{x}$, $\boldsymbol{h}$ and $\boldsymbol{n}$ represent the noisy speech, clean speech, channel and additive noise MFCC vectors, respectively. $\boldsymbol{C}$ and $\boldsymbol{C}^{-1}$ denote the discrete cosine transform (DCT) matrix and its inverse (IDCT). According to this equation, the distorted features are obtained by means of a non-linear function of the channel, the additive noise, and the clean speech.

In order to compensate the transformations introduced by the noise in the speech signal, we can assume that the statistics of the distortion process are given by the conditional pdf $p(\boldsymbol{x}|\boldsymbol{y})$. Therefore, the MMSE estimate of the clean feature vector, $\hat{\boldsymbol{x}}$, can be obtained as,

$$\hat{\boldsymbol{x}} = E[\boldsymbol{x}|\boldsymbol{y}] = \int_{\boldsymbol{x}} \boldsymbol{x}\cdot p(\boldsymbol{x}|\boldsymbol{y})d\boldsymbol{x} \quad (2)$$

where $E[\cdot]$ is the expectation operator.

The main problem of the above expression is modeling the posterior $p(\boldsymbol{x}|\boldsymbol{y})$. To solve this problem, we can assume that the clean and distorted features are modeled by means of pdf mixtures in the following way,

$$p(\boldsymbol{x}) = \sum_{k_x} p(\boldsymbol{x}|k_x)P(k_x) \quad (3)$$

$$p(\boldsymbol{y}) = \sum_{k_y} p(\boldsymbol{y}|k_y)P(k_y) \quad (4)$$

where k_x and k_y are the components (e.g., Gaussians) of the pdf mixtures that model the clean and noisy spaces, respectively.

Applying this decomposition, the conditional probability $p(\boldsymbol{x}|\boldsymbol{y})$ can be expressed as,

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{y}) &= \sum_{k_x}\sum_{k_y} p\left(\boldsymbol{x},k_x,k_y|\boldsymbol{y}\right) \\ &= \sum_{k_x}\sum_{k_y} p\left(\boldsymbol{x}|k_x,k_y,\boldsymbol{y}\right) p\left(k_x|k_y,\boldsymbol{y}\right) p\left(k_y|\boldsymbol{y}\right) \end{aligned} \quad (5)$$

Therefore, the MMSE estimate in eqn. (2) can be computed as,

$$\hat{\boldsymbol{x}} = \sum_{k_x}\sum_{k_y} E[\boldsymbol{x}|k_x,k_y,\boldsymbol{y}]P(k_x|k_y,\boldsymbol{y})P(k_y|\boldsymbol{y}) \quad (6)$$

where $P(k_y|\boldsymbol{y})$ is the *a posteriori* probability of the noisy mixture component k_y given the observation $\boldsymbol{y}$, $P(k_x|k_y,\boldsymbol{y})$ is the conditional probability of the clean mixture component k_x given the noisy model mixture component k_y and the observation $\boldsymbol{y}$, and $E[\boldsymbol{x}|k_x,k_y,\boldsymbol{y}]$ is the estimate of the clean

feature vector $\boldsymbol{x}$ given the mixture components (k_x, k_y) and the observation $\boldsymbol{y}$.

An alternative way to express the above estimation is derived from eqn. (1), where an additive correction vector $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{x}$ can be applied to compensate the input feature vector. In such a case,

$$\hat{\boldsymbol{x}} = \boldsymbol{y} - \hat{\boldsymbol{r}}(\boldsymbol{y}) \tag{7}$$

where the estimate of the correction vector is obtained as,

$$\hat{\boldsymbol{r}}(\boldsymbol{y}) = \sum_{k_x}\sum_{k_y} E[\boldsymbol{r}|k_x, k_y, \boldsymbol{y}]P(k_x|k_y, \boldsymbol{y})P(k_y|\boldsymbol{y}) \tag{8}$$

Note that if the expectation term in eqn. (6) is set to

$$E[\boldsymbol{x}|k_x, k_y, \boldsymbol{y}] = \boldsymbol{y} - E[\boldsymbol{r}|k_x, k_y, \boldsymbol{y}] \tag{9}$$

then this equation becomes (7).

A. GMM-based MMSE estimation

From the above general framework, we can derive many of the MMSE compensation techniques based on stereo data that can be found in the literature. For example, multivariate Gaussian-based cepstral normalization (RATZ) [13] models the clean feature space by means of a GMM. That is, the conditional pdfs in eqn. (3) are Gaussians. Since RATZ only models the clean feature space, it can be derived that,

$$E[\boldsymbol{r}|k_x, k_y, \boldsymbol{y}] = E[\boldsymbol{r}|k_x] = \boldsymbol{r}_{k_x} \tag{10}$$

$$\hat{\boldsymbol{r}}(\boldsymbol{y}) = \sum_{k_x} \boldsymbol{r}_{k_x} P(k_x|\boldsymbol{y}) \tag{11}$$

where $\boldsymbol{r}_{k_x}$ is the correction vector computed for the clean Gaussian k_x (e.g., using stereo data), and $P(k_x|\boldsymbol{y})$ is the *a posteriori* probability of k_x given the observation $\boldsymbol{y}$. This posterior is computed from eqn. (3) assuming an additive effect of the noise on the MFCC domain, i.e., $p(\boldsymbol{y}|k_x) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_{k_x} + \boldsymbol{r}_{k_x}, \boldsymbol{\Sigma}_{k_x} + \boldsymbol{\Sigma}_{\boldsymbol{r}_{k_x}})$, where $\boldsymbol{\mu}_{k_x}$ and $\boldsymbol{\Sigma}_{k_x}$ are the mean and covariance of Gaussian k_x, and $\boldsymbol{\Sigma}_{\boldsymbol{r}_{k_x}}$ is the covariance of the correction bias for the same Gaussian. Finally, RATZ computes the clean vector estimate by means of eqn. (7).

In contrast to RATZ, stereo-based piecewise linear compensation for environments (SPLICE) [20] models the distorted feature space with a GMM. For each Gaussian k_y in the noisy feature space, SPLICE computes a correction vector $\boldsymbol{r}_{k_y}$ using stereo data. Thus, eqn. (7) becomes

$$\hat{\boldsymbol{x}} = \boldsymbol{y} - \hat{\boldsymbol{r}}(\boldsymbol{y}) = \boldsymbol{y} - \sum_{k_y} \boldsymbol{r}_{k_y} P(k_y|\boldsymbol{y}) \tag{12}$$

where $P(k_y|\boldsymbol{y})$ is the *a posteriori* probability of the noisy Gaussian k_y given the input feature vector $\boldsymbol{y}$. This term is computed using eqn. (4).

While the above techniques only model one feature space, clean or noisy, the multi-environment model based linear normalization (MEMLIN) [21] models both. In this case, the estimation formulae are,

$$E[\boldsymbol{r}|k_x, k_y, \boldsymbol{y}] = E[\boldsymbol{r}|k_x, k_y] = \boldsymbol{r}_{k_x k_y} \tag{13}$$

$$\hat{\boldsymbol{r}}(\boldsymbol{y}) = \sum_{k_y}\sum_{k_x} \boldsymbol{r}_{k_x k_y} P(k_x|k_y)P(k_y|\boldsymbol{y}) \tag{14}$$

Again, assuming a perfect knowledge of the distortion process, the correction vector associated to each pair of Gaussians, $\boldsymbol{r}_{k_x k_y}$, can be computed using stereo data. Finally, MEMLIN computes the clean vector estimate $\hat{\boldsymbol{x}}$ by means of eqn. (7).

More recently, a novel MMSE approach based on a joint modeling of the clean and noisy features spaces has been researched. Thus, methods such as front-end joint uncertainty decoding (FE-Joint) [22] or stereo-based stochastic mapping (SSM) [17], [23] model the joint variable $\boldsymbol{z} \equiv (\boldsymbol{x}, \boldsymbol{y})$ by means of a GMM as,

$$p(\boldsymbol{z}) = \sum_k P(k)\mathcal{N}\left(\boldsymbol{z}; \boldsymbol{\mu}_z^{(k)}, \boldsymbol{\Sigma}_z^{(k)}\right) \tag{15}$$

where the mean and covariance of each Gaussian k can be partitioned as

$$\boldsymbol{\mu}_z^{(k)} = \begin{pmatrix} \boldsymbol{\mu}_x^{(k)} \\ \boldsymbol{\mu}_y^{(k)} \end{pmatrix} \tag{16}$$

$$\boldsymbol{\Sigma}_z^{(k)} = \begin{pmatrix} \boldsymbol{\Sigma}_x^{(k)} & \boldsymbol{\Sigma}_{xy}^{(k)} \\ \boldsymbol{\Sigma}_{yx}^{(k)} & \boldsymbol{\Sigma}_y^{(k)} \end{pmatrix} \tag{17}$$

Then, the MMSE estimate (for the case of SSM) can be obtained as,

$$\hat{\boldsymbol{x}} = \sum_k E[\boldsymbol{x}|k, \boldsymbol{y}]P(k|\boldsymbol{y}) \tag{18}$$

where the posterior $P(k|\boldsymbol{y})$ is computed using the marginal distribution $p(\boldsymbol{y})$, and the expectation term $E[\boldsymbol{x}|k, \boldsymbol{y}]$ is defined as,

$$E[\boldsymbol{x}|k, \boldsymbol{y}] = \boldsymbol{\mu}_x^{(k)} + \boldsymbol{\Sigma}_{xy}^{(k)}\left(\boldsymbol{\Sigma}_y^{(k)}\right)^{-1}\left(\boldsymbol{y} - \boldsymbol{\mu}_y^{(k)}\right) \tag{19}$$

where $\boldsymbol{\mu}_x^{(k)}$, $\boldsymbol{\mu}_y^{(k)}$ are the mean vectors of the clean and noisy marginal distributions for the k^{th} Gaussian, respectively, $\boldsymbol{\Sigma}_y^{(k)}$ is the covariance matrix of the noisy marginal distribution, and $\boldsymbol{\Sigma}_{xy}^{(k)}$ is the clean-noisy cross-covariance matrix.

III. VQ-based MMSE estimation

We have seen that the traditional MMSE estimation techniques model the clean and distorted spaces using GMMs. In contrast, we propose to characterize both feature spaces by means of two sets of prototype vectors. Each set can be considered as a VQ codebook that partitions its corresponding feature space into a set of disjoint cells. The sets of VQ cells representing the clean space X and noisy space Y will be referred to as $\{C_X^{(i)}\ (i = 1, \ldots, M)\}$ and $\{C_Y^{(j)}\ (j = 1, \ldots, N)\}$, respectively. These cells will play now the role of the pdfs k_x and k_y in equations (6) and (8).

