MMSE-based missing-feature reconstruction with temporal modeling for robust speech recognition

José A. González, Antonio M. Peinado, *Senior Member, IEEE*, Ning Ma, Angel M. Gómez, and Jon Barker

***Abstract*—This paper addresses the problem of feature compensation in the log-spectral domain by using the missing-data (MD) approach to noise robust speech recognition, that is, the log-spectral features can be either almost unaffected by noise or completely masked by it. First, a general MD framework based on minimum mean square error (MMSE) estimation is introduced which exploits the correlation across frequency bands to reconstruct the missing features. This framework allows the derivation of different MD imputation approaches and, in particular, a novel technique taking advantage of truncated Gaussian distributions is presented. While the proposed technique provides excellent results at high and medium signal-to-noise ratios (SNRs), its performance diminishes at low SNRs where very few reliable features are available. The reconstruction technique is therefore extended to exploit temporal constraints using two different approaches. In the first approach, time-frequency patches of speech containing a number of consecutive frames are modeled using a Gaussian mixture model (GMM). In the second one, the sequential structure of speech is alternatively modeled by a hidden Markov model (HMM). The proposed techniques are evaluated on Aurora-2 and Aurora-4 databases using both oracle and estimated masks. In both cases, the proposed techniques outperform the recognition performance obtained by the baseline system and other related techniques. Also, the introduction of a temporal modeling turns out to be very effective in reconstructing spectra at low SNRs. In particular, HMMs show the highest capability of accounting for time correlations and, therefore, achieve the best results.**

***Index Terms*—Robust speech recognition, spectral reconstruction, missing-feature, minimum mean square error estimation.**

I. Introduction

The performance of automatic speech recognition (ASR) systems degrades rapidly when they operate under conditions that differ from those used for training. One source of the mismatch that still remains as a major issue among the ASR research community is additive noise [1]–[3]. Challenging acoustic environments such as those related with ASR in mobile devices can heavily deteriorate the performance of these systems. Therefore, accomplishing noise robustness is a key issue to make these systems deployable in real world conditions.

Among various approaches to robust ASR, one recent approach that has proved to be effective in noise compensation is based on the missing-data (or missing-feature) theory [4], [5]. This approach has been motivated by studies which show that human listeners can take advantage of the redundancy in speech signals to perform decoding even if large portions of speech spectra are masked by noise [6]–[8]. Missing-data techniques consider that, when speech is corrupted by additive noise, some regions of the noisy spectra are less affected by noise than others (e.g. formant frequencies are typically quite resilient to noise). In the log-spectral domain, while some regions of the spectra are completely masked by the noise energy and are considered unreliable or missing, other parts where the speech energy dominates are considered reliable. The reliable parts can be directly used for speech recognition since they constitute a good approximation of the unknown clean spectra [9], but the unreliable parts need a special treatment to perform recognition with them.

The classification of noisy spectra according to its degree of reliability can be either deterministic or probabilistic. Deterministic classification results in a binary missing-data mask that distinguishes between completely unreliable features and reliable ones. To obtain such a mask, the SNR of each feature is usually computed and compared with a fixed threshold. Given that SNR estimation is not always accurate in real word conditions, the probabilistic classification provides instead a soft mask with values within the interval $[0, 1]$. These soft values can be either obtained by applying a sigmoid compression to the estimated SNR level for every feature [10]–[12] or by means of a statistical classifier [13]–[16]. An in-depth discussion about mask estimation can be found in [17].

Two different methods have been considered within the missing-data framework to perform speech recognition with incomplete data: *marginalization* [4], [18], [19] and *imputation* [5], [20]–[27]. In marginalization, speech decoding relies on the reliable parts of noisy spectra, while the unreliable parts are discarded or marginalized up to the observed values. The imputation method involves the estimation of the unreliable features, so that decoding can be performed as usual. To estimate the missing features, the redundancy in speech signals is exploited by using statistical models that capture the correlation among features.

Although marginalization performs optimal decoding with missing features, it suffers from two main drawbacks [5], [27]. First, the standard decoding algorithm must be modified to account for missing features. Second, recognition has to

This work has been supported by an FPU grant from the Spanish Ministry of Education and by project MICINN-FEDER TEC2010-18009.

J. A. González, A. M. Peinado, and A. M. Gómez are with the Department of Teoría de la Señal Telemática y Comunicaciones, Universidad de Granada, Granada, 18071, Spain (e-mail: joseangl@ugr.es; amp@ugr.es; amgg@ugr.es).

N. Ma is with the MRC Institute of Hearing Research, Nottingham NG7 2RD, U.K. (e-mail: n.ma@ihr.mrc.ac.uk).

J. Barker is with the Department of Computer Science, University of Sheffield, Sheffield S1 4DP, U.K. (e-mail: j.barker@dcs.shef.ac.uk).

be carried out with spectral features. However, it is well known that cepstral features outperform spectral ones for speech recognition [5]. Moreover, since spectral features are used, the acoustic model (hidden Markov models) needs to employ Gaussian mixtures with full covariance matrices or an increased number of Gaussians with diagonal covariance. This could be computationally prohibitive and might need more training data in order to robustly estimate parameters in some cases (e.g. in large vocabulary continuous speech recognition systems).

The performance of missing-data reconstruction is poor in low SNR conditions where few, if any, reliable features are present [22], [24]. In order to improve spectral reconstruction in these conditions, *temporal redundancy* of speech signals has been exploited which is complementary to the information expressed in the frequency domain [22], [28]–[31]. In [5], the sequence of feature vectors is assumed to be a Gaussian wide-sense stationary process. Then, missing features are imputed using their known correlation with the reliable features of neighboring frames. However, it is shown that this approach is inferior to an alternative imputation technique only exploiting across-frequency correlation [5]. In [31], an HMM-based approach combining noise estimation and time-frequency modeling is proposed. Alternatively, Kim and Hansen [22] propose an approach in which the temporal trajectories in every filterbank channel are independently modeled. The final reconstruction for the missing features is then obtained as a combination of the original frequency-based estimation and the time-based one.

This paper presents a novel spectral imputation framework that provides full-band reconstructed spectra based on truncated Gaussian distributions. To do so, a GMM is used to represent clean speech and a minimum mean square error (MMSE) criterion is adopted to obtain suitable estimates of the unreliable features. The imputation framework is also extended to employ both frequency and temporal information using two novel approaches. First, a *patch-based imputation* using a sliding-window approach to model spectro-temporal patches of speech is presented. In this way, frequency and short-term temporal correlation of speech are jointly modeled and used for spectral reconstruction, without requiring any *ad hoc* fusion function to combine both estimates. The second technique uses HMMs different from those employed by the recognizer. Unlike the work in [31], our proposal does not directly require noise estimates, but mask estimates. Although good noise estimates can be helpful for speech reconstruction, as shown in [31], they can be difficult to obtain, specially for non-stationary noises. In the missing-data approaches which only rely on masks, noise estimates can be also required for mask estimation, although other speech features resistant to noise distortion (e.g. pitch, spectral flatness) may be also employed to compute such masks.

This paper is organized as follows. In the next section we present the general framework for missing-feature spectral reconstruction. Section III is devoted to temporal correlation exploitation within the reconstruction framework. First, an study of the limitations of the proposed reconstruction is carried out. Then, the two aforementioned techniques exploiting time-frequency correlations are presented. Section IV describes the experimental framework that has been adopted to evaluate the proposed techniques, whereas the speech recognition results are reported in Section V. Finally, Section VI concludes this paper and discusses some future directions.

II. MMSE- GMM-BASED IMPUTATION

Let us consider the feature vectors $\boldsymbol{y}$, $\boldsymbol{x}$, and $\boldsymbol{n}$ corresponding to log-filterbank energies (e.g. log-Mel) for noisy speech, clean speech, and additive noise, respectively. The model that relates these vectors is given by [32],

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

where the log and exponentiation operations are applied element-wise, and $\boldsymbol{r}$ is a small residual vector that depends on the phase relationship between speech and noise. Making the usual assumption that $\boldsymbol{r}$ is negligible compared to the other terms in (1) [33]–[35], the above model can be simplified to

$$\boldsymbol{y} = \log(e^{\boldsymbol{x}} + e^{\boldsymbol{n}}) \approx \max(\boldsymbol{x}, \boldsymbol{n}) \tag{2}$$

where $\max(\cdot)$ operates element-wise, and the *log-max* approximation has been considered to further simplify the model (i.e. $\log(e^x + e^n) \approx \max(x, n)$) [9], [36]–[38].

Eq. (2) is known as the *noise masking model* in the missing-data approach. According to this model, after noise distortion some spectro-temporal energies of the original clean speech are masked, while others remain almost uncorrupted. Thus, $\boldsymbol{y}$ can be rearranged into $\boldsymbol{y} \equiv (\boldsymbol{y}_r, \boldsymbol{y}_u)$, where $\boldsymbol{y}_r \approx \boldsymbol{x}_r$ denote the reliable features where the speech energy dominates, and $\boldsymbol{y}_u$ ($-\infty \leq \boldsymbol{x}_u \leq \boldsymbol{y}_u$) are the unreliable noisy observations in which speech is masked by noise.

Given the noisy observation $\boldsymbol{y}$, missing-data imputation techniques estimate the underlying clean speech $\boldsymbol{x}_u$ corresponding to $\boldsymbol{y}_u$, provided that $\boldsymbol{y}_r$ is a good approximation of $\boldsymbol{x}_r$. To do so, these techniques usually exploit the correlation between different features represented in the form of joint statistical models. In the following subsection, the framework for missing-data imputation will be presented.

A. General framework

Our starting point will be the MMSE estimator for the unreliable features given the noise masking model in (2):

$$\hat{\boldsymbol{x}}_u = E\left[\boldsymbol{x}_u | \boldsymbol{x}_r, \boldsymbol{x}_u \leq \boldsymbol{y}_u\right] = \int_{-\infty}^{\boldsymbol{y}_u} \boldsymbol{x}_u p\left(\boldsymbol{x}_u | \boldsymbol{x}_r, \boldsymbol{y}_u\right) d\boldsymbol{x}_u \tag{3}$$

where a simplified notation is used to denote the multidimensional integral within $(-\infty, y_{u,1}] \times (-\infty, y_{u,2}] \times \ldots$

In order to obtain the posterior distribution in (3), we assume that the clean speech feature distribution can be modeled using a GMM as,

$$p(\boldsymbol{x}) = \sum_{k=1}^{M} P(k) \mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k\right) \tag{4}$$

where $\boldsymbol{\mu}^k$, $\boldsymbol{\Sigma}^k$ and $P(k)$ are the mean vector, covariance matrix and *a priori* probability of the kth component of the GMM, respectively. In this case, the distribution of the

unreliable features in (3) is obtained by marginalizing over the Gaussian components as follows,

$$p\left(\boldsymbol{x}_u|\boldsymbol{x}_r,\boldsymbol{y}_u\right)=\sum_{k=1}^{M}p\left(\boldsymbol{x}_u,k|\boldsymbol{x}_r,\boldsymbol{y}_u\right) \tag{5}$$

Applying Bayes' rule, $p\left(\boldsymbol{x}_u,k|\boldsymbol{x}_r,\boldsymbol{y}_u\right)$ can be expressed as,

$$p\left(\boldsymbol{x}_u,k|\boldsymbol{x}_r,\boldsymbol{y}_u\right)=p\left(\boldsymbol{x}_u|\boldsymbol{x}_r,\boldsymbol{y}_u,k\right)P\left(k|\boldsymbol{x}_r,\boldsymbol{y}_u\right) \tag{6}$$

Then, applying (5) and (6) to (3), the resulting MMSE estimate is given by

$$\hat{\boldsymbol{x}}_u=\sum_{k=1}^{M}P\left(k|\boldsymbol{x}_r,\boldsymbol{y}_u\right)\underbrace{\int_{-\infty}^{\boldsymbol{y}_u}\boldsymbol{x}_u p\left(\boldsymbol{x}_u|\boldsymbol{x}_r,\boldsymbol{y}_u,k\right)d\boldsymbol{x}_u}_{\hat{\boldsymbol{x}}_u^k} \tag{7}$$

As can be seen, the MMSE estimate requires the computation of the posterior $P(k|\boldsymbol{x}_r,\boldsymbol{y}_u)$ and the partial estimate $\hat{\boldsymbol{x}}_u^k\equiv E[\boldsymbol{x}_u|\boldsymbol{x}_r,\boldsymbol{x}_u\leq\boldsymbol{y}_u,k]$ for every Gaussian component k. Using Bayes' rule again, the posterior probability can be expressed as:

$$P\left(k|\boldsymbol{x}_r,\boldsymbol{y}_u\right)=\frac{p\left(\boldsymbol{x}_r,\boldsymbol{y}_u|k\right)P(k)}{\sum_{k'=1}^{M}p\left(\boldsymbol{x}_r,\boldsymbol{y}_u|k'\right)P(k')} \tag{8}$$

where the joint observation probability $p\left(\boldsymbol{x}_r,\boldsymbol{y}_u|k\right)$ can be computed by marginalizing over the unreliable features $\boldsymbol{x}_u$ up to the observed values $\boldsymbol{y}_u$,

$$\begin{aligned}p\left(\boldsymbol{x}_r,\boldsymbol{y}_u|k\right)&=\int_{-\infty}^{\boldsymbol{y}_u}p\left(\boldsymbol{x}_r,\boldsymbol{x}_u|k\right)d\boldsymbol{x}_u\\&=p(\boldsymbol{x}_r|k)\int_{-\infty}^{\boldsymbol{y}_u}p(\boldsymbol{x}_u|\boldsymbol{x}_r,k)d\boldsymbol{x}_u\end{aligned} \tag{9}$$

For later use, we define here the mean square error (MSE) associated with the derived estimator. This term will be employed to compare different estimates according to their predicted accuracy. The MSE is defined as the trace of the estimator covariance matrix. For the estimator in (3), this matrix is:

$$\Sigma_{\hat{\boldsymbol{x}}_u}=E\left[(\boldsymbol{x}_u-\hat{\boldsymbol{x}}_u)(\boldsymbol{x}_u-\hat{\boldsymbol{x}}_u)^T|\boldsymbol{x}_r,\boldsymbol{x}_u\leq\boldsymbol{y}_u\right] \tag{10}$$

Assuming again that clean speech can be well modeled by a GMM, the covariance corresponding to (7) can be computed as,

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_u}=\sum_{k=1}^{M}P\left(k|\boldsymbol{x}_r,\boldsymbol{y}_u\right)\left(\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_u^k}+(\hat{\boldsymbol{x}}_u^k-\hat{\boldsymbol{x}}_u)(\hat{\boldsymbol{x}}_u^k-\hat{\boldsymbol{x}}_u)^T\right) \tag{11}$$

where $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_u^k}$ is the covariance corresponding to $\hat{\boldsymbol{x}}_u^k$, that is,

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_u^k}=E\left[(\boldsymbol{x}_u-\hat{\boldsymbol{x}}_u^k)(\boldsymbol{x}_u-\hat{\boldsymbol{x}}_u^k)^T|\boldsymbol{x}_r,\boldsymbol{x}_u\leq\boldsymbol{y}_u,k\right] \tag{12}$$

The integrals in (7) and (9) cannot be solved analytically for Gaussian probability density functions (pdfs) with non-diagonal covariance matrices. Therefore, some approximations are needed to make these integrals tractable. Next section will show how the proposed reconstruction technique can be derived from the above framework by making different assumptions. Then, other approaches that can be found in the literature will also be presented and compared with the proposed approach.

B. Truncated-Gaussian based imputation

The imputation framework described above makes two assumptions: (i) a GMM is used to model the clean speech statistics and (ii) the MMSE criterion is adopted to derive the estimator. The resulting estimator in (7) can be seen as a linear combination of partial estimates $\hat{\boldsymbol{x}}_u^k$ weighted by their corresponding posterior probabilities $P\left(k|\boldsymbol{x}_r,\boldsymbol{y}_u\right)$ $(k=1,\ldots,M)$. During the computation of these terms, integrals with no closed-form solution appear for full-covariance GMMs. We will see along this section how this problem can be partially avoided to finally obtain the reconstructed speech. For the derivation of our proposal, which will be referred to as truncated-Gaussian based imputation (TGI), the truncated Gaussian pdf will be extensively used. A brief overview of this probability distribution can be found in Appendix A.

First consider the computation of $P\left(k|\boldsymbol{x}_r,\boldsymbol{y}_u\right)$ in (7). As shown in (8) and (9), this posterior depends on $p(\boldsymbol{x}_r|k)$ and $p(\boldsymbol{x}_u|\boldsymbol{x}_r,k)$. These last two pdfs can be shown to be Gaussian distributed: the marginal distribution $p(\boldsymbol{x}_r|k)=\mathcal{N}(\boldsymbol{x}_r;\boldsymbol{\mu}_r^k,\boldsymbol{\Sigma}_{rr}^k)$ is obtained from the original pdf $\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}^k,\boldsymbol{\Sigma}^k)$ by partitioning $\boldsymbol{\mu}^k$ and $\boldsymbol{\Sigma}^k$ according to the the missing-data mask as follows,

$$\boldsymbol{\mu}^k=\begin{pmatrix}\boldsymbol{\mu}_r^k\\\boldsymbol{\mu}_u^k\end{pmatrix} \tag{13}$$

$$\boldsymbol{\Sigma}^k=\begin{pmatrix}\boldsymbol{\Sigma}_{rr}^k&\boldsymbol{\Sigma}_{ru}^k\\\boldsymbol{\Sigma}_{ur}^k&\boldsymbol{\Sigma}_{uu}^k\end{pmatrix}, \tag{14}$$

and the parameters of the conditional pdf $p(\boldsymbol{x}_u|\boldsymbol{x}_r,k)=\mathcal{N}(\boldsymbol{x}_u;\boldsymbol{\mu}_{u|r}^k,\boldsymbol{\Sigma}_{u|r}^k)$ are given by,

$$\boldsymbol{\mu}_{u|r}^k=\boldsymbol{\mu}_u^k+\boldsymbol{\Sigma}_{ur}^k\left(\boldsymbol{\Sigma}_{rr}^k\right)^{-1}\left(\boldsymbol{y}_r-\boldsymbol{\mu}_r^k\right) \tag{15}$$

$$\boldsymbol{\Sigma}_{u|r}^k=\boldsymbol{\Sigma}_{uu}^k-\boldsymbol{\Sigma}_{ur}^k\left(\boldsymbol{\Sigma}_{rr}^k\right)^{-1}\boldsymbol{\Sigma}_{ru}^k \tag{16}$$