The partition into cells requires the definition of a distance measure. We will employ a weighted Euclidean distance measure that, for the case of the noisy space, is defined as,

$$\mathrm{d}\left(\boldsymbol{y}, C_Y^{(j)}\right) = \left(\boldsymbol{\mu}_Y^{(j)} - \boldsymbol{y}\right)^T \mathrm{diag}\left(\boldsymbol{\Sigma}_Y^{(j)}\right)^{-1}\left(\boldsymbol{\mu}_Y^{(j)} - \boldsymbol{y}\right) \tag{20}$$

where the operator $\mathrm{diag}(\cdot)$ returns a diagonal matrix with the elements of the main diagonal of its argument, $\boldsymbol{\mu}_Y^{(j)}$ is the mean vector (centroid) of cell $C_Y^{(j)}$, and $\boldsymbol{\Sigma}_Y^{(j)}$ is the corresponding covariance matrix. A similar distance measure can be defined

for the clean space. Thus, the VQ cell corresponding to an input feature vector $\boldsymbol{y}$ can be determined as follows,

$$C_Y^*(\boldsymbol{y}) = \underset{j}{\operatorname{argmin}}\, \mathrm{d}\left(\boldsymbol{y}, C_Y^{(j)}\right) \tag{21}$$

For the sake of simplicity, we will simply write C_Y^* for $C_Y^*(\boldsymbol{y})$. Given that the noisy feature space is VQ-partitioned, the posterior of the noisy mixture component $P(k_y|\boldsymbol{y})$ of equations (6) and (8) can be defined now as follows,

$$P\left(C_Y^{(j)}\middle|\,\boldsymbol{y}\right) = \begin{cases} 1 & C_Y^{(j)} = C_Y^* \\ 0 & otherwise \end{cases} \tag{22}$$

This expression involves that the double sum of the MMSE-based estimation (eqn. (6)) is reduced to a single sum. Additionally, the conditional probability of the clean component $P(k_x|k_y, \boldsymbol{y})$ in eqns. (6) and (8) is simplified to $P(C_X^{(i)}|C_Y^{(j)})$. With these modifications, the VQ-based MMSE estimation can be written as,

$$\hat{\boldsymbol{x}} = \sum_{i=1}^{M} E\left[\boldsymbol{x}\middle|C_X^{(i)}, C_Y^*, \boldsymbol{y}\right] P\left(C_X^{(i)}\middle|\,C_Y^*\right) \tag{23}$$

Alternatively, we can express this estimator in terms of correction vectors as,

$$\hat{\boldsymbol{x}} = \boldsymbol{y} - \hat{\boldsymbol{r}}(\boldsymbol{y}) \tag{24}$$

$$\hat{\boldsymbol{r}}(\boldsymbol{y}) = \sum_{i=1}^{M} E\left[\boldsymbol{r}\middle|C_X^{(i)}, C_Y^*, \boldsymbol{y}\right] P\left(C_X^{(i)}\middle|\,C_Y^*\right) \tag{25}$$

As can be observed, the application of VQ to eqn. (6) implies that only the nearest VQ cell C_Y^* to the input $\boldsymbol{y}$ is involved in the estimation. One could argue that a similar behavior can be obtained if, for example, GMM modeling is applied and only the most likely Gaussian is selected. However, there are important differences that must be pointed out. Thus, although the distance measure of eqn. (20) can be compared to the evaluation of a Gaussian pdf, they are, in fact, different. Also, different (specific) training techniques can be applied to build GMMs and VQ codebooks (e.g., expectation-maximization (EM) for GMM and k-means for VQ). These differences result in different representations of the feature spaces and, thus, different parameters for compensation.

In the following subsections, we develop several MMSE estimation techniques based on the proposed VQ framework of eqn. (23). As we will see, the proposed techniques mainly differ in the way that the expected value $E[\boldsymbol{x}|C_X^{(i)}, C_Y^*, \boldsymbol{y}]$ is computed. It is important to note that the original noisy input vector $\boldsymbol{y}$ is explicitly maintained in this expected value in spite of the VQ partition. That is, the VQ partition is applied to compute the probabilities required by the MMSE estimation, but does not necessarily involve a VQ quantization of the input vector which could lead us to a performance reduction.

A. Fully discrete MMSE estimation

The VQ-based MMSE estimation of eqn. (23) is a common approach in digital transmission systems [28], where, in fact, the transmitted vectors are quantized before transmission. As first approach, we assume that both feature spaces, clean and distorted, are VQ-quantized. For the case of the noisy space, this means that the input feature vector can be approximated by the corresponding VQ centroid, so that $\boldsymbol{y} \approx \boldsymbol{\mu}_Y^*$, where $\boldsymbol{\mu}_Y^*$ is the mean of cell C_Y^*. On the other hand, the VQ quantization involves that all feature vectors inside every clean cell $C_X^{(i)}$ are represented by centroid $\boldsymbol{\mu}_X^{(i)}$. Additionally, we assume that both feature spaces are independent. Under these assumptions, the clean vector expectation of eqn. (23) is approximated by

$$E\left[\boldsymbol{x}\middle|C_X^{(i)}, C_Y^*, \boldsymbol{y}\right] = E\left[\boldsymbol{x}\middle|C_X^{(i)}\right] = \boldsymbol{\mu}_X^{(i)} \tag{26}$$

Equation (26) yields a *fully discrete MMSE* (FD-MMSE) estimator that is derived from eqn. (23) as,

$$\hat{\boldsymbol{x}}_{FD} = \sum_{i=1}^{M} \boldsymbol{\mu}_X^{(i)} P\left(C_X^{(i)}\middle|\,C_Y^*\right) \tag{27}$$

where $P(C_X^{(i)}|C_Y^*)$ can be estimated using stereo data as frequencies of appearance.

Now, we can express the FD-MMSE estimation of eqn. (27) in terms of correction vectors as it appears in eqn. (24). Given the aforementioned assumptions (i.e., both feature spaces are VQ-quantized and independent of each other), the correction vector for the pair of cells $C_X^{(i)}$ and $C_Y^{(j)}$ can be obtained as,

$$\begin{aligned} E\left[\boldsymbol{r}\middle|C_X^{(i)}, C_Y^{(j)}, \boldsymbol{y}\right] &= E\left[\boldsymbol{r}\middle|C_X^{(i)}, C_Y^{(j)}\right] \\ &= E\left[\boldsymbol{y} - \boldsymbol{x}\middle|C_X^{(i)}, C_Y^{(j)}\right] \\ &= E\left[\boldsymbol{y}\middle|C_X^{(i)}, C_Y^{(j)}\right] - E\left[\boldsymbol{x}\middle|C_X^{(i)}, C_Y^{(j)}\right] \\ &\approx E\left[\boldsymbol{y}\middle|C_Y^{(j)}\right] - E\left[\boldsymbol{x}\middle|C_X^{(i)}\right] = \boldsymbol{\mu}_Y^{(j)} - \boldsymbol{\mu}_X^{(i)} \end{aligned} \tag{28}$$

Then, using eqn. (28) in eqn. (25), we obtain

$$\begin{aligned} \hat{\boldsymbol{r}}(\boldsymbol{\mu}_Y^*) &= \sum_{i=1}^{M} E\left[\boldsymbol{r}\middle|C_X^{(i)}, C_Y^*\right] P\left(C_X^{(i)}\middle|\,C_Y^*\right) \\ &= \boldsymbol{\mu}_Y^* - \sum_{i=1}^{M} \boldsymbol{\mu}_X^{(i)} P\left(C_X^{(i)}\middle|\,C_Y^*\right) \\ &= \boldsymbol{\mu}_Y^* - \hat{\boldsymbol{x}}_{FD} \end{aligned} \tag{29}$$

where, as can be noted, we assume that $\hat{\boldsymbol{r}}$ is constant for all the points inside a noisy cell, that is, $\hat{\boldsymbol{r}}(\boldsymbol{y}) \approx \hat{\boldsymbol{r}}(\boldsymbol{\mu}_Y^*)$. Finally, the expression for the FD-MMSE estimation expressed in terms of correction vectors is

$$\hat{\boldsymbol{x}}_{FD} = \boldsymbol{\mu}_Y^* - \hat{\boldsymbol{r}}(\boldsymbol{\mu}_Y^*) \tag{30}$$

B. Basic bias MMSE estimation

Let us examine the estimation proposed for FD-MMSE in eqn. (30). As can be seen, the VQ quantization involves that the input feature vector $\boldsymbol{y}$ is replaced by $\boldsymbol{\mu}_Y^*$. However, unlike in transmission applications, the input of a robust ASR application is not usually quantized. In order to avoid the quantization distortion, we should apply the correction vector directly to $\boldsymbol{y}$ instead of $\boldsymbol{\mu}_Y^*$. In this sense, eqns. (24) and (25) will help us to derive an estimator which minimizes this distortion. As the computation of a correction vector $\hat{\boldsymbol{r}}(\boldsymbol{y})$ for every input feature vector is unfeasible, we propose to compute

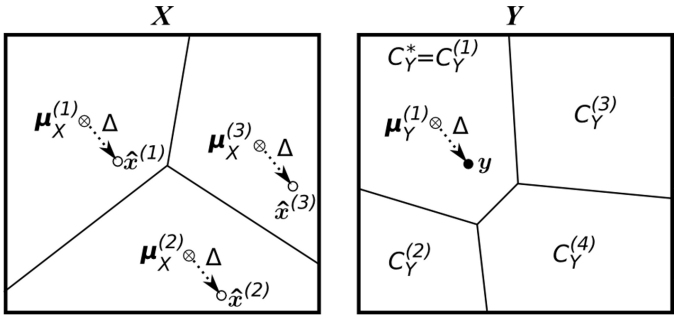

Fig. 1. Scheme of the compensation carried out by *BB-MMSE*.

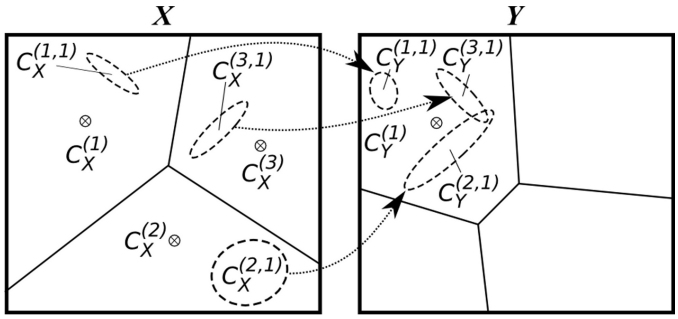

Fig. 2. Mapping between subregions of the clean and noisy feature spaces.

an unique correction vector for every noisy feature vector corresponding to the same cell. To do so, we approximate $\hat{\boldsymbol{r}}(\boldsymbol{y})$ by $\hat{\boldsymbol{r}}(\boldsymbol{\mu}_Y^*)$ (as computed in equation (29) for the fully discrete approach) by considering that $\hat{\boldsymbol{r}}$ is constant in cell C_Y^*. Then, eqn. (24) yields the following estimate

$$\hat{\boldsymbol{x}}_{BB} = \boldsymbol{y} - \hat{\boldsymbol{r}}(\boldsymbol{\mu}_Y^*) \tag{31}$$

We will refer to this approach as *basic bias MMSE* (BB-MMSE) estimation in the following. Alternatively, by using eqn. (29) we can express this estimator as,

$$\begin{aligned} \hat{\boldsymbol{x}}_{BB} &= \boldsymbol{y} - (\boldsymbol{\mu}_Y^* - \hat{\boldsymbol{x}}_{FD}) \\ &= \hat{\boldsymbol{x}}_{FD} + (\boldsymbol{y} - \boldsymbol{\mu}_Y^*) \end{aligned} \tag{32}$$

where $\hat{\boldsymbol{x}}_{FD}$ is the FD-MMSE estimate computed according to eqn. (27). As we can see, the resulting BB-MMSE estimate is the FD-MMSE one biased by the quantization error $\Delta = \boldsymbol{y} - \boldsymbol{\mu}_Y^*$.

We can see that the estimation proposed for BB-MMSE in eqns. (31) and (32) is quite similar to the one proposed for SPLICE in eqn. (12). Both techniques compute a partial clean estimate for every component k_y in the noisy space by subtracting a correction vector r_{k_y} to the input $\boldsymbol{y}$. Nevertheless, only the estimate computed for the best noisy cell C_Y^* is considered in BB-MMSE. In addition, SPLICE only models the noisy feature space, while we model both domains. This dual modeling means that the projection space associated to a correction vector transformation in BB-MMSE is smaller than in SPLICE [21], so that we can consider that our technique is more accurate in this sense.