As can be seen, a full covariance is obtained for $p(\boldsymbol{x}_u|\boldsymbol{x}_r,k)$. In order to make the integral in (9) tractable, only the diagonal elements of $\boldsymbol{\Sigma}_{u|r}^k$ are retained in this work. With this approximation, a closed-form solution can be obtained for the multivariate integral, whereas some correlation between features can still be exploited. On the other hand, the accuracy of the approximation will depend on the degree of non-diagonality of the covariance matrix $\boldsymbol{\Sigma}_{u|r}^k$. Applying this approximation, $p\left(\boldsymbol{x}_r,\boldsymbol{y}_u|k\right)$ in (9) can now be evaluated as,

$$\begin{aligned}p\left(\boldsymbol{x}_r,\boldsymbol{y}_u|k\right)&\approx p(\boldsymbol{x}_r|k)\prod_{l\in\boldsymbol{s}_u}\int_{-\infty}^{y_{u,l}}\mathcal{N}\left(x_u;\mu_{u|r,l}^k,\sigma_{u|r,l}^k\right)dx_u\\&=\mathcal{N}\left(\boldsymbol{x}_r;\boldsymbol{\mu}_r^k,\boldsymbol{\Sigma}_{rr}^k\right)\prod_{l\in\boldsymbol{s}_u}\Phi\left(\overline{y}_{u,l}^k\right)\end{aligned} \tag{17}$$

where $\boldsymbol{s}_u$ is the set with the indexes of the unreliable features, $\Phi(\cdot)$ is the cumulative distribution function (CDF) for the standard Gaussian distribution, i.e.

$$\Phi(x)=\int_{-\infty}^{x}\mathcal{N}(u)du, \tag{18}$$

and $\overline{y}_{u,l}^k$ is the standardized value for $y_{u,l}$ regarding the kth Gaussian:

$$\overline{y}_{u,l}^k=\frac{y_{u,l}-\mu_{u|r,l}^k}{\sigma_{u|r,l}^k} \tag{19}$$

Next, consider the computation of $\hat{\boldsymbol{x}}_u^k$ in (7). Again, the integral has a closed-form only for Gaussian pdfs with diagonal covariance matrices. Thus, assuming independence among unreliable features given the reliable ones and approximating $p\left(\boldsymbol{x}_u|\boldsymbol{x}_r,\boldsymbol{y}_u,k\right)$ by $p\left(\boldsymbol{x}_u|\boldsymbol{x}_r,k\right)$, $\hat{x}_{u,l}^k$ $(l \in \boldsymbol{s}_u)$ corresponds to the first moment of a right-truncated Gaussian distribution defined in the interval $(-\infty, y_{u,l}]$. This value can be computed as (see eqn. (51)),

$$\hat{x}_{u,l}^k = \mu_{u|r,l}^k - \sigma_{u|r,l}^k \frac{\mathcal{N}\left(\overline{y}_{u,l}^k\right)}{\Phi\left(\overline{y}_{u,l}^k\right)} \tag{20}$$

Finally, the covariance of the proposed estimator is obtained according to (11), being the lth element of the diagonal matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_u^k}$ given by (see eqn. (52)),

$$\left(\tilde{\sigma}_{u|r,l}^k\right)^2 = \left(\sigma_{u|r,l}^k\right)^2 \left[1 - \frac{\mathcal{N}(\overline{y}_{u,l}^k)}{\Phi(\overline{y}_{u,l}^k)} \left(\overline{y}_{u,l}^k + \frac{\mathcal{N}(\overline{y}_{u,l}^k)}{\Phi(\overline{y}_{u,l}^k)}\right)\right] \tag{21}$$

C. Comparison with related techniques

Many of the techniques proposed in the literature for missing-feature imputation can be inferred from the above framework by adopting different assumptions. In this section, their relationship with the proposed technique TGI is discussed.

One of the first techniques devoted to imputation of partial speech spectra was Raj et al.'s cluster-based reconstruction (CBR) technique [5]. Again, clean speech is modeled by a GMM. However, CBR assumes diagonal covariances for the Gaussian pdfs. In this case, the observation probability $p(\boldsymbol{x}_r,\boldsymbol{y}_u|k)$ defined in (9), which is required to compute the posterior $P(k|\boldsymbol{x}_r,\boldsymbol{y}_u)$ of the general estimation in (7), can be computed as,

$$\begin{aligned} p\left(\boldsymbol{x}_r,\boldsymbol{y}_u|k\right) &= \prod_{d\in\boldsymbol{s}_r} p(x_{r,d}|k,d) \prod_{l\in\boldsymbol{s}_u} \int_{-\infty}^{y_{u,l}} p(x_u|k,l)dx_u \\ &= \prod_{d\in\boldsymbol{s}_r} \mathcal{N}\left(x_{r,d};\mu_{r,d}^k,\sigma_{r,d}^k\right) \prod_{l\in\boldsymbol{s}_u} \Phi\left(\frac{y_{u,l}-\mu_{u,l}^k}{\sigma_{u,l}^k}\right) \end{aligned} \tag{22}$$

with $\boldsymbol{s}_r$ being the set with the frequency indexes corresponding to the reliable features. The parameters of the marginal distributions $\mathcal{N}(x_{r,d};\mu_{r,d}^k,\sigma_{r,d}^k)$ and $\mathcal{N}(x_{u,l};\mu_{u,l}^k,\sigma_{u,l}^k)$ are obtained from (13) and (14) according to the information provided by the missing-data mask.

The other factor required by the estimator in (7) is the Gaussian-conditional estimate $\hat{\boldsymbol{x}}_u^k$. To compute this term, an iterative procedure called *bounded MAP estimation* is proposed for CBR in [5]. This procedure, however, can be computationally expensive, so in practice it is simplified. In this simplification, the upper bound limit of the integral in (7) is discarded, so that, assuming independence among features, this integral reduces to the conditional mean of the unreliable features, that is,

$$\hat{\boldsymbol{x}}_u^k = \int \boldsymbol{x}_u p\left(\boldsymbol{x}_u|\boldsymbol{x}_r,k\right) d\boldsymbol{x}_u = \boldsymbol{\mu}_{u|r}^k \tag{23}$$

Finally, the partial estimate computed in (23) is usually postprocessed to fulfill the model in (2), i.e. the observed values $\boldsymbol{y}_u$ are an upper bound for $\boldsymbol{x}_u$. Hence, the final estimation is given by,

$$\hat{\boldsymbol{x}}_u^k = \min\left(\boldsymbol{y}_u, \boldsymbol{\mu}_{u|r}^k\right) \tag{24}$$

where $\min(\cdot)$ operates element-wise.

The following differences between CBR and TGI can be observed. First, CBR assumes independence among features to obtain closed-form solutions for the multivariate integrals that appear during the estimation. Consequently, this assumption may cause this technique to be unable to fully exploit the correlation among features. Second, the noise masking model in (2) involves that $\boldsymbol{x}_u$ is upper-bounded by $\boldsymbol{y}_u$. As we have seen, CBR postprocesses its estimates to acommplish this restriction, whereas this restriction is fully embedded in TGI by using truncated Gaussian distributions.

Recently, an alternative technique for missing-feature reconstruction was proposed in [26]. This technique is similar to TGI except in the way the multivariate integrals in (7) and (9) are computed. In [26], a linear transformation is applied to diagonalize the covariance matrix of $p(\boldsymbol{x}_u|\boldsymbol{x}_r,k)$. After the transformation, the multivariate integral reduces to univariate integrals in the transformed domain, so that it can be easily computed. Let $\boldsymbol{H}$ be the upper triangular matrix obtained after the Cholesky factorization of the precision matrix of $p(\boldsymbol{x}_u|\boldsymbol{x}_r,k)$, that is,

$$\left(\boldsymbol{\Sigma}_{u|r}^k\right)^{-1} = \boldsymbol{H}\boldsymbol{H}^T \tag{25}$$

Then, the transformation used for diagonalization of the covariance matrix is $\tilde{\boldsymbol{x}}_u = \boldsymbol{H}(\boldsymbol{x}_u - \boldsymbol{\mu}_{u|r}^k)$. After this transformation, $p(\boldsymbol{x}_u|\boldsymbol{x}_r,k)$ in (9) is shown to be given by (see [26] for more details),

$$p\left(\boldsymbol{x}_r,\boldsymbol{y}_u|k\right) \approx \mathcal{N}\left(\boldsymbol{x}_r;\boldsymbol{\mu}_r^k,\boldsymbol{\Sigma}_r^k\right) \prod_{l\in\boldsymbol{s}_u} \Phi\left(\tilde{y}_{u,l}^k\right) \tag{26}$$

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

Similarly, after the aforementioned transformation, $\hat{\boldsymbol{x}}_u^k$ in (7) is approximated in [26] as,

$$\hat{\boldsymbol{x}}_u^k \approx \boldsymbol{\mu}_{u|r}^k - \boldsymbol{H}^{-1}\left[\frac{\mathcal{N}\left(\tilde{y}_{u,1}^k\right)}{\Phi\left(\tilde{y}_{u,1}^k\right)}, \frac{\mathcal{N}\left(\tilde{y}_{u,2}^k\right)}{\Phi\left(\tilde{y}_{u,2}^k\right)}, \cdots\right]^T \tag{27}$$

It can be seen that TGI discards the non-diagonal elements of the covariance matrix, whereas the work in [26] diagonalizes this matrix. However, the integration limits are also modified by the transformation used to diagonalize the matrix, and the basic assumption introduced by the noise masking model in (2), i.e. $-\infty \leq \boldsymbol{x}_u \leq \boldsymbol{y}_u$, does not necessarily hold.

Other approaches to compute the multivariate Gaussian integral have been also employed in the missing-feature framework. For example, Monte-Carlo integration [39] or the grid-based sampling technique proposed in [40] can be used to this end. Although appealing, these numerical approaches become computationally prohibitive as the dimension of $\boldsymbol{y}_u$ increases. Another simple, yet effective, solution to this problem is also proposed in [40]. In this work, the integral of the multivariate Gaussian is approximated by its integrand's maximum within the integration region. It is shown that, when compared with

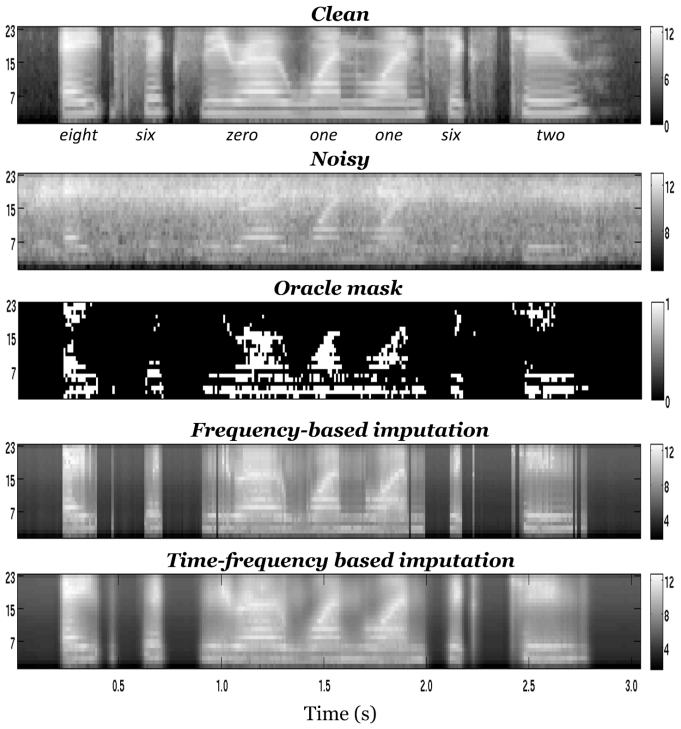

Fig. 1. From top to bottom: log-Mel spectrogram for the clean utterance *eight six zero one one six two* of the Aurora-2 database [42], spectrogram for the utterance corrupted by subway noise at 0dB, oracle mask (white means reliable and black unreliable), reconstructed spectrogram obtained by the TGI technique of Section II-B, and the estimation obtained by the patch-based imputation technique proposed in Section III-A exploiting temporal and frequency correlations.

numerical integration, this second approach yield similar performance and is faster.

Finally, a technique for spectral reconstruction based on an occlusion model (SRO) has recently been proposed in [41]. As in the proposed missing-feature reconstruction framework, SRO is based on MMSE estimation of noise-corrupted log-spectral features. Furthermore, the noise occlusion model in (2) and a GMM for clean speech are also assumed to derive SRO. However, SRO uses noise estimates to compensate the speech features instead of using missing-data masks.

III. Exploiting temporal correlation

The imputation techniques presented in Section II only make use of the correlation across frequency and no temporal information is exploited. However, it is well known that temporal and frequency information are complementary, so that better performance can be expected if both are jointly used [22], [28]–[31].

To better understand how temporal correlations could improve imputation, Fig. 1 shows examples of log-Mel spectrograms for clean and noisy utterances of the Aurora-2 database [42], together with the oracle mask and two different reconstructions obtained for the noisy utterance. Subway noise at 0 dB SNR has been added to obtain the noisy utterance. The oracle mask is obtained by direct comparison between the clean and noisy utterances using a threshold of 7 dB SNR. Two reconstructions are shown: frequency-based and time-frequency based imputations. The first one is obtained by the TGI technique of Section II-B. To obtain the second one, both time and frequency correlations are exploited by the patch-based imputation technique presented in Section III-A. By comparing the clean and frequency-based reconstructed spectra, the three following issues can be observed.

First, the frames with no reliable features are reconstructed as silence. We can understand this behavior by looking into the general estimation equation (7). If all the features are unreliable, then the partial estimate $\hat{\boldsymbol{x}}_u^k$ corresponds to mean of the kth Gaussian, i.e.,

$$\hat{\boldsymbol{x}}_u^k = \mu^k \tag{28}$$

For the posterior $P(k|\boldsymbol{x}_r, \boldsymbol{y}_u)$ in (7), this term becomes $P(k|\boldsymbol{y}_u)$ when $\boldsymbol{x}_r$ is empty. Applying Bayes' rule, $P(k|\boldsymbol{y}_u)$ is proportional to $P(k)p(\boldsymbol{y}_u|k)$ and, finally, $p(\boldsymbol{y}_u|k)$ can be obtained by marginalizing the distribution $p(\boldsymbol{x}_u|k)$ up to $\boldsymbol{y}_u$:

$$p\left(\boldsymbol{y}_u|k\right) = \int_{-\infty}^{\boldsymbol{y}_u} p\left(\boldsymbol{x}_u|k\right) d\boldsymbol{x}_u \approx \prod_{l \in \boldsymbol{s}_u} \Phi(\overline{y}_{u,l}^k) \tag{29}$$

where independence among the unreliable features is assumed. Now it is clear why the completely unreliable frames are reconstructed as silence frames. As the Gaussian components of the GMM representing silence usually have low-energy mean vectors and small variances, the CDF computation in (29) will approach one for these components. For the components representing speech, the cumulative probability will be smaller, since these components have higher energy mean-vectors and broader variances. Hence, the silence components will have higher posterior probabilities than speech components. Note that this behavior can also be predicted from the log-max approximation of (2): this model assumes that, for unreliable features, noise energy is much higher than speech energy.

Second, a noticeable vertical banding effect is present in the reconstructed spectrogram indicating discontinuities between neighboring frames (e.g. last spoken digit around time 2.5 s). These discontinuities are known to be harmful for ASR, as they yield substitution and insertion errors during the recognition. The vertical banding is due to two different reasons. First, frames are processed independently and no temporal constraints are imposed for the reconstruction. Hence, a non-smooth reconstruction could be obtained for low SNR conditions. Second, as previously mentioned, totally unreliable frames are reconstructed as silence frames. This generates discontinuities when totally unreliable frames are surrounded by frames with reliable features (e.g. time instants 0.98, 1.92 and 2.72 s), or when frames with reliable features are surrounded by totally unreliable frames (e.g. time instants 0.47 and 2.23 s).

Third, some parts of the speech spectra are completely lost due to noise distortion (e.g. high frequencies for the /s/ phonemes at 0.57 s and 2.07 s). This information is irrecoverable by the imputation techniques presented above, unless temporal constraints are used.

The temporal modeling of the speech can improve the missing-feature reconstruction accuracy. As can be seen in the

bottom plot of Fig. 1, the exploitation of temporal and frequency correlation helps to ameliorate some of the aforementioned problems that occur when only frequency correlations are accounted for. This paper proposes two novel approaches to jointly consider both correlations within the missing-data reconstruction framework: (i) temporal modeling of speech patches and (ii) HMM-based modeling of speech. These approaches will be discussed in the following subsections.