An alternative way to express the BB-MMSE estimation can be derived from eqn. (23), where the expectation term is computed (applying eqns. (9) and (29)) as,

$$\begin{aligned} E\left[\boldsymbol{x}\left|C_X^{(i)}, C_Y^*, \boldsymbol{y}\right.\right] &= \boldsymbol{y} - E\left[\boldsymbol{r}\left|C_X^{(i)}, C_Y^*\right.\right] \\ &= \boldsymbol{y} - \left(\boldsymbol{\mu}_Y^* - \boldsymbol{\mu}_X^{(i)}\right) \\ &= \boldsymbol{\mu}_X^{(i)} + (\boldsymbol{y} - \boldsymbol{\mu}_Y^*) \end{aligned} \tag{33}$$

The graphical interpretation of eqn. (33) is depicted in Fig. 1. As can be seen, the clean and noisy feature spaces are VQ-partitioned. In the case of the clean feature space, three cells, with centroids $\boldsymbol{\mu}_X^{(1)}$, $\boldsymbol{\mu}_X^{(2)}$, and $\boldsymbol{\mu}_X^{(3)}$, are considered. On the other hand, the noisy space is partitioned into four cells. The BB-MMSE estimate of the clean feature vector, $\hat{\boldsymbol{x}}_{BB}$, is obtained as a weighted average of $\hat{\boldsymbol{x}}^{(i)}$ $(i = 1, 2, 3)$. Each point $\hat{\boldsymbol{x}}^{(i)}$ in the figure represents the expected value $E[\boldsymbol{x}|C_X^{(i)}, C_Y^*, \boldsymbol{y}]$, which is obtained by applying the quantization displacement $\Delta = \boldsymbol{y} - \boldsymbol{\mu}_Y^*$ to every centroid $\boldsymbol{\mu}_X^{(i)}$.

From eqn. (29), it can be observed that the bias vector $\hat{\boldsymbol{r}}\,(\boldsymbol{\mu}_Y^*)$ used in FD-MMSE and BB-MMSE is obtained as a piecewise linear combination of several correction vectors computed for every clean VQ cell. However, the approximation of $\boldsymbol{y}$ by $\boldsymbol{\mu}_Y^*$ involves a quantization distortion that could lead to a degradation in the recognition accuracy. BB-MMSE tries to mitigate this distortion by means of the Δ displacement. We will see in the following subsections how this distortion can be further reduced by means of a more reliable estimation of the expected values $E[\boldsymbol{x}|C_X^{(i)}, C_Y^*, \boldsymbol{y}]$ and $E[\boldsymbol{r}|C_X^{(i)}, C_Y^*, \boldsymbol{y}]$.

C. Refined bias MMSE estimation

The correction vector $\hat{\boldsymbol{r}}\,(\boldsymbol{\mu}_Y^*)$ used in FD-MMSE and BB-MMSE (see eqn. (29)) is an approximation which exclusively uses VQ centroids. In order to obtain a more accurate representation of the transformation between cells defined by $E[\boldsymbol{r}|C_X^{(i)}, C_Y^{(j)}, \boldsymbol{y}]$, we introduce in the following the concept of subregions in a VQ cell. We will consider that every clean cell $C_X^{(i)}$ includes a set of subregions $\{C_X^{(i,j)}, (j = 1, \ldots, N)\}$, where $C_X^{(i,j)}$ represents all the clean feature vectors whose corresponding distorted ones belong to the noisy cell $C_Y^{(j)}$. Similarly, $C_Y^{(i,j)}$ represents the subregion of $C_Y^{(j)}$ where those distorted vectors are mapped. It must be pointed out that these subregions do not involve a hard partition of every cell (as the one due to the VQ partition) since, in fact, they will be strongly overlapped due to the random nature of the noise. A scheme of the mapping between subregions of the clean and noisy feature spaces is depicted in Fig. 2.

Using the notation introduced for subregions, the expected values $E[\boldsymbol{r}|C_X^{(i)}, C_Y^{(j)}, \boldsymbol{y}]$ required for the computation of $\hat{\boldsymbol{r}}(\boldsymbol{y})$ (see equations (24) and (25)) can be more accurately estimated as follows,

$$\begin{aligned} E\left[\boldsymbol{r}\left|C_X^{(i)}, C_Y^{(j)}, \boldsymbol{y}\right.\right] &= E\left[\boldsymbol{r}\left|C_X^{(i)}, C_Y^{(j)}\right.\right] \\ &= E\left[\boldsymbol{y}\left|C_X^{(i)}, C_Y^{(j)}\right.\right] - E\left[\boldsymbol{x}\left|C_X^{(i)}, C_Y^{(j)}\right.\right] \\ &= E\left[\boldsymbol{y}\left|C_Y^{(i,j)}\right.\right] - E\left[\boldsymbol{x}\left|C_X^{(i,j)}\right.\right] \\ &= \boldsymbol{\mu}_Y^{(i,j)} - \boldsymbol{\mu}_X^{(i,j)} \end{aligned} \tag{34}$$

where $\boldsymbol{\mu}_X^{(i,j)}$ and $\boldsymbol{\mu}_Y^{(i,j)}$ represent the mean vectors of subregions $C_X^{(i,j)}$ and $C_Y^{(i,j)}$, respectively. These vectors can be easily obtained from a stereo database for a given acoustic environment using the feature vectors assigned to each subregion.

The subregion-based approach introduced above leads us to a novel estimation method that we will refer to as *refined bias MMSE* (RB-MMSE). This estimator employs equations (24), (25) and (34) to compute a clean estimate $\hat{\boldsymbol{x}}_{RB}$ from the noisy feature vector $\boldsymbol{y}$. In order to express it in the form of eqn. (23), the required expectation values $E[\boldsymbol{x}|C_X^{(i)}, C_Y^{(j)}, \boldsymbol{y}]$ are obtained as

$$\begin{aligned} E\left[\boldsymbol{x}\left|C_X^{(i)}, C_Y^{(j)}, \boldsymbol{y}\right.\right] &= \boldsymbol{y} - E\left[\boldsymbol{r}\left|C_X^{(i)}, C_Y^{(j)}\right.\right] \\ &= \boldsymbol{y} - \left(\boldsymbol{\mu}_Y^{(i,j)} - \boldsymbol{\mu}_X^{(i,j)}\right) \\ &= \boldsymbol{\mu}_X^{(i,j)} + \left(\boldsymbol{y} - \boldsymbol{\mu}_Y^{(i,j)}\right) \end{aligned} \quad (35)$$

Again, we can see that the clean vector estimate obtained for every clean VQ cell is computed modifying a mean vector $\boldsymbol{\mu}_X^{(i,j)}$ by a quantization displacement $\boldsymbol{y} - \boldsymbol{\mu}_Y^{(i,j)}$. Moreover, the fact of considering subregions inside a cell provides us with a more flexible modeling of the noise distortion (a correction vector is computed to compensate for the transformation between every pair of cells in the clean and noisy domains). As can be noted, this approach resembles the double feature space modeling of MEMLIN, although in our case this is carried out more efficiently.

D. MMSE estimation with covariance normalization

In the previous bias-based MMSE techniques, BB-MMSE and RB-MMSE, the noisy input vector is compensated by subtracting a correction vector to every observed feature vector. Thus, in a sense, these methods could be considered a sort of cell-dependent cepstral mean normalization (CMN) techniques where the shift in the mean introduced by the noise is corrected by applying a bias correction vector to every noisy vector. However, it is well known that the effect of the environmental noise in the speech signal is not only a shift in the mean, but also a change in the covariance of the distributions that model the speech [13], [29]. This effect is also depicted in Fig. 2, where the shapes of the subregions in the clean space can be modified by the acoustic noise in the distorted space.

In order to model these covariance modifications, we propose to compensate every feature vector by means of a linear transformation. The proposed transformation assumes that the subregions in the clean and noisy feature spaces follow Gaussian pdfs $C_X^{(i,j)} \sim \mathcal{N}(\boldsymbol{\mu}_X^{(i,j)}, \boldsymbol{\Sigma}_X^{(i,j)})$ and $C_Y^{(i,j)} \sim \mathcal{N}(\boldsymbol{\mu}_Y^{(i,j)}, \boldsymbol{\Sigma}_Y^{(i,j)})$, respectively, where $\boldsymbol{\mu}_X^{(i,j)}$,$\boldsymbol{\mu}_Y^{(i,j)}$ are the mean vectors and $\boldsymbol{\Sigma}_X^{(i,j)}$,$\boldsymbol{\Sigma}_Y^{(i,j)}$ are the corresponding covariance matrices. Then, the transformation of the noisy subregion into the clean one is achieved by means of the following expression,

$$\begin{aligned} E\left[\boldsymbol{x}\left|C_X^{(i)}, C_Y^{(j)}, \boldsymbol{y}\right.\right] &= \boldsymbol{\mu}_X^{(i,j)} + \\ &+ \left(\boldsymbol{\Sigma}_X^{(i,j)}\right)^{1/2} \left(\boldsymbol{\Sigma}_Y^{(i,j)}\right)^{-1/2} \left(\boldsymbol{y} - \boldsymbol{\mu}_Y^{(i,j)}\right) \end{aligned} \quad (36)$$

where the exponent $1/2$ that raises both covariance matrices denotes the square root of the matrix[1].

The covariance matrices of eqn. (36) can be computed from a stereo database using the feature vectors assigned to each subregion. As can be noted, if the product of covariance matrices is the identity matrix (i.e., the shapes of the clean subregions are not modified by the noise), then eqn. (36) is the same as eqn. (35) from RB-MMSE.

As can be seen, the transformation proposed in eqn. (36) is quite similar to the one applied by SSM in eqn. (19). However, SSM uses a cross-covariance matrix between the clean and noisy spaces, while our method only uses the covariance matrices of each space independently. In order to evaluate both transformations under the VQ-based estimation framework proposed in eqn. (23), we derive a new set of estimators using eqn. (19) instead of eqn. (36). Preliminary experiments showed that our transformation clearly outperforms (in terms of recognition accuracy) the one used in SSM. This can be explained as follows. The normalization applied in eqn. (19) assumes joint Gaussian on the noisy and clean features to derive the conditional distribution of the clean speech. While this assumption is accurate in SSM, where both feature spaces are jointly modeled by means of a GMM, in our case is not totally true, since every feature space is modeled independently.

Finally, we can use the estimated values computed by means of eqn. (36) in eqn. (23), resulting in two estimation methods. The first one, which will be called *mean and full covariance MMSE* (fMV-MMSE), assumes a full covariance matrix for every subregion. On the other hand, the second method, which will be referred to as *mean and diagonal covariance MMSE* (dMV-MMSE), assumes diagonal covariance matrices. In this case, the square root of the matrices is computed simply by taking the square root of each element. Furthermore, matrix products become element-wise vector products with the consequent computational saving.