A. Patch-based modeling of short-term temporal correlations

A novel approach is proposed here to exploit the short-term temporal correlation of speech. By using a sliding window, time-frequency patches of speech are first modeled and then used for reconstruction. Let us define the time-frequency patch for time instant t as a super-vector $\boldsymbol{z}(t)$ containing τ consecutive feature vectors as follows,

$$\boldsymbol{z}(t) = \left[\boldsymbol{x}(\delta\cdot(t-1)+1)^T, \ldots, \boldsymbol{x}(\delta\cdot(t-1)+\tau)^T\right]^T \quad (30)$$

where τ is the window length and δ is the frame displacement. For example, if $\tau = 3$ and $\delta = 2$, then $\boldsymbol{z}(1) = [\boldsymbol{x}(1)^T, \boldsymbol{x}(2)^T, \boldsymbol{x}(3)^T]^T$, $\boldsymbol{z}(2) = [\boldsymbol{x}(3)^T, \boldsymbol{x}(4)^T, \boldsymbol{x}(5)^T]^T$, etc. As can be noted, consecutive patches $\boldsymbol{z}(t)$ and $\boldsymbol{z}(t+1)$ are overlapped $\tau - \delta$ frames.

Two elements are required to apply the imputation techniques to $\boldsymbol{z}(t)$: a missing data mask in the augmented feature space, and a model representing the augmented features. The missing data mask is obtained by applying the sliding window approach (30) to the original missing data mask $\boldsymbol{m}$. A GMM can be trained for the new features $\boldsymbol{z}$. However, the parameters of this GMM will be coarsely estimated given the high dimensionality of $\boldsymbol{z}$ (the size of $\boldsymbol{z}$ is $\tau \times n$, where n is the number of filterbank channels, typically 23). Hence, a dimensionality reduction technique is needed to robustly train the GMM. In this work, we use principal component analysis (PCA) [43].

Let $\boldsymbol{P}$ be the matrix containing the n^* principal components (eigenvectors) of the covariance $\boldsymbol{\Sigma}_z$ in the augmented space. To achieve dimensionality reduction, PCA projects every feature vector $\boldsymbol{z}(t)$ onto the orthogonal space defined by $\boldsymbol{P}$. Prior to the projection, PCA requires that the mean of the training data $\boldsymbol{\mu}_z$ has been subtracted from $\boldsymbol{z}(t)$. Thus, the new feature vector $\boldsymbol{z}'(t)$ in the PCA domain is defined as,

$$\boldsymbol{z}'(t) = \boldsymbol{P}^T\left(\boldsymbol{z}(t) - \boldsymbol{\mu}_z\right) \quad (31)$$

Once PCA is applied to the training data, a GMM can be trained to represent the projected features in the PCA domain. As a result of applying the orthogonal projection defined by PCA, the features in the projected space are decorrelated, so that they can be better modeled with a diagonal covariance GMM. However, the uncertainty of the unreliable features in the log-filterbank domain is spread over all the dimensions of the PCA domain. Thus, no reliable/unreliable decision can be taken in this domain. To solve this problem, we adopt a similar solution to that proposed by [21], in which the parameters of the GMM in the PCA domain are transformed back to the original space. In our case, the parameters of kth Gaussian

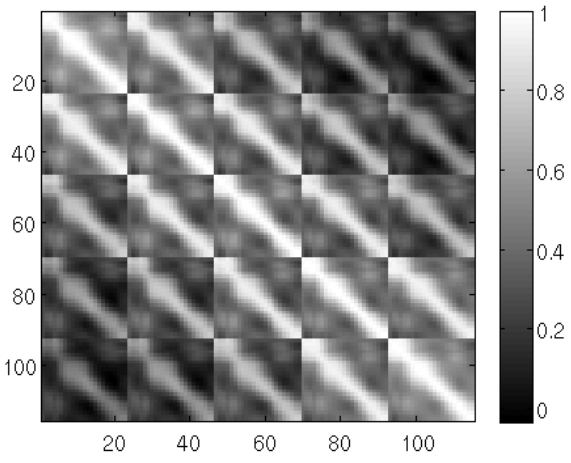

Fig. 2. Correlation matrix of a Gaussian component. Features are obtained applying an sliding window with $\tau = 5$ frames with a displacement $\delta = 1$.

$\mathcal{N}\left(\boldsymbol{z}'; \boldsymbol{\mu}^k_{z'}, \boldsymbol{\Sigma}^k_{z'}\right)$ are transformed to the log-filterbank domain as follows,

$$\boldsymbol{\mu}^k_z = \boldsymbol{P}\boldsymbol{\mu}^k_{z'} + \boldsymbol{\mu}_z \quad (32)$$

$$\boldsymbol{\Sigma}^k_z = \boldsymbol{P}\boldsymbol{\Sigma}^k_{z'}\boldsymbol{P}^T + \boldsymbol{\Sigma}_r \quad (33)$$

where $\boldsymbol{\Sigma}_r$ is the residual variance of the training data not considered in the PCA domain. This term, which is added to avoid rank-deficiency in the covariance matrix $\boldsymbol{\Sigma}^k_z$, is computed as

$$\boldsymbol{\Sigma}_r = \boldsymbol{R}\boldsymbol{R}^T\boldsymbol{\Sigma}_z\boldsymbol{R}\boldsymbol{R}^T \quad (34)$$

with $\boldsymbol{R}$ being the matrix containing the remaining eigenvectors not considered in $\boldsymbol{P}$, so that $[\boldsymbol{P}\boldsymbol{R}]$ is the eigenvector matrix of $\boldsymbol{\Sigma}_z$.

Fig. 2 shows an example of the correlation matrix for a given Gaussian component after the transformation in (33). The GMM is first trained with PCA features ($n = 23$, $\tau = 5$, and $\delta = 1$) and, then, transformed back to the log-Mel domain using (32) and (33). As can be seen, high correlation exists between close features, both in frequency or time. As expected, when the distance between features (in frequency or time) is increased, this correlation diminishes.

Given the frame-overlap between consecutive patches, several estimates for a particular unreliable feature x_u will be obtained (one for every patch containing that specific feature). We will denote these estimates as $\hat{x}^{(n)}_u$ $(n = 1, \ldots, N)$. These estimates are combined to obtain a final reconstruction as follows,

$$\hat{x}_u = \sum_{n=1}^{N} c_n \hat{x}^{(n)}_u \quad (35)$$

where c_n is the confidence assigned to the nth estimate ($\sum_{n=1}^{N} c_n = 1$). In this work, we consider c_n as being inversely proportional to the variance of $\hat{x}^{(n)}_u$. These variances are computed as described by equations (11) and (21). Then, the confidence will be heuristically obtained as,

$$c_n = \begin{cases} 1 & N = 1 \\ \frac{1-\epsilon_n}{N-1} & otherwise \end{cases} \quad (36)$$

where ϵ_n is the normalized error for the nth estimate. This value is computed from the variances $(\sigma_{\hat{x}_u}^{(n)})^2$ as follows,

$$\epsilon_n = \frac{\left(\sigma_{\hat{x}_u}^{(n)}\right)^2}{\sum_{n'=1}^{N}\left(\sigma_{\hat{x}_u}^{(n')}\right)^2} \tag{37}$$

Fig. 1 (bottom plot) shows an example of the spectral reconstruction obtained by the patch-based approach. As can be seen, some reconstruction errors appearing in the frequency-based imputation are effectively compensated when the temporal redundancy is exploited. Moreover, a more smooth reconstruction is obtained.

B. Temporal modeling using HMMs

HMMs are a well known method to model the sequential structure of speech [44], and have been shown as a powerful tool for MMSE-based feature reconstruction [28]–[31]. In fact, the acoustic HMM models of the recognition engine can be applied for this task. Works such as [4], [21] follow this approach which is referred to as *state-based imputation*. However, decoupling the feature compensation and recognition processes can be attractive for several reasons. First, the recognizer does not need to be modified to account for missing-feature inputs. Also, the feature compensation is independent of the recognizer architecture and the recognition task complexity (e.g., large vocabulary vs. digit recognition). For these reasons, we will use the second approach in this work.

An ergodic HMM with single-mixture state-conditional pdfs will be applied for spectral reconstruction. To obtain this HMM, the clean speech GMM that was previously used for imputation is augmented with transition probabilities, so that each HMM state corresponds to one component of the original GMM. It must be pointed out that a more sophisticated model could be applied (e.g. a mixture of Gaussians per state). For the sake of simplicity, this is not considered here. Let $\boldsymbol{A}$ be a matrix representing the transition probabilities between states (Gaussian components) of the HMM. The elements $a_{ij}(1 \leq i,j \leq M)$ of this matrix can be estimated as,

$$\begin{aligned} a_{ij} &= P\left(s_{t+1}=j|s_t=i\right) \\ &= \frac{\sum_{t=1}^{T} p\left(\boldsymbol{x}_{t+1}|s_{t+1}=j\right) p\left(\boldsymbol{x}_t|s_t=i\right) P(j)P(i)}{\sum_{t'=1}^{T} p\left(\boldsymbol{x}_{t'}|s_{t'}=i\right) P(i)} \end{aligned} \tag{38}$$

where i and j are states of the HMM, and $s_1, s_2, \ldots, s_T$ defines the state sequence along time.

In order to account for the temporal feature evolution, the estimation in (7) is rewritten as [29],

$$\hat{\boldsymbol{x}}_{u,t} = \sum_{k=1}^{M} P\left(s_t=k|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_T\right)\hat{\boldsymbol{x}}_{u,t}^{k} \tag{39}$$

where $P(s_t=k|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_T) \equiv \gamma_t(k)$, i.e. the posterior of the kth HMM state at time t given the whole sequence of feature vectors, is the only term accounting for temporal correlations. This posterior can be efficiently computed by means of the forward-backward algorithm [28], [44] as,

$$\gamma_t(k) = \frac{\alpha_t(k)\beta_t(k)}{\sum_{i=1}^{M}\alpha_t(i)\beta_t(i)} \tag{40}$$

where the forward and backward variables $\alpha_t(i)$ and $\beta_t(i)$, respectively, are defined as follows,

$$\alpha_t(i) = p\left(s_t=i|\boldsymbol{y}_1,\ldots,\boldsymbol{y}_t\right) \tag{41}$$

$$\beta_t(i) = p\left(\boldsymbol{y}_{t+1},\ldots,\boldsymbol{y}_T|s_t=i\right) \tag{42}$$

These probabilities can be computed recursively as,

$$\alpha_t(i) = \left[\sum_{j=1}^{M}\alpha_{t-1}(j)a_{ji}\right] p\left(\boldsymbol{y}_t|i\right) \tag{43}$$

$$\beta_t(i) = \sum_{j=1}^{M} a_{ij}\beta_{t+1}(j)p\left(\boldsymbol{y}_{t+1}|i\right) \tag{44}$$

where the observation probability $p(\boldsymbol{y}_t|i)$ is equivalent to the conditional probability (9) and it is computed by using (17) in TGI and (22) in CBR.

The forward and backward variables are initialized as,

$$\alpha_1(i) = P(i)p\left(\boldsymbol{y}_1|i\right) \tag{45}$$

$$\beta_T(i) = 1 \tag{46}$$

C. Computational complexity

In this section, a broad analysis of the computational complexity of the proposed reconstruction techniques will be presented. Generally speaking, the computational cost of these techniques mainly depends on the following terms: (i) the number of Gaussians in the GMM (or HMM) M, (ii) the number of reliable features r, and (iii) the number of unreliable features u. For the TGI approach of Section II-B, the computational cost is mainly dominated by the computation of the $(u \times r)$ regression matrix $\boldsymbol{\Sigma}_{ur}^{k}\left(\boldsymbol{\Sigma}_{rr}^{k}\right)^{-1}$ that appears in (15) and (16). A naive computation of this matrix takes $O(r^3 + r^2u)$ operations (multiplications and additions), from which the inversion of $\boldsymbol{\Sigma}_{rr}^{k}$ involves $O(r^3)$ operations and the multiplication of both matrices takes $O(r^2u)$. In addition, the computation of $p(\boldsymbol{x}_r, \boldsymbol{y}_u|k)$ defined in (17) is also required. For this probability, the evaluation of $\mathcal{N}(\boldsymbol{x}_r; \boldsymbol{\mu}_r^k, \boldsymbol{\Sigma}_{rr}^k)$ includes a quadratic form in $(\boldsymbol{\Sigma}_{rr}^k)^{-1}$, which approximately requires $O(r^3)$ operations. Moreover, the evaluation of the CDF $\Phi(\cdot)$ in (17) can be efficiently implemented through a look-up table. Here, we will assume that the cost of the search algorithm is negligible and, hence, the cost of evaluating the CDF for all the unreliable components takes $O(u)$ operations. In summary, the computational complexity of TGI is, roughly speaking, upper bounded by $O(Mr^3 + Mr^2u + Mu)$.

Similarly, the computational cost of the HMM-based approach is nearly the same as that of TGI. The main difference between both approaches is that the forward-backward probabilities $\gamma_t(k)$ in (40) are computed by the HMM-based approach. For an ergodic HMM, the order of complexity of the forward-backward algorithm is $O(M^2T)$.

Finally, as for TGI, the computational cost of the patch-based approach is dominated by the computation of $\mathbf{\Sigma}_{ur}^{k}\left(\mathbf{\Sigma}_{rr}^{k}\right)^{-1}$ and the evaluation of the Gaussian pdfs and CDFs. Note that in this case the covariance matrices represent the speech statistics in the augmented space, so that a high dimension is expected for this matrices. Nevertheless, some of the computations can be made offline since all parameters are known (e.g. (32) and (33)).

IV. Experimental framework

The proposed spectral reconstruction techniques were evaluated on the Aurora-2 [42] and Aurora-4 [45] databases. Aurora-2 is a small vocabulary recognition task consisting of utterances of English connected digits with artificially added noise. The clean training dataset comprises 8440 utterances with 55 male and 55 female speakers. Three test sets are defined: set A, B, and C. The utterances of every set are artificially contaminated by four types of additive noise (two types for set C) at seven SNR values (clean, 20, 15, 10, 5, 0, and -5 dB). The utterances of set C are also filtered using a different channel response. Aurora-4 is a large vocabulary database which is based on the Wall Street Journal (WSJ0) 5000-word recognition task. A total of 14 hours of speech data corresponding to 7138 utterances from 83 speakers are included in the clean training dataset. Fourteen different test sets are defined. The first seven sets, from T-01 to T-07, are generated by adding seven different noise types (clean condition, car, babble, restaurant, street, airport, and train) to 330 utterances from eight speakers. The SNR values considered range from 5 dB to 15 dB. The last seven sets are obtained in the same way, but the utterances are recorded with different microphones than the one used for recording the training set.

The European Telecommunication Standards Institute front-end (ETSI FE, ES 201 108) [46] is used to extract acoustic features from the speech signal. Twelve Mel-Frequency Cepstral Coefficients (MFCCs) along with the 0th order coefficient and their respective velocity and acceleration form the 39 dimensional feature vector used by the recognizer. For spectral reconstruction, 23-component feature vectors corresponding to the outputs of the log-Mel filterbank are used. After reconstruction, the discrete cosine transform (DCT) is applied to obtain the final cepstral parameters. Cepstral mean normalization (CMN) is applied in all cases to improve the robustness of the system to channel mismatches.

Acoustic models trained on clean speech are employed in each task. For Aurora-2, left to right continuous density HMMs with 16 states and 3 Gaussians per state are used to model each digit. Silences and short pauses are modeled by HMMs with 3 and 1 states, respectively, and 6 Gaussians per state. In Aurora-4, continuous cross-word triphone models with 3 tied states and a mixture of 6 Gaussians per state are used. The language model is the standard bigram for the WSJ0 task.

Spectral reconstruction is performed using a 256-component GMM (or HMM) with full covariance matrices. GMM training is performed by the expectation-maximization (EM) algorithm on the same dataset used for acoustic model training. Oracle and estimated masks are tested. We initially develop a set of oracle experiments in which a perfect knowledge of the SNR of the input signal is assumed. To do so, clean and noisy spectra are compared to obtain oracle missing-data masks. These experiments allow us to evaluate the potential of the proposed techniques under no mask estimation errors. Then, more realistic experiments are performed. In these experiments, missing-data masks are computed from noise estimates obtained through linear interpolation of initial noise statistics extracted from the beginning and final frames of every utterance. These experiments will provide us an estimation of the actual performance that can be expected from the reconstruction techniques when combined with estimated masks. Thus, the robustness to mask estimation errors can be evaluated. In both cases (oracle and estimated masks), a SNR threshold is applied to obtain the final binary masks.

For comparative purposes, the performance of the time-frequency missing-feature reconstruction (TF-MFR) technique proposed in [22] is also evaluated. In this technique, GMMs are used to model the temporal correlation for every filterbank channel. Then, frequency-based and temporal-based reconstructions are combined according to an heuristic rule in order to obtain the final estimate for the missing features. The parameters used in the implementation of this technique were those proposed in [22].

The optimal values for the parameters τ and δ of the sliding window in (30) used by the patch-based approach in Section III-B were empirically derived. To do so, development sets created for both databases (Aurora-2 and Aurora-4) were used. In both cases, the optimal values were found to be $\tau = 5$ and $\delta = 1$ when a 95 % of the variance is retained by PCA. Surprisingly, the value for the window length $\tau = 5$ frames (65 ms) approximately corresponds to the mean duration of a English phone [47].

V. Results

A. Experiments using oracle masks

Table I shows the performance obtained by the proposed spectral reconstruction techniques for the Aurora-2 database. The three test sets of this database are considered for computing the average results for each SNR. In addition, the average accuracy over all SNRs (Avg.) and the corresponding relative improvement (R.I.) regarding the baseline (MFCC features plus CMN) for every technique are also shown. For comparison, the recognition results obtained by the ETSI advanced front-end (AFE) [48], which is especially designed for noise robustness, have also been included as a second baseline.