E. Computational complexity

In this section, an analysis of the computational complexity of the proposed VQ-based MMSE estimators is carried out. Also, we compare the complexity of our techniques with respect to other well-known GMM-based approaches (SPLICE and MEMLIN). To do so, we simplify the expression obtained for every technique taking into account that some terms can be grouped together and precomputed off-line. For example, the expression obtained for fMV-MMSE (see eqns. (23) and

[1]The square root of a matrix $\boldsymbol{\Sigma}$ is the matrix $\boldsymbol{\Sigma}^{1/2}$ which fulfills $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\Sigma}^{1/2}$. If the matrix $\boldsymbol{\Sigma}$ is symmetric and positive definite, which is the case of the covariance matrices, the square root can be computed as follows,

$$\boldsymbol{\Sigma}^{1/2} = \boldsymbol{V}\,\mathrm{sqrt}(\boldsymbol{D})\boldsymbol{V}^T \quad (37)$$

where $\boldsymbol{V}$ is the matrix of eigenvectors of $\boldsymbol{\Sigma}$ and $\boldsymbol{D}$ is the diagonal matrix of eigenvalues of $\boldsymbol{\Sigma}$. The notation $\mathrm{sqrt}(\boldsymbol{D})$ stands for the element-wise square root of $\boldsymbol{D}$.

Technique	Simplified expression	Complexity
SPLICE	$\hat{\boldsymbol{x}} = \boldsymbol{y} - \sum_{k_y} \boldsymbol{r}_{k_y} P(k_y\|\boldsymbol{y})$	$O(k_y D)$
MEMLIN	$\hat{\boldsymbol{x}} = \boldsymbol{y} - \sum_{k_y} \boldsymbol{r}_{k_y} P(k_y\|\boldsymbol{y})$	$O(k_y D)$
FD-MMSE	$\hat{\boldsymbol{x}} = \boldsymbol{c}$	$O(1)$
BB-MMSE	$\hat{\boldsymbol{x}} = \boldsymbol{y} - \boldsymbol{r}$	$O(D)$
RB-MMSE	$\hat{\boldsymbol{x}} = \boldsymbol{y} - \boldsymbol{r}$	$O(D)$
dMV-MMSE	$\hat{\boldsymbol{x}} = \boldsymbol{a}\boldsymbol{y} + \boldsymbol{b}$	$O(D)$
fMV-MMSE	$\hat{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b}$	$O(D^2)$

TABLE I
COMPARISON OF THE COMPUTATIONAL COMPLEXITY BETWEEN THE PROPOSED VQ AND GMM COMPENSATION TECHNIQUES.

(36)) can be simplified as follows,

$$\begin{aligned}\hat{\boldsymbol{x}}_{fMV} &= \sum_{i=1}^{M} E\left[\boldsymbol{x}\left|C_X^{(i)}, C_Y^*, \boldsymbol{y}\right.\right] P\left(C_X^{(i)}\left|C_Y^*\right.\right) \\ &= \sum_{i=1}^{M} \left(\boldsymbol{A}^{(i,*)}\boldsymbol{y} + \boldsymbol{b}^{(i,*)}\right) P\left(C_X^{(i)}\left|C_Y^*\right.\right) \\ &= \underbrace{\left(\sum_{i=1}^{M} P\left(C_X^{(i)}\left|C_Y^*\right.\right)\boldsymbol{A}^{(i,*)}\right)}_{\boldsymbol{A}}\boldsymbol{y} + \underbrace{\sum_{i=1}^{M} P\left(C_X^{(i)}\left|C_Y^*\right.\right)\boldsymbol{b}^{(i,*)}}_{\boldsymbol{b}} \\ &= \boldsymbol{A}\boldsymbol{y} + \boldsymbol{b} \end{aligned} \tag{38}$$

where $\boldsymbol{A}$ and $\boldsymbol{b}$ can be precomputed off-line for every VQ cell of the noisy space. $\boldsymbol{A}^{(i,*)}$ and $\boldsymbol{b}^{(i,*)}$ are obtained from eqn. (36) as,

$$\boldsymbol{A}^{(i,j)} = \left(\boldsymbol{\Sigma}_X^{(i,j)}\right)^{1/2}\left(\boldsymbol{\Sigma}_Y^{(i,j)}\right)^{-1/2} \tag{39}$$

$$\boldsymbol{b}^{(i,j)} = \boldsymbol{\mu}_X^{(i,j)} - \left(\boldsymbol{\Sigma}_X^{(i,j)}\right)^{1/2}\left(\boldsymbol{\Sigma}_Y^{(i,j)}\right)^{-1/2}\boldsymbol{\mu}_Y^{(i,j)} \tag{40}$$

these terms can also be precomputed.

Table I compares the complexity of the different feature compensation techniques presented in this work. The complexity accounts for the number of operations needed by every technique to carry out the compensation and it is represented in the asymptotic notation. To do so, the cost of the evaluation of the *a posteriori* probability of every component in the codebook (i.e., evaluation of a Gaussian pdf or the distance defined in eqn. (20)) is ignored and we focus on the complexity of the estimate computation. We will comment this below. Additionally, simplified expression is outlined for every method. For the sake of notation, we consider that products between vectors are element-wise. As can be seen, the complexity is a function that depends on the dimension of the feature vector D and, for GMM-based techniques, the number of Gaussians k_y in the noisy space. The proposed estimations based on VQ are more efficient than those derived from GMM (provided $D << k_y$, as usual), since only the closest VQ cell to the input is used during the estimation. This way, many operations can be precomputed.

As also shown in Table I, the subregion modeling does not necessarily introduce additional costs in computation or memory (e.g., BB-MMSE and RB-MMSE are equally efficient, but, as we will see, RB-MMSE outperforms BB-MMSE in recognition accuracy). Thus, the subregion modeling is an efficient way to approximate a codebook with squared number of cells. While M^2 comparisons are needed to choose the closest VQ cell in a codebook with M^2 cells, only M comparisons are needed in a codebook with subregions. Additionally, subregions improve the recognition accuracy by means of a more accurate modeling of the distortion introduced by acoustic noise.

Finally, some comments are needed about the computational cost involved by the *a posteriori* probability of every component in the codebook. In the GMM-based approaches, the likelihood of a feature vector given a Gaussian can be computed in the logarithmic domain where no exponential function is involved. In such a case, the cost of the computation of the log-likelihood for each Gaussian is almost the same as VQ distance of eqn. (20). However, when several Gaussians are involved in the estimation, a $logAdd$ function is usually applied to add the likelihoods of such Gaussians in the logarithmic domain. In this case, the exponential function is still required, leading to a whole computational cost even higher.

F. Experimental framework and results

The experimental setup is based on the framework proposed by ETSI STQ-Aurora working group using the Aurora-2 database [18]. This database consists of utterances of connected English digits. The vocabulary consists of 11 digits between 0 and 9 (zero can be uttered as 'zero' and 'oh'). In addition, two silence models (i.e., normal silence and short pause) are used. For our purposes, we selected the clean training set and the clean utterances from test *set A*. The clean training set comprises 8440 utterances from 55 male and 55 female speakers. On the other hand, the clean test subset consists of 4004 utterances containing a total of 13159 words.

A set of 9 environmental noises is chosen to test the proposed MMSE methods, namely: airport, highway, babble, bar, beach, pedestrian street, restaurant, street, and train station. The reason to employ these noises instead of the ones included in the Aurora-2 database is twofold. First, they allow us to test the different methods in a great number of realistic scenarios. Second, some of the noises included in the Aurora-2 database are too short for our purposes (the average length of our recordings is approximately 282 seconds). Every noise recording from this set is split into two parts: two-thirds are employed to train the MMSE estimators while the remaining third is reserved for testing. The training part of the 9 environmental noises has been added to the *Clean* training set of Aurora-2 at 6 different SNRs (20, 15, 10, 5, 0, and -5 dB), resulting in 54 environmental noisy training conditions plus the clean condition. Similarly, 55 test conditions are defined by artificially contaminating the clean test set of Aurora-2 with the testing part of the 9 environmental noises. The SNRs considered here are the same that in the training stage.

Speech features are extracted according with the European Telecommunication Standards Institute front-end (ETSI FE, ES 201 108) [30]. The final feature vector employed by the recognizer consists of 12 MFCCs (the 0th order cepstral

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg.
Baseline	99.02	90.79	75.53	50.70	25.86	11.27	6.18	50.83
Matched	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
FD-MMSE	96.19	93.72	90.21	81.24	61.82	31.33	14.39	71.66
BB-MMSE	99.02	97.93	96.28	90.57	74.70	43.02	18.57	80.50
RB-MMSE	99.02	98.23	96.79	91.60	76.82	46.60	20.02	82.01
dMV-MMSE	99.02	98.33	97.06	92.43	78.70	48.88	20.26	83.08
fMV-MMSE	99.02	98.37	97.15	92.88	79.61	50.04	20.89	83.61

TABLE II

WORD ACCURACY RESULTS (IN %) OBTAINED BY DIFFERENT SYSTEMS: 1) CLEAN ACOUSTIC MODELS WITH NO COMPENSATION (BASELINE); 2) MATCHED ACOUSTIC MODELS ARE APPLIED WITH NO COMPENSATION (MATCHED); 3) CLEAN ACOUSTIC MODELS WITH GMM-BASED MMSE COMPENSATION (SPLICE AND MEMLIN); 4) CLEAN ACOUSTIC MODELS WITH VQ-BASED MMSE COMPENSATION (FD-MMSE, BB-MMSE, RB-MMSE, DMV-MMSE, AND FMV-MMSE). THE RESULTS ARE DETAILED FOR EACH SNR (INCLUDING THE AVERAGE BETWEEN 0 DB AND 20 DB (AVG.)). THE MMSE-BASED SYSTEMS ARE TESTED USING 256 COMPONENTS (GAUSSIANS OR VQ CELLS) PER FEATURE SPACE.

coefficient is discarded), plus the log-energy feature and its delta and delta-delta coefficients, resulting in a 39 dimensional feature vector.

The recognizer is the one provided by Aurora-2 based on the HTK toolkit [31]. It uses whole word acoustic models trained on clean speech. Each digit is modeled by means of a 16-state continuous HMM with 3 Gaussians per state. The silence and short pause are modeled by means of HMMs with 3 and 1 states, respectively, and 6 Gaussians per state.

A different VQ codebook is trained for every available training condition using only the static speech features (12 MFCCs and the log-energy). Training is carried out by means of a k-means algorithm applying the weighted Euclidean distance defined in eqn. (20). From this set of codebooks, the compensation parameters are estimated for every VQ-based MMSE method proposed in this section using stereo data. These compensation parameters take into account the possible transformations (due to the environmental noise) between the clean feature space and the distorted one, both modeled by means of VQ codebooks with the same number of cells (i.e., $M = N$). The utterance compensation follows a frame-by-frame basis, where an estimate of the clean speech parameters is computed using only the static speech features from each frame. Finally, the dynamic parameters (delta and delta-delta) are computed for each frame using the previously compensated static features. In order to compare our proposal with other MMSE-based techniques (SPLICE and MEMLIN), GMMs are also trained. Thus, one GMM with diagonal covariance is estimated for every available training condition using iterative splitting (only for the static speech features).

Table II shows the performance in terms of word accuracy (WAcc) (%) achieved by different techniques in the aforementioned recognition task. For each technique, the average WAcc (%) of the 9 noises is shown for every SNR (i.e., from the clean condition to -5 dB). In addition, the average WAcc between 0 dB and 20 dB is considered in column *Avg*.

The *baseline* results are obtained applying acoustic models trained with clean speech and no compensation. As expected, the addition of noise to speech seriously degrades the performance of the recognizer. In order to obtain an upper bound estimate of the recognition accuracy, we also test a *matched* system which employs a set of acoustic models trained under the same noise conditions as for testing.

In the case of the MMSE techniques, the experiments are carried out using oracle information about the environmental noise, i.e., each utterance is compensated using a set of compensation parameters trained under an identical noise condition. It is worth noting that in practical situations this information is not available. However, the oracle results will give us an estimate of the best performance that could be expected from each method. In Section V we will extend the proposed methods in order to deal with unknown environmental conditions.

The VQ-based MMSE techniques are tested using 256 centroids per VQ codebook. As can be seen, all these techniques greatly improve the performance achieved by the baseline system. The successive refinements in the VQ-based approach are reflected in a better recognition accuracy. The results show that all the bias-derived techniques (BB-MMSE, RB-MMSE, dMV-MMSE and fMV-MMSE) achieve significant improvements in comparison with the classic FD-MMSE method.

Not surprisingly, the more accurate correction vectors used in RB-MMSE provides a performance improvement with respect to BB-MMSE. Moreover, the compensation of both mean and variance in the methods dMV-MMSE and fMV-MMSE provides further improvements in the recognition accuracy. As can be observed, in this case the use of full covariance matrices in the normalization achieves the best recognition results. However, the little gain in recognition accuracy might not be justified by the extra computational burden introduced by fMV-MMSE. In addition, full covariance matrices are coarsely estimated in case of insufficient training data.

For comparison purposes, Table II also shows the recognition results achieved by SPLICE and MEMLIN. For the sake of a fair comparison with the VQ-based estimation methods, both are tested using 256 Gaussians per GMM. The slight degradation in performance of RB-MMSE regarding MEMLIN (RB-MMSE is the proposed technique more similar to MEMLIN) is justified by the hard decision of VQ in contrast to the soft selection of MEMLIN. However, it must be pointed out that RB-MMSE is computationally simpler than MEMLIN

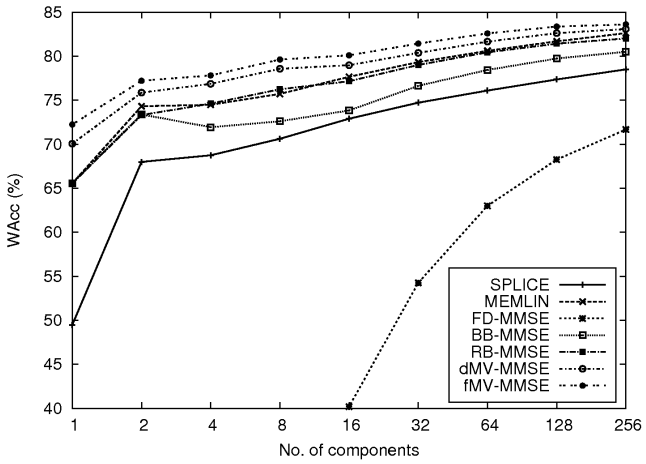

Fig. 3. Recognition accuracy for different MMSE-based methods regarding the number of components (Gaussians or VQ cells) employed.

as shown in Table I. Hard decision (HD) versions of SPLICE and MEMLIN are also tested. In these versions only the most likely Gaussian is involved in the estimation. This way, we approximate the behavior of VQ by means of GMM. As can be seen, SPLICE and MEMLIN suffer a performance reduction when making hard decision. Furthermore, we can see that RB-MMSE provides better results than HD-MEMLIN. Therefore, we show that, in spite of the hard decision carried out by VQ, the performance of our techniques is comparable to those obtained the GMM-based approaches.

Fig. 3 shows the performance in terms of WAcc for the different MMSE-based estimators versus the number of components (Gaussians or VQ cells) used for the compensation. As expected, the more components are used, the better are the recognition results obtained since the feature spaces are better represented and the compensation parameters are better estimated. However, negligible improvements are obtained with codebooks greater than 256 components. The case of FD-MMSE is slightly different since this method introduces a quantization error. Thus, if we consider FD-MMSE along with 1-centroid VQ codebooks, all the frames in the utterance are substituted by the mean feature vector in the clean space. This case provides a 7.67 % of WAcc in the recognition accuracy. On the contrary, the other methods work as basic compensation methods in this trivial case (e.g., BB-MMSE with a 1-centroid VQ codebook for each feature space can be viewed as a noise-dependent CMN technique).

IV. HMM-BASED MMSE ESTIMATION

Let us recall the general formula proposed in equation (23) for the VQ-based MMSE estimation. This and other MMSE approaches (e.g., SPLICE and MEMLIN) for robust speech recognition do not consider the time evolution of speech in the estimation. However, it is well known that, modeling the source evolution, time correlations can be exploited and better features estimates can be obtained [25]. By means of this explicit modeling, $P(C_X^{(i)}|C_Y^*)$ in eqn. (23) is modified as $P(C_X^{(i)}|\mathcal{Y}_t)$, where $\mathcal{Y}_t = (\boldsymbol{y}_{t-k1}, \ldots, \boldsymbol{y}_t, \ldots, \boldsymbol{y}_{t+k2})$ is a sequence of feature vectors used at time t. This sequence is obtained by applying a time window around the current input vector $\boldsymbol{y}_t$ starting at time $t-k_1$ and ending at $t+k_2$. In general, k_1 and k_2 depend on t. For example, if the window takes the whole utterance, then $k_1(t) = t+1$ and $k_2(t) = T-t$. With the proposed modifications, we can rewrite eqn. (23) as,

$$\hat{\boldsymbol{x}}_t = \sum_{i=1}^{M} E\left[\boldsymbol{x}\left|C_X^{(i)}, C_{Y,t}^*, \boldsymbol{y}_t\right.\right] P\left(C_X^{(i)}\left|\,\mathcal{Y}_t\right.\right) \qquad (41)$$

where $\hat{\boldsymbol{x}}_t$ is the clean vector estimate obtained at time t and $C_{Y,t}^*$ is the VQ cell which the input feature vector $\boldsymbol{y}_t$ belongs to.

In order to obtain the required probabilities $P(C_X^{(i)}|\mathcal{Y}_t)$ we will employ an HMM as a source model. HMMs are versatile tools that have been successfully applied to reconstruction of unreliable data. For example in [32] HMM-based algorithms for reconstructing unreliable spectral components are proposed. These algorithms exploit the time and spectral frequency correlations of the speech signal in order to improve the reconstruction. Similarly, in [33], HMM-based methods are developed to compensate distorted speech signal in the MFCC-domain. Another field where HMMs have also been successfully employed is in digital transmission. Works such as [25], [34]–[37] combat the speech distortions due to channel degradation in network transmissions by modeling speech signal dynamics with HMMs.

In this work, ergodic HMMs are employed to model the effects of the environmental noise in the speech signal. An HMM λ is trained for every environmental condition and consists of a set of states where each state represents a clean VQ cell $C_X^{(i)}$ $(i = 1, \ldots, M)$. It is important to note the discrete nature of the set of states in an HMM and how the discretization applied in our estimation framework to the clean feature vector space allows a straightforward application of HMMs. An HMM is defined as a tuple $\lambda = \langle \boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi} \rangle$, where $\boldsymbol{A}$ is the matrix with the transition probabilities between states $a_{ij} = P(s_t = C_X^{(j)}|s_{t-1} = C_X^{(i)})(i, j = 1, \ldots, M)$ (i.e., the probability that the state at time t (s_t) is $C_X^{(j)}$, given that the previous state (s_{t-1}) was $C_X^{(i)}$), matrix $\boldsymbol{B}$ provides the observation probabilities $b_i(\boldsymbol{y}_t \in C_Y^{(j)}) = P(C_Y^{(j)}|C_X^{(i)})$ $(i = 1, \ldots, M; j = 1, \ldots, N)$ (i.e., the probability of the noisy cell $C_Y^{(j)}$ given that the current state is the clean cell $C_X^{(i)}$), and $\boldsymbol{\pi}$ is a vector with the *a priori* probability of each state $\pi_i = P(C_X^{(i)})$.

In the same way as we did for the VQ-based estimation, the HMM-based one employs a discrete set of prototypes for the noisy feature space, which involves a partition of this space into a set of VQ cells $C_Y^{(j)}$ $(j = 1, \ldots, N)$. Again, this discretization will lead us to a computationally simpler estimator that will not involve any performance reduction when a sufficient number of VQ cells is employed. It must be considered that, unlike the clean space, this discretization is not mandatory in the proposed HMM framework and continuous HMMs could be alternatively applied (at the cost of a increased computational complexity).

Applying the defined model, the conditional probabilities used in the MMSE estimate of eqn. (41) are approximated

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg.
fMV-MMSE	99.02	98.37	97.15	92.88	79.61	50.04	20.89	83.61
FD-HMMSE	97.12	94.36	92.01	86.22	71.45	44.28	20.12	77.66
BB-HMMSE	99.22	97.50	95.82	90.82	77.37	50.92	22.64	82.49
RB-HMMSE	99.22	98.27	96.94	92.79	80.63	53.99	23.47	84.52
dMV-HMMSE	99.22	98.28	96.98	93.03	81.46	55.70	24.32	85.09
fMV-HMMSE	99.22	98.25	96.98	93.05	81.46	55.87	24.18	85.12

TABLE III

PERFORMANCE, IN WACC (%), OF THE PROPOSED HMM-BASED COMPENSATION METHODS. CLEAN ACOUSTIC MODELS ARE APPLIED AND 256-COMPONENTS VQ CODEBOOKS ARE USED. THE RESULTS ARE DETAILED FOR EACH SNR. THE AVERAGE WACC (AVG.) BETWEEN 0 DB AND 20 DB IS ALSO SHOWN.

by $P(C_X^{(i)}|\mathcal{Y}_t) \approx P(s_t = C_X^{(i)}|\mathcal{Y}_t^*)$. This is the probability of being at the HMM state $C_X^{(i)}$ at time t given $\mathcal{Y}_t^*$, which is the sequence of VQ cells corresponding to $\mathcal{Y}_t$, i.e., $\mathcal{Y}_t^* = \{C_{Y,t-k1}^*, \ldots, C_{Y,t}^*, \ldots, C_{Y,t+k2}^*\}$. This posterior can be efficiently computed by means of the forward-backward algorithm [25], [38] as,

$$P\left(s_t = C_X^{(i)} \middle| \mathcal{Y}_t^*\right) = \gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{M}\alpha_t(j)\beta_t(j)} \quad (42)$$

where $\alpha_t(i)$ and $\beta_t(i)$ $(i = 1, \ldots, M)$, known as the forward and backward variables, respectively, model the following distributions,

$$\alpha_t(i) = P\left(s_t = C_X^{(i)} \middle| C_{Y,t-k1}^*, \ldots, C_{Y,t}^*\right) \quad (43)$$

$$\beta_t(i) = P\left(C_{Y,t+1}^*, \ldots, C_{Y,t+k2}^* \middle| s_t = C_X^{(i)}\right) \quad (44)$$

These probabilities can be computed through the following forward and backward recursions,

$$\alpha_t(i) = \left[\sum_{j=1}^{M}\alpha_{t-1}(j)a_{ji}\right] b_i(\boldsymbol{y}_t)/K_t \quad (45)$$

$$\beta_t(i) = \sum_{j=1}^{M} a_{ij}b_j\left(\boldsymbol{y}_{t+1}\right)\beta_{t+1}(j) \quad (46)$$

where K_t is a normalization factor at time t and $b_i\left(\boldsymbol{y}_t\right)$ is the observation probability of the feature vector $\boldsymbol{y}_t$ given the clean cell $C_X^{(i)}$. This probability is approximated by the conditional probability of $C_{Y,t}^*$ given $C_X^{(i)}$, i.e., $b_i(\boldsymbol{y}_t) \approx P(C_{Y,t}^*|C_X^{(i)})$.