Two reconstruction techniques using no temporal information are evaluated: the CBR approach of [5] and our proposal TGI presented in Section II-B. Given that TGI achieves better performance than CBR, the exploitation of the temporal correlation of speech to improve spectral reconstruction is only considered for TGI. Three different approaches to account for time dependency within spectral reconstruction are evaluated: the PATCH and HMM approaches proposed in Sections III-A and III-B, respectively, and the TF-MFR technique from [22].

		Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg.	R.I.%
	Baseline	99.11	97.29	92.55	75.56	42.82	22.69	12.92	63.28	-
	AFE	99.23	98.15	96.66	93.23	84.57	62.50	30.33	80.67	27.48
No temporal	CBR	99.11	98.90	98.52	97.41	94.78	88.05	71.48	92.61	46.35
information	TGI	99.11	99.01	98.75	97.99	96.11	90.90	77.34	94.17	48.83
Using	TF-MFR+TGI	99.11	99.01	98.84	98.29	97.00	93.67	84.65	95.79	51.39
temporal	PATCH+TGI	99.11	99.02	98.81	98.12	96.70	92.84	83.39	95.43	50.81
information	HMM+TGI	99.11	99.01	98.75	98.22	96.77	92.80	85.99	95.81	51.41

TABLE I
WORD ACCURACY RESULTS (%) ACHIEVED BY THE PROPOSED RECONSTRUCTION TECHNIQUES FOR THE AURORA-2 DATABASE WHEN ORACLE MASK ARE USED.

		T-01	T-02	T-03	T-04	T-05	T-06	T-07	T-08	T-09	T-10	T-11	T-12	T-13	T-14	Avg.	R.I.%
	Baseline	87.69	75.30	53.24	53.15	46.80	56.36	45.38	77.04	64.24	45.30	42.07	36.15	47.43	36.67	54.77	-
	AFE	88.25	81.41	69.14	64.80	67.44	66.34	68.78	80.57	74.76	61.89	56.47	58.75	60.13	59.87	68.47	25.01
No temporal	CBR	87.69	86.59	82.98	83.82	81.95	85.65	81.30	79.17	77.15	71.87	71.44	68.48	73.81	69.81	78.69	43.67
information	TGI	87.69	87.07	84.81	85.02	83.22	86.10	83.11	79.97	78.82	73.87	73.40	71.74	75.32	72.41	80.18	46.39
Using	TF-MFR+TGI	87.69	87.24	84.98	85.97	84.66	86.79	84.31	80.03	79.13	76.91	74.78	73.85	77.64	74.59	81.32	48.47
temporal	PATCH+TGI	87.69	86.87	85.02	84.83	84.12	86.51	84.35	80.48	79.86	76.70	74.99	73.94	77.81	75.19	81.31	48.45
information	HMM+TGI	87.69	86.27	85.02	85.75	84.83	86.66	84.70	79.99	79.81	77.56	77.13	76.20	78.63	76.95	81.94	49.60

TABLE II
WORD ACCURACY RESULTS (%) ACHIEVED BY THE PROPOSED RECONSTRUCTION TECHNIQUES USING ORACLE MASKS FOR THE DIFFERENT TEST SETS OF THE AURORA-4 DATABASE.

It can be observed in Table I that both techniques CBR and TGI clearly outperform the baseline and AFE sytems when oracle masks are used. This improvement is especially noticeable at medium and low SNRs. Thereby, the mismatch due to noise can be effectively reduced with only knowledge of the masking pattern. TGI outperforms CBR at all the SNRs. This difference can be explained by the two following reasons. First, the correlation between features is better exploited by TGI. Second, the upper bound imposed by the noisy observation to the estimated clean features is fully embedded within TGI by using truncated Gaussian pdfs, while CBR simply post-processes the estimate to ensure this condition.

Relative improvements of 1.72 %, 1.34 % and 1.74 % regarding basic TGI are achieved when this technique is combined with the TF-MFR, PATCH and HMM approaches, respectively. Thus, the exploitation of the speech temporal redundancy improves the reconstruction performance. Again, the improvements are particularly noticeable at low SNR conditions where few reliable features are present and, thus, the accuracy of the basic CBR and TGI approaches can be limited. The three time-based approaches TF-MFR, PATCH and HMM yield similar performance for Aurora-2. Thus, no complicated temporal modeling is required in case of simple recognitions tasks such as Aurora-2. Other technique that can be found in the literature combining missing-data imputation and temporal modeling is that proposed by Börgstrom et al. [31]. This technique employs HMMs and noise estimates to reconstruct noise-corrupted features. When oracle masks are used, the average recognition accuracy over SNRs between 0-20 dB obtained by this technique for the test Set A of Aurora-2 is 94.83 % (see Table II in [31]), whereas our proposed HMM+TGI achieves 96.59 %. This result confirms that, under proper knowledge of the masking pattern, the mismatch introduced by the noise can be significantly palliated just by a suitable exploitation of source correlations (represented by prior speech models).

The results obtained for the Aurora-4 database are summarized in Table II. Again the proposed reconstruction techniques suffer little degradation with respect to the clean condition (T-01) for those conditions in which only additive noise is considered (T-02 to T-07). For the conditions where additive and convolutional noises are present (from T-08 to T-14), the performance of all the techniques drops. Hence, CMN is not completely capable to compensate for channel mismatches and, therefore, a more powerful technique is needed.

HMM+TGI presents the best recognition performance. This result can be attributed to the more sophisticated temporal modeling applied by HMM+TGI in comparison with TF-MFR+TGI and PATCH+TGI. While only a small number of frames are involved during the estimation of a given missing feature in TF-MFR+TGI or PATCH+TGI, HMM+TGI uses the whole utterance. Thus, HMM+TGI is expected to be less sensitive to mask estimation errors than the other two techniques. The average performance yielded by TF-MFR+TGI and PATCH+TGI is almost the same. In fact, both techniques are quite similar in the sense that information provided by neighboring frames is considered for missing-data imputation. The patch-based technique mainly differs from TF-MFR+TGI in the way that frequency and temporal information is combined. For PATCH+TGI, time-frequency patches of speech are modeled and later used for reconstruction. In TF-MFR+TGI, both sources of information are independently modeled and then combined to obtain the final estimation.

B. Experiments using estimated masks

Table III presents the recognition accuracy results for Aurora-2 when estimated masks are applied. These results confirm the advisability of introducing temporal information in the imputation. Furthermore, it can be seen that HMM+TGI outperforms TF-MFR+TGI and PATCH+TGI again. The differ-

		Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg.	R.I.%
	Baseline	99.11	97.29	92.55	75.56	42.82	22.69	12.92	63.28	-
	AFE	99.23	98.15	96.66	93.23	84.57	62.50	30.33	80.67	27.48
No temporal	CBR	98.88	97.58	95.28	89.58	77.19	52.77	23.63	76.42	20.76
information	TGI	98.91	97.44	95.03	89.12	77.23	53.97	24.57	76.61	21.07
Using	TF-MFR+TGI	98.94	97.76	95.59	90.28	78.42	55.35	25.96	77.47	22.43
temporal	PATCH+TGI	98.93	97.73	95.49	90.15	77.92	54.03	25.38	77.09	21.83
information	HMM+TGI	98.92	97.49	95.12	89.80	78.96	58.01	29.93	78.32	23.77

TABLE III
WORD ACCURACY RESULTS (%) FOR THE AURORA-2 DATABASE USING ESTIMATED MASKS.

		T-01	T-02	T-03	T-04	T-05	T-06	T-07	T-08	T-09	T-10	T-11	T-12	T-13	T-14	Avg.	R.I.%
	Baseline	87.69	75.30	53.24	53.15	46.80	56.36	45.38	77.04	64.24	45.30	42.07	36.15	47.43	36.67	54.77	-
	AFE	88.25	81.41	69.14	64.80	67.44	66.34	68.78	80.57	74.76	61.89	56.47	58.75	60.13	59.87	68.47	25.01
No temporal	CBR	87.07	81.32	59.61	53.15	59.01	58.10	61.72	79.36	74.54	53.93	47.36	50.12	51.62	55.80	62.34	13.81
information	TGI	87.43	81.39	60.00	55.86	58.29	60.36	59.44	79.26	74.63	55.45	47.38	48.38	54.04	54.94	62.63	14.35
Using	TF-MFR+TGI	87.45	82.44	64.66	58.79	62.81	62.10	66.22	80.72	75.71	57.39	50.33	52.03	55.09	58.27	65.29	19.19
temporal	PATCH+TGI	87.15	82.01	64.73	60.00	63.35	63.65	66.28	79.71	75.49	59.78	52.25	54.59	56.62	58.60	66.02	20.53
information	HMM+TGI	87.15	81.41	64.73	60.92	64.06	63.80	66.63	79.22	75.44	60.64	54.39	56.85	57.44	60.36	66.65	21.68

TABLE IV
WORD ACCURACY RESULTS (%) FOR THE AURORA-4 DATABASE USING ESTIMATED MASKS.

ence in performance between TF-MFR+TGI and PATCH+TGI is again small.