The initialization of both variables is,

$$\alpha_{t-k1}(i) = \pi_i b_i(\boldsymbol{y}_{t-k1})/K_{t-k1} \quad (47)$$

$$\beta_{t+k2}(i) = 1 \quad (48)$$

Finally, the HMM-based MMSE estimation formula of eqn. (41) becomes,

$$\hat{\boldsymbol{x}}_t = \sum_{i=1}^{M} E\left[\boldsymbol{x}\middle|C_X^{(i)}, C_{Y,t}^*, \boldsymbol{y}_t\right] P\left(s_t = C_X^{(i)}\middle|\mathcal{Y}_t^*\right) \quad (49)$$

A. Results

In order to evaluate the proposed HMM-based MMSE estimator, the same experimental framework described in subsection III-F is used. Five different methods are derived. These methods differ in the computation of the expected value $E[\boldsymbol{x}|C_X^{(i)}, C_{Y,t}^*, \boldsymbol{y}_t]$ of eqn. (49). In the first method, noted as *fully discrete HMM-based MMSE estimation* (FD-HMMSE), the expectation term is computed by means of eqn. (26). The second one, which is referred to as *basic bias HMM-based MMSE estimation* (BB-HMMSE), employs eqn. (33) to this end. The use of the more accurate bias computation of eqn. (35) involves a new method called *refined bias HMM-based MMSE estimation* (RB-HMMSE). Finally, the two last estimation methods use eqn. (36) in the expectation computation. We will refer to these methods as *mean and diagonal covariance HMM-based MMSE estimation* (dMV-HMMSE) when applying diagonal covariance matrices, and *mean and full covariance HMM-based MMSE estimation* (fMV-HMMSE) in the case of full covariance matrices.

Table III shows the recognition results achieved by these methods. In all cases, clean acoustic models are used and oracle information about the environmental noise is assumed. VQ codebooks with 256 centroids are applied. The results are obtained initializing the forward-backward algorithm at the beginning and end of each utterance, i.e., all the feature vectors of the utterance are used for the computation of the conditional probability $P(s_t = C_X^{(i)}|\mathcal{Y}_t^*)$ of eqn. (49). For comparison purposes, the results obtained applying fMV-MMSE, which is the best VQ-based method of Table II, are also included.

The results show that considering temporal correlations of the speech signal improves the estimates in all the cases. As we commented above, these results are obtained using the whole utterance in the probability calculation of the forward-backward algorithm. However, in real-time applications, where the time of response is critical, the delay introduced by this algorithm could be undesirable, specially for long utterances. For this reason, we propose two bounded delay versions of the forward-backward algorithm. The first one consists of using a symmetric window of radius δ centered at time t. Thus, the forward-backward algorithm only employs the observations $\mathcal{Y}_t = \{\boldsymbol{y}_{t-\delta}, \ldots, \boldsymbol{y}_t, \ldots, \boldsymbol{y}_{t+\delta}\}$ in order to compensate the feature vector $\boldsymbol{y}_t$. The second version takes advantage of the fact that only the backwards probabilities involve a delay in the computation. Therefore, this second version initializes the forward probabilities at the beginning of the utterance with the objective of exploiting all the previous history of the observed signal. Under this approach, the window employed to compensate $\boldsymbol{y}_t$ is composed of $\mathcal{Y}_t = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_t, \ldots, \boldsymbol{y}_{t+\delta}\}$. Both approaches introduce a delay of δ frames and they will be referred to as symmetric window (SW) and asymmetric window (AW) forward-backward algorithms.

Table IV shows the average results obtained by the two

	Delay (frames)						
	0	1	3	5	15	25	50
SW	83.08	85.07	85.18	85.16	85.12	85.10	85.09
AW	84.59	84.83	84.97	85.03	85.11	85.09	85.09

TABLE IV
AVERAGE WACC RESULTS FOR THE TWO BOUNDED DELAY VERSIONS OF THE DMV-HMMSE METHOD.

	MMSE		HMMSE	
	Avg.	Imp.	Avg.	Imp.
SPLICE	83.82	6.78	-	-
MEMLIN	87.67	6.11	-	-
FD	73.65	-0.35	83.74	7.83
BB	83.75	4.04	87.37	5.92
RB	86.28	5.21	88.19	4.34
dMV	87.00	4.72	88.50	4.00
fMV	87.31	4.43	88.66	4.16

TABLE V
AVERAGE RECOGNITION RESULTS (AVG.) AND RELATIVE IMPROVEMENT (IMP.) ACHIEVED BY THE DIFFERENT TECHNIQUES WHEN THE STATIC AND DYNAMIC FEATURES OF THE SPEECH ARE COMPENSATED.

versions of the forward-backward algorithm. In both cases the compensation method is dMV-HMMSE with 256 centroids per VQ codebook. This method is chosen because its results are comparable to those ones obtained by the best method (fMV-HMMSE), but with a lower computational complexity. The results are obtained for different values of delay measured in number of frames (every frame involves 10 ms). As shown, the SW version with no delay (0 frames) obtains the same results as dMV-MMSE, since the temporal information is ignored (see Table II). This is not the case of AW with no delay, which provides better recognition results since all previous frames are considered in addition to the current one. A significant improvement over the baseline is achieved when short time correlations (1, 3, and 5 frames) are considered. As the delay increases, the recognition accuracy converges to the one obtained using the whole utterance (85.09 %). The fact of considering all past frames in the AW algorithm does not involve any improvements regarding SW. These results support the idea that the speech model (ergodic HMM) does not properly capture the long-term correlations. Therefore, we can achieve a good compromise between complexity, delay, and recognition accuracy by means of an SW approach with few frames ($\delta = 3, 5$).

B. Results compensating the dynamic features

We have seen that the introduction of *a priori* information about the speech dynamics during the MMSE estimation yields a significant improvement in the recognition accuracy. Along with HMMs, another efficient and simple way to exploit the short temporal correlations of the speech is by means of dynamic features. That is, the static and dynamic features of the speech signal can be jointly modeled by means of VQ codebooks so the whole input vector (including the static and dynamic features) is compensated by means of the different MMSE proposals.

Table V shows the average results obtained for every proposed technique when the dynamic features are taken into account in the estimation. In addition, the relative improvement (*Imp.*), in percent, regarding the case of only compensating the static features is also shown. VQ codebooks with 256 centroids are applied in all the cases. As can be seen, significant improvements are achieved when the dynamic features are incorporated into the MMSE estimation. The only exception is the FD-MMSE technique which suffers a negligible relative performance reduction of 0.35 %. Note that in this case (the only one which actually applies an VQ quantization) we are employing the same number of VQ cells to model spaces with more dimensions (from 13 to 39 features), so that the quantization error is increased. For the rest of the techniques, the recognition results are consistent. Therefore, the introduction of the dynamic features in the MMSE estimations improves the VQ-based baselines, and further improvements are obtained when HMMs are applied in the estimation process.

V. ACCOUNTING FOR UNKNOWN ENVIRONMENTAL NOISES

The compensation techniques presented until now assume that the characteristics of the environmental noise that contaminates the speech signal are available. However, this assumption is not true in general. In realistic scenarios the environmental noise is frequently unknown and time-varying. In order to consider this fact, we present a method to obtain clean speech estimates for unknown environments. Although this method provides us with suitable estimates of the clean vectors, we must take into account that they cannot be considered fully reliable. For this reason, at the end of this section we also show how to improve the recognition performance by providing reliability measures about the estimation process to the recognizer.

A. Multiple-model based compensation

In the literature a vast number of methods to cope with unknown environmental noises can be found. Some of these techniques represent the noise effects by means of a model which is updated during the silence periods of the speech signal. In order to do so, this type of techniques employs a voice activity detector (VAD) to identify silence segments. The recognizer adaptation to the testing conditions is usually carried out by combining the clean speech acoustic models with the aforementioned noise model. Methods such as parallel model combination (PMC) [6] perform this combination in the filter bank domain, thus a considerable amount of computational resources for the conversion between this domain and cepstrum is required. In addition, the performance of these techniques directly relies on the VAD accuracy, which tends to decrease for low-SNR scenarios.

Another alternative, which avoids some of the aforementioned drawbacks, is the multiple-model framework [26], [27]. Under this approach, a set of feature-domain compensation vectors is trained off-line for several environmental noise conditions. In the case of the MMSE-based techniques, each compensation set accounts for the speech signal transformation for an specific noise condition. Thus, an estimate can be computed for every environment using its corresponding

compensation vectors. The final clean feature vector is obtained as a combination of the estimates computed for every environment. The advantages of this approach, as opposed to others, are mainly two. First, it provides high performance for noise conditions considered during the training stage [26]. Thus, it is usually applied in applications where the total number of environments is bounded, e.g., in-car recognition [27]. Second, its computational complexity is fair less than that of other methods, since the compensation can be obtained as a combination of the off-line precomputed compensation vectors [27].

Applying the set of feature-domain compensation vectors computed for environment e, a clean speech estimate $\hat{\boldsymbol{x}}_t^e$ can be obtained for this environment. In our case, this estimate is obtained by means of one of the MMSE techniques (VQ- or HMM-based) described in Sections III and IV. The final clean speech estimate is computed as a weighted combination of the estimates obtained for all the environments as,

$$\hat{\boldsymbol{x}}_t = \sum_e P(e|\boldsymbol{y}_t)\hat{\boldsymbol{x}}_t^e \tag{50}$$

where $P(e|\boldsymbol{y}_t)$ is the *a posteriori* probability of environment e. This posterior is obtained by a classifier which assigns a probability for every environment given the noisy observation $\boldsymbol{y}_t$. A common approach for computing this posterior is by assuming that each environment e is modeled by means of a GMM as follows [21], [27],

$$p_e(\boldsymbol{y}) = \sum_{k_y^e} P(k_y^e)\mathcal{N}\left(\boldsymbol{y}; \boldsymbol{\mu}_{k_y^e}, \boldsymbol{\Sigma}_{k_y^e}\right) \tag{51}$$

where k_y^e denotes a Gaussian component of the GMM for the environment e, and $\boldsymbol{\mu}_{k_y^e}$ and $\boldsymbol{\Sigma}_{k_y^e}$ are the mean vector and the covariance matrix of this Gaussian, respectively. Using the GMMs trained for every environment, the posterior $P(e|\boldsymbol{y}_t)$ is obtained as follows,

$$P(e|\boldsymbol{y}_t) = \frac{p_e(\boldsymbol{y}_t)}{\sum_e p_e(\boldsymbol{y}_t)} \tag{52}$$

B. Dealing with estimate uncertainty

The clean speech estimation carried out by eqn. (50) allows the proposed VQ- and HMM-based MMSE estimators to deal with unknown environmental noises, even when these noises are not considered during the training stage. In this regard, it is expected that the environment classifier returns high probabilities $P(e|\boldsymbol{y}_t)$ for those learned environments that are more similar to the unknown ones. Nevertheless, the more the test environment differs from the training ones, the poorer will be the performance of the recognizer. Increasing the number of training environments in order to reduce the mismatch between training and testing is not a desirable solution, since it would increase the computational cost. Hence, we must take into account that the resulting clean estimates may be not fully reliable. For this reason, we consider two methods devoted to incorporate the uncertainty of the estimates in the recognition decoding process: soft-data decoding (SD) and weighted Viterbi algorithm (WVA). In a previous work [37] we successfully applied these methods to packet loss concealment for remote speech recognition.

Both methods, SD and WVA, rely on the probability distributions employed for the computation of the MMSE estimates. Thus, we will consider, for a given environment e and time instant t, the following distributions: $P(C_X^{(i)}|\mathcal{Y}_{e,t}^*, e)$ for the HMM-based estimates of eqn. (49) and $P(C_X^{(i)}|C_{Y,e,t}^*, e)$ for the VQ-based estimates of eqn. (23). It must be pointed out that while the VQ codebooks for the noisy feature spaces are different for every environment, the VQ codebook that models the clean feature space is shared by all the environments and it can be viewed as a common anchor to model the transformations between the noisy and clean features spaces. We will use this property to compute the reliability of the clean speech estimates.

As a first step to derive SD and WVA approaches, we assume that the recognition process is carried out by the Viterbi algorithm (VA) [1], [38]. Thus, the state metrics update equation used in the decoding stage is

$$\phi_t(s_j) = \max_{s_i}\left\{\phi_{t-1}(s_i)a_{ij}\right\} p(\boldsymbol{x}_t|s_j) \tag{53}$$

where s_i and s_j are states of the HMM acoustic models, $\phi_t(s_j)$ is the maximum likelihood of observing the feature vector $\boldsymbol{x}_t$ in state s_j at time instant t, and a_{ij} and $p(\boldsymbol{x}_t|s_j)$ correspond to the transition and observation probabilities of the acoustic model, respectively.

The first method devoted to incorporate the uncertainty of the estimates in the recognition process is soft-data decoding (SD) [1], [29], [37]. Under this approach instead of assuming that we are dealing with deterministic data, we consider that the clean vector estimate has an associated Gaussian evidence pdf so that the observation probability used in the VA is modified. Thus, the uncertainty of the estimation is taken into account by adding the covariances $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_t}$ of the MMSE estimate to the covariances of the Gaussian mixtures which model every HMM state from the recognizer. Finally, the observation probability for a given state s is computed as follows,

$$p(\boldsymbol{x}_t|s) = \sum_k P(k|s)\mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_s^{(k)}, \boldsymbol{\Sigma}_s^{(k)} + \boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_t}\right) \tag{54}$$

where $\boldsymbol{\mu}_s^{(k)}$ and $\boldsymbol{\Sigma}_s^{(k)}$ are the mean vector and covariance matrix of the k^{th} Gaussian component of the GMM which models state s.

Assuming Gaussian pdfs with diagonal covariance matrices, the variance of the j^{th} feature in the clean vector estimate can be easily computed as,

$$\sigma_{\hat{\boldsymbol{x}}_t}^2(j) = E\left[\left(\boldsymbol{x}_t(j) - \hat{\boldsymbol{x}}_t(j)\right)^2 |\Lambda_t\right] \tag{55}$$

where Λ_t is the *a priori* information applied in the above expectation. In the case of the HMM-based estimates, we can apply the following approximation,

$$\sigma_{\hat{\boldsymbol{x}}_t}^2(j) \approx \sum_e \sum_{i=1}^{M} P\left(C_X^{(i)}, e\middle|\mathcal{Y}_{e,t}^*\right)\left(E\left[\boldsymbol{x}(j)\middle|C_X^{(i)}, C_{Y,e,t}^*, \boldsymbol{y}_t, e\right] - \hat{\boldsymbol{x}}_t(j)\right)^2 \tag{56}$$

where $E\left[\boldsymbol{x}(j)\middle|C_X^{(i)}, C_{Y,e,t}^*, \boldsymbol{y}_t, e\right]$ is the expected value employed for the computation of the HMM-based MMSE estimate (see eqn. (49)) and the joint probability $P(C_X^{(i)}, e|\mathcal{Y}_{e,t}^*)$

can be approximated by

$$P\left(C_X^{(i)}, e\middle|\,\mathcal{Y}_{e,t}^*\right) \approx P\left(C_X^{(i)}\middle|\,\mathcal{Y}_{e,t}^*, e\right) P(e|\boldsymbol{y}_t) \tag{57}$$

Similar expressions could be obtained for the VQ-based estimates.

WVA is an alternative way to introduce reliability information of the estimates during Viterbi decoding [37], [39], [40]. The idea is to incorporate a time-varying weighting factor $\rho_t \in [0,1]$ that reduces the contribution of unreliable feature vectors in the decoding stage. Thus, eqn. (53) is modified as follows,

$$\phi_t(s_j) = \max_{s_i}\left\{\phi_{t-1}(s_i)a_{ij}\right\} p(\boldsymbol{x}_t|s_j)^{\rho_t} \tag{58}$$

As can be seen, WVA is a very efficient method to incorporate the reliability of the estimates in the decoding process provided that the weighting factor ρ_t becomes a simple multiplication in the logarithmic domain. This made WVA very attractive in comparison with soft-data decoding.

Our problem now is to determine a reliability function for the MMSE estimates. The reliability of an estimate $\hat{\boldsymbol{x}}_t$ directly depends on the probability distribution obtained during the MMSE estimation. In such a way, the flatter the obtained distribution is, the less reliable is the estimate. Thus, it is expected that when this distribution approaches a delta, the estimate should be close to the original clean vector. In such a case, we can consider that the estimate is fully reliable. In this work, we propose to use the marginal distributions of the clean VQ cells for this purpose. Using the joint distribution of eqn. (57), the marginal distribution of the clean cell $C_X^{(i)}$ at time t can be approximated by

$$\begin{aligned} P_t\left(C_X^{(i)}\right) \equiv P_t\left(C_X^{(i)}\middle|\,\mathcal{Y}_t\right) &= \sum_e P\left(C_X^{(i)}, e\middle|\,\mathcal{Y}_t\right) \\ &\approx \sum_e P\left(C_X^{(i)}, e\middle|\,\mathcal{Y}_{e,t}^*\right) \end{aligned} \tag{59}$$

In order to measure the uncertainty of the above marginal distribution, in a previous work [37] we proposed to use the entropy function. Thus, we define the instantaneous entropy $H_t(\hat{\boldsymbol{x}})$ at time t as,

$$H_t(\hat{\boldsymbol{x}}) = -\sum_{i=1}^{M} P_t\left(C_X^{(i)}\right)\log_2 P_t\left(C_X^{(i)}\right) \tag{60}$$

Now, we can easily define a relation with the reliability factor ρ_t by means of the following expression,

$$\rho_t = 1 - \left(\frac{H_t(\hat{\boldsymbol{x}})}{\log_2 M}\right)^{\varphi} \tag{61}$$

where $\log_2 M$ is the maximum entropy that a discrete probability distribution with M different values can take and φ is a factor which is experimentally determined.

As can be noted, when the MMSE estimation does not provide information to estimate the clean feature vector, $P_t(C_X^{(i)})$ will present a uniform distribution, so that the entropy is maximum and equals to $\log_2 M$. In this case, the estimate is considered unreliable by setting $\rho_t = 0$ and the decoding process of eqn. (58) is only guided by the transition probabilities. In the opposite case, when $P_t(C_X^{(i)})$ is a delta function, then the corresponding entropy is zero, and the reliability factor becomes one, so that eqn. (58) becomes the usual Viterbi algorithm.

C. Results

In order to evaluate the proposed techniques under unknown environments, two different test sets are defined. The first test set, called *test Set A*, is intended to show the performance of the different methods when considering the same environments used for training. This test set consists of the same 55 environmental conditions (9 noises at 6 SNR values plus a clean condition) as described in subsection III-F. The second test set, called *test Set B*, is created in the same way, but using five new different noises (pedestrian square, car, bus station, heavy sea, and heavy traffic avenue) at 5 new different SNRs (17.5, 12.5, 7.5, 2.5, and -2.5 dB). Thus, we can evaluate the influence of considering different environments than the ones used for training.

For training the compensation methods, the same 55 environments considered in the oracle experiments are used. The clean condition is also considered as another environment. This allows a sort of interpolation of SNRs not considered in the training stage. Every environment is modeled by means of a GMM with 256 Gaussians and diagonal covariance matrices. It must be pointed out that these GMMs are the same as those employed in SPLICE and MEMLIN although a more sophisticated environment modeling could be applied. Full-dimension feature vectors (including static and dynamic features) are employed. Factor φ (see eqn. (61)) has been heuristically set to 0.1 in the WVA approach.

Two sets of acoustic models are defined. First, acoustic models are trained on clean speech as usual. Second, multi-style trained (MST) acoustic models are also considered. Such models are trained with distorted speech in order to robust the ASR system against noisy conditions. The data employed to train the MST models correspond to the 9 noise types considered until now at 4 different SNRs (20, 15, 10, and 5 dB) plus a clean condition (37 acoustic environments in total). These data are compensated with the corresponding MMSE estimator and used to train the MST models. Finally, the multiple-model approach described in Section V-A is applied in tandem with MST models during decoding.

Table VI shows the recognition results achieved by different MMSE techniques under the multiple-model framework for the two aforementioned test sets. The average word accuracy obtained for the two sets (Avg.) and the relative improvement over the baseline (Imp.), in percent, are also shown. As in the previous experiments, the baseline systems use acoustic models (clean or MST models) trained with no compensation.

As can be seen, SPLICE and MEMLIN suffer a performance degradation regarding the oracle results when clean models are applied (see Table V). This degradation is caused by mismatches in the environment identification. However, our method dMV-HMMSE does not suffer any degradation. The introduction of temporal information through HMM modeling achieves better compensations for every learned environment,

		Set A	Set B	Avg.	Imp.
Clean models	Baseline	50.83	40.28	45.56	–
	SPLICE	83.07	73.90	78.49	72.28
	MEMLIN	86.77	75.52	81.15	78.12
	dMV-HMMSE	89.10	77.69	83.40	83.06
	dMV-HMMSE+SD	89.36	79.72	84.54	85.56
	dMV-HMMSE+WVA	89.83	81.41	**85.62**	87.93
MST models	Baseline	88.73	79.78	84.26	–
	SPLICE	88.87	80.92	84.90	0.77
	MEMLIN	89.78	80.92	85.35	1.30
	dMV-HMMSE	90.86	82.73	86.80	3.01
	dMV-HMMSE+SD	90.67	82.99	86.83	3.06
	dMV-HMMSE+WVA	90.82	85.34	**88.08**	4.54

TABLE VI
AVERAGE WORD ACCURACY RESULTS (%) OF THE MULTIPLE-MODEL APPROACHES FOR SET A AND SET B TESTING CONDITIONS (INCLUDING THE AVERAGE RESULTS IN THE TWO CONDITIONS AND THE RELATIVE IMPROVEMENT (%) OVER THE BASELINE) IN CLEAN AND MST ACOUSTIC MODELS.

which can be understood as a better identification of the instantaneous SNR. We must consider that, although in the oracle experiments every utterance is compensated using an MMSE estimator trained under its same environment (environmental noise and global SNR value), a better compensation could be obtained if we consider the instantaneous SNR. For example, given a particular feature vector, it could be better to compensate it with an MMSE estimator trained under the same noise but at a different SNR value.

As expected, the relative improvement achieved by feature compensation is smaller when MST models are used. In addition, all the methods yield poorer recognition results for test *Set B*. This is one of the lacks of the multiple-model framework: the performance drops in mismatch situations [26]. Nevertheless, we can counteract this reduction in performance by employing reliability measures of the MMSE estimates in the decoding stage of the recognizer. The relative improvement achieved by the SD approach (dMV-HMMSE+SD) for *Set B* regarding do not taking into account the reliability of the estimates (dMV-HMMSE) is 2.61 % in clean training and 0.31 % in MST training. The WVA approach (dMV-HMMSE+WVA) yields relative improvements of 4.79 % and 3.15 % for clean and MST models, respectively. On the other hand, the improvement achieved by SD and WVA for test *Set A* is smaller since the uncertainty of the estimates in non-mismatch conditions is lower. Furthermore, a negligible performance reduction appears when MST models are applied. Finally, it can be noted that WVA yields better recognition results than the SD approach. This could be explained by the lack of validity of some adopted assumptions in the SD approach (i.e., the estimates are Gaussian distributed) [37].

VI. CONCLUSIONS

In this paper, we have presented a novel framework for feature compensation based on MMSE estimation and stereo data for robust speech recognition. As a result, a piecewise linear mapping between the noisy feature space and the clean one is obtained. Initially, both feature spaces are modeled by means of VQ codebooks. Five different VQ-based MMSE estimators are derived from this framework: FD-MMSE, BB-MMSE, RB-MMSE, dMV-MMSE, and fMV-MMSE. These methods make different assumptions, so that different compensation functions are obtained. In addition, a novel subregion-based approach is proposed in order to accurately model the transformation introduced by acoustic noise in the speech. Although only applied to the proposed VQ-based estimators, the subregion modeling can also be extended to other approaches (e.g., GMM-based estimators). The relationship between our technique and other MMSE-based methods has also been discussed. The experimental results show the importance of modeling both feature spaces (clean and noisy) in order to obtain a more accurate probability model for the MMSE estimation. Furthermore, the compensation of feature vectors taking into account more complex transformations introduced by the noise (e.g., mean and covariance modifications), leads to further improvements.

In order to exploit the temporal correlations of the speech in the estimation, the proposed techniques are extended in two ways. Firstly, the short-term correlations are taken into account by means of the dynamic features, so that the whole input feature vector (including the static and dynamic features) is compensated. Secondly, the speech evolution is modeled by means of an HMM. This HMM acts as a source model which helps to improve the probability distributions required to obtain the MMSE estimates. When these extensions are jointly applied to our estimators, an averaged relative improvement of 5.25 % is achieved.

Finally, a multiple-model framework is proposed in order to deal with unknown environmental conditions. Using this framework, a set of clean feature vector estimates is computed for a set of learned environments. The final estimate is obtained as a linear combination of the previous environment-specific estimates weighted by the *a posteriori* probabilities provided by an environment classifier. In addition, two techniques to improve the recognizer performance by exploiting the uncertainty of the estimation process are derived. The first technique is based on a soft-data approach where the covariances of the estimates are added to the covariances of the GMM mixtures that model every HMM state in the recognizer. The second technique modifies the Viterbi algorithm by including a weighting factor that reduces the contribution of unreliable estimates in the decoding stage. In order to test the proposed techniques, two noisy digit recognition tasks are defined. The first one contains the same environmental conditions used for training. In the second one, a new set of noises are used to artificially contaminate the test utterances. When the reliability measures are not considered, the proposal leads to relative improvements of 83.06% and 3.01% for clean and MST models, respectively, over the uncompensated results. Further improvements are achieved by employing reliability measures. In this case, we obtain improvements of 85.56% and 87.93% for the soft-data and weighted Viterbi approaches, respectively, when clean acoustic models are applied. In MST training, improvements of 3.06% and 4.54% are achieved by soft-data and weighted Viterbi, respectively. Furthermore, important improvements over other well-known MMSE techniques (e.g., SPLICE and MEMLIN) have been obtained.

Future work includes improving our estimation methods for the cases when stereo data are not available. In this aspect, missing-data techniques based on imputation methods are very appealing, since few assumptions about the environmental noise which contaminates the speech signal are needed. Also, we think that an improved instantaneous environment classification could significantly contribute to a better performance.

REFERENCES

[1] A. M. Peinado and J. C. Segura, *Speech recognition over digital channels: Robustness and standars*. Wiley, July 2006.

[2] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Commun.*, vol. 16, no. 3, pp. 261 – 291, Apr. 1995.

[3] M. Benzeghiba, R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Commun.*, vol. 49, no. 10-11, pp. 763 – 786, Nov. 2007.

[4] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[5] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, vol. 9, no. 2, pp. 171 – 185, Apr. 1995.

[6] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 5, pp. 352–359, Sep. 1996.

[7] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[8] P. Loizou, *Speech enhancement: Theory and practice*. CRC, 2007.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[10] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, pp. 1304–1312, June 1974.

[11] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez-Cordoba, M. C. Benitez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 355–366, May 2005.

[12] J. C. Segura, M. C. Benitez, A. de la Torre, A. J. Rubio, and J. Ramirez, "Cepstral domain segmental nonlinear feature transformations for robust speech recognition," *IEEE Signal Process. Lett.*, vol. 11, no. 5, pp. 517–520, May 2004.

[13] P. Moreno, "Speech recognition in noisy environments," Ph.D. dissertation, Deparment of Electrical and Computer Engineering, Carnegie Mellon University, 1996.

[14] D. Y. Kim, C. K. Un, and N. S. Kim, "Speech recognition in noisy environments using first-order vector Taylor series," *Speech Commun.*, vol. 24, no. 1, pp. 39 – 49, Apr. 1998.

[15] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "Robust speech recognition using a cepstral minimum-mean-square-error-motivated noise suppressor," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 5, pp. 1061–1070, Jul. 2008.

[16] L. Deng, J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 568–580, Nov. 2003.

[17] M. Afify, X. Cui, and Y. Gao, "Stereo-based stochastic mapping for robust speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 7, pp. 1325–1334, Sep. 2009.

[18] H. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ICSLP*, 2000, pp. 29–32.

[19] A. Acero, *Acoustical and environmental robustness in automatic speech recognition*. Kluwer Academic Publishers, 1993.

[20] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora2 database," in *Proc. Eurospeech 2001*, Aalborg, Denmark, 2001, pp. 217–220.

[21] L. Buera, E. Lleida, A. Miguel, A. Ortega, and O. Saz, "Cepstral vector normalization based on stereo data for robust speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 3, pp. 1098–1113, Mar. 2007.

[22] H. Liao and M. J. F. Gales, "Issues with uncertainty decoding for noise robust automatic speech recognition," *Speech Commun.*, vol. 50, no. 4, pp. 265 – 277, 2008.

[23] X. Cui, M. Afify, and Y. Gao, "MMSE-based stereo feature stochastic mapping for noise robust speech recognition," in *Proc. ICASSP 2008*, Apr. 2008, pp. 4077–4080.

[24] J. A. Gonzalez, A. M. Peinado, A. M. Gomez, J. L. Carmona, and J. A. Morales-Cordovilla, "Efficient VQ-based MMSE estimation for robust speech recognition," in *Proc. ICASSP 2010*, Mar. 2010, pp. 4558 – 4561.

[25] A. M. Peinado, V. Sanchez, J. L. Perez-Cordoba, and A. de la Torre, "HMM-based channel error mitigation and its application to distributed speech recognition," *Speech Commun.*, vol. 41, no. 4, pp. 549 – 561, Nov. 2003.

[26] H. Xu, P. Dalsgaard, Z.-H. Tan, and B. Lindberg, "Noise condition-dependent training based on noise classification and SNR estimation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 8, pp. 2431–2443, Nov. 2007.

[27] W. Kim and J. H. L. Hansen, "Feature compensation in the cepstral domain employing model combination," *Speech Commun.*, vol. 51, no. 2, pp. 83 – 96, Feb. 2009.

[28] A. M. Peinado, V. Sanchez, J. L. Perez-Cordoba, and A. J. Rubio, "Efficient MMSE-based channel error mitigation techniques. Application to distributed speech recognition over wireless channels," *IEEE Trans. Wireless Commun.*, vol. 4, no. 1, pp. 14–19, Jan. 2005.

[29] H. Liao, "Uncertainty decoding for noise robust speech recognition," Ph.D. dissertation, University of Cambridge, 2007.

[30] *ETSI ES 201 108 - Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms*, Std., 2000.

[31] S. Young, G. Everman, M. J. F. Gales, T. Hain, D. Kershaw, D. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book - Version 3.4*. Cambridge University Engineering Department, Dec. 2006.

[32] B. J. Borgstrom and A. Alwan, "HMM-based estimation of unreliable spectral components for noise robust speech recognition," in *Proc. Interspeech*, Brisbane, Australia, Sep. 2008, pp. 1769–1772.

[33] V. Stouten, H. V. hamme, and P. Wambacq, "Model-based feature enhancement with uncertainty decoding for noise robust ASR," *Speech Commun.*, vol. 48, no. 11, pp. 1502 – 1514, Nov. 2006.

[34] B. J. Borgstrom and A. Alwan, "An efficient approximation of the forward-backward algorithm to deal with packet loss, with applications to remote speech recognition," in *Proc. ICASSP*, Apr. 2008, pp. 4425–4428.

[35] C. A. Rodbro, M. N. Murthi, S. V. Andersen, and S. H. Jensen, "Hidden Markov model-based packet loss concealment for voice over IP," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 5, pp. 1609–1623, Sep. 2006.

[36] T. Fingscheidt and P. Vary, "Softbit speech decoding: A new approach to error concealment," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 240–251, Mar. 2001.

[37] J. L. Carmona, A. M. Peinado, J. L. Perez-Cordoba, and A. M. Gomez, "MMSE-based packet loss concealment for CELP-coded speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 6, pp. 1341–1353, Aug. 2010.

[38] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.

[39] N. B. Yoma, F. R. McInnes, and M. A. Jack, "Weighted Viterbi algorithm and state duration modelling for speech recognition in noise," in *Proc. ICASSP*, vol. 2, May 1998, pp. 709–712.

[40] A. M. Gomez, A. M. Peinado, V. Sanchez, and A. J. Rubio, "Combining media-specific FEC and error concealment for robust distributed speech recognition over loss-prone packet channels," *IEEE Trans. Multimedia*, vol. 8, no. 6, pp. 1228–1238, Dec. 2006.

José A. Gónzalez received the M.Sc. degree in computing science from the University of Granada (UGR), Spain, in 2006. He is currently working towards the Ph.D. degree on statistical methods for robust speech recognition at UGR. Since 2007, he has been with the Research Group on Signals, Networking and Communications (GSTC), Department of Signal Theory, Networking and Communications, UGR, under a research grant. His research interests are in robust speech recognition and coding and signal processing.

Antonio M. Peinado (M95SM05) received the M.S. and the Ph.D. degrees in physic sciences from the University of Granada, Granada, Spain, in 1987 and 1994, respectively. Since 1988, he has been working with the GSTC Research Group, University of Granada, where has led or participated in several research projects related with speech recognition, coding and transmission. In 1989, he was a Consultant in the Speech Research Department, AT&T Bell Labs. Since 1996, he has been an Associate Professor in the Department of Signal Theory, Networking, and Communications, University of Granada. He is the author of numerous publications and coauthor of the book *Speech Recognition over Digital Channels* (Wiley, 2006), and has served as reviewer for several international journals and conferences. His research interests are in distributed and robust speech recognition, and speech and audio coding and transmission.

Angel M. Gómez received the M.A.Sc. and the Ph.D. degrees in computing science from the University of Granada, Spain, in 2001 and 2006, respectively. Since 2002, he has been with the Research Group on Signals, Networking and Communications (GSTC) of the University of Granada. Since 2006, he is an Assitant Lecturer at the Dpt. of Signal Theory, Networking and Communications of the University of Granada. His research interests are in robust speech recognition and coding and signal processing.

José L. Carmona received the Master degree from the University of Malaga, Spain, in telecommunication engineering in 2004 and the Ph.D. degree from the University of Granada, Spain, in 2009. Since 2005, he has been with the Research Group on Signals, Networking and Communications (GSTC) of the University of Granada. His research interests include speech coding, robust speech recognition, and signal processing.