The results obtained for Aurora-4 using estimated masks are summarized in Table IV. The exploitation of temporal correlation results in a more effective noise compensation. Thus, relative improvements of 4.25 %, 5.41 % and 6.32 % are achieved by TF-MFR+TGI, PATCH+TGI, and HMM+TGI with respect to TGI, respectively. For this task, PATCH+TGI outperforms TF-MFR+TGI. The more complex modeling applied by PATCH+TGI permits better discrimination and reconstruction of the unreliable features. For both databases, the superiority of HMMs for temporal modeling is indicative of its higher robustness against mask estimation errors.

It should be noted that when the estimated masks are used, the best MD imputation results achieved (HMM+TGI) are still lower than those obtained with AFE for both Aurora-2 (more clearly for low SNRs) and Aurora-4 (in all cases except for the quasi-stationary car noises, T-02 and T-09). This poor performance is largely due to the simple noise estimation technique employed, which can not suitably account for non-stationarity. However, this simple noise estimator has been useful to demonstrate the utility of the proposed temporal modeling for the MD approaches, which is able to improve the ASR performance over other MD systems consistently using the estimated masks as well as the oracle masks. In particular, the large relative improvement yielded by the proposed MD approach when the oracle masks are used, as shown in Tables I and II, clearly demonstrates the effectiveness of improved temporal modeling and makes it possible to outperform AFE with a better noise estimator.

VI. CONCLUSIONS

In this paper, a common framework for missing-feature spectral reconstruction has been proposed. Assuming that knowledge about feature reliability is provided in advance, MMSE estimation of the missing features is computed by exploiting the known across-frequency correlation with the reliable features. During the derivation of the estimator, some expressions not analytically solvable appear. Hence, different assumptions for the computation of these expressions are taken, leading to the proposed truncated-Gaussian based imputation. Furthermore, the relationship of the proposed method to some similar techniques is also discussed.

A qualitative analysis of the proposed imputation techniques reveals that, under low SNR conditions, these techniques exhibit poor performance since few reliable features can be found at these conditions. In order to improve the performance in such cases, this paper has proposed two novel approaches to additionally exploit the speech temporal redundancy. First, time-frequency patches of speech containing a few frames are modeled, so that short-term temporal correlation within acoustic units such as phones can be accounted for. These patches are later applied for spectral reconstruction. Second, an HMM modeling is considered to represent the speech sequential structure. This HMM acts as a source model which helps to improve the probability distributions required to obtain the MMSE estimates.

The proposed techniques were tested on the Aurora-2 digit recognition task and on the Aurora-4 large vocabulary task. In all the cases, the proposed spectral reconstruction techniques provide significative gains in recognition accuracy over the baseline system for both oracle and estimated masks. Moreover, further improvements in the recognition accuracy are obtained when temporal correlations are also considered. In particular, the HMM-based approach consistently presents the best overall performance of all the methods included in this research, which is an indicative of its robustness against mask estimation errors.

APPENDIX A
TRUNCATED GAUSSIAN DISTRIBUTION

Let Y be a Gaussian-distributed random variable with parameters $Y \sim \mathcal{N}(\mu, \sigma)$ and X a subset of Y taking values in the interval $X \in [a, b]$. Then, X is a truncated Gaussian

distributed variable and its pdf can be defined as,

$$p\left(x|a\leq x\leq b,\mu,\sigma\right)\equiv\mathcal{T}\left(x;\mu,\sigma,a,b\right)$$
$$=\begin{cases}\gamma_{a,b}\mathcal{N}(x;\mu,\sigma) & a\leq x\leq b\\ 0 & otherwise\end{cases} \tag{47}$$

where $\gamma_{a,b}$ is a normalization factor that makes the integral of the truncated distribution equal to one. This factor is

$$\gamma_{a,b}=\frac{1}{\Phi(\beta)-\Phi(\alpha)} \tag{48}$$

being α and β the standardized values corresponding to a and b, respectively. That is,

$$\alpha=\frac{a-\mu}{\sigma} \tag{49}$$

$$\beta=\frac{b-\mu}{\sigma} \tag{50}$$

The first two moments of the truncated distribution are given by [49],

$$E[x|a\leq x\leq b,\mu,\sigma]=\mu-\sigma\gamma_{a,b}\left[\mathcal{N}(\beta)-\mathcal{N}(\alpha)\right] \tag{51}$$

$$Var[x|a\leq x\leq b,\mu,\sigma]=\sigma^2\left[1-\gamma_{a,b}\left(\beta\mathcal{N}(\beta)-\alpha\mathcal{N}(\alpha)\right)\right.$$
$$\left.-\gamma_{a,b}^2\left(\mathcal{N}(\alpha)-\mathcal{N}(\beta)\right)^2\right] \tag{52}$$

REFERENCES

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

[2] J. M. Baker, L. Deng, J. Glass, S. Khudanpur, C.-H. Lee, N. Morgan, and D. O'Shaughnessy, "Research developments and directions in speech recognition and understanding, Part 1," *IEEE Signal Process. Mag.*, vol. 26, no. 3, pp. 75–80, May 2009.

[3] J. M. Baker, L. Deng, S. Khudanpur, C.-H. Lee, J. Glass, N. Morgan, and D. O'Shaughnessy, "Updated MINDS report on speech recognition and understanding, Part 2," *IEEE Signal Process. Mag.*, vol. 26, no. 4, pp. 78–85, July 2009.

[4] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Commun.*, vol. 34, no. 3, pp. 267–285, June 2001.

[5] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Commun.*, vol. 43, no. 4, pp. 275–296, Sep. 2004.

[6] H. Fletcher, *Speech and hearing in communication*. New York: Van Nostrand Co., 1953.

[7] R. Warren, K. Riener, J. Bashford, and B. Brubaker, "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," *Percept. Psychophys.*, vol. 57, no. 2, pp. 175–182, Feb. 1995.

[8] M. Cooke, "A glimpsing model of speech perception in noise," *J. Acoust. Soc. Am.*, vol. 119, no. 3, pp. 1562–1573, Mar. 2006.

[9] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. ICASSP*, 1990, pp. 845–848.

[10] J. Barker, L. Josifovski, M. P. Cooke, and P. D. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. ICSLP*, 2000, pp. 373–376.

[11] N. Ma, P. Green, J. Barker, and A. Coy, "Exploiting correlogram structure for robust speech recognition with multiple speech sources," *Speech Commun.*, vol. 49, no. 12, pp. 874 – 891, Dec. 2007.

[12] J. A. Morales-Cordovilla, N. Ma, V. E. Sanchez, J. L. Carmona, A. M. Peinado, and J. Barker, "A pitch based noise estimation technique for robust speech recognition with missing data," in *Proc. ICASSP*, 2011, pp. 4808–4811.

[13] M. L. Seltzer, B. Raj, and R. M. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Commun.*, vol. 43, no. 4, pp. 379 – 393, Sep. 2004.

[14] B. J. Borgström and A. Alwan, "A statistical approach to Mel-domain mask estimation for missing-feature ASR," *IEEE Signal Process. Lett.*, vol. 17, no. 11, pp. 941–944, Nov. 2010.

[15] W. Kim and R. M. Stern, "Mask classification for missing-feature reconstruction for robust speech recognition in unknown background noise," *Speech Commun.*, vol. 53, no. 1, pp. 1 – 11, Jan. 2011.

[16] S. Badiezadegan and R. C. Rose, "Mask estimation in non-stationary noise environments for missing feature based robust speech recognition," in *Proc. Interspeech*, 2010, pp. 2062–2065.

[17] C. Cerisara, S. Demange, and J.-P. Haton, "On noise masking for automatic missing data speech recognition: A survey and discussion," *Comput. Speech Lang.*, vol. 21, no. 3, pp. 443 – 457, July 2007.

[18] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sources," *Speech Commun.*, vol. 45, no. 1, pp. 5 – 25, Jan. 2005.

[19] N. Ma, J. Barker, H. Christensen, and P. Green, "Combining speech fragment decoding and adaptive noise floor modelling," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 3, pp. 818–827, Mar. 2012.

[20] H. Van Hamme, "PROSPECT features and their application to missing data techniques for robust speech recognition," in *Proc. Interspeech*, 2004, pp. 101–104.

[21] M. Van Segbroeck and H. Van Hamme, "Advances in missing feature techniques for robust large-vocabulary continuous speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 1, pp. 123–137, Jan. 2011.

[22] W. Kim and J. Hansen, "Missing-feature reconstruction by leveraging temporal spectral correlation for robust speech recognition in background noise conditions," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 8, pp. 2111–2120, Nov. 2010.

[23] B. Raj and R. Singh, "Reconstructing spectral vectors with uncertain spectrographic masks for robust speech recognition," in *Proc. ASRU*, 2005, pp. 65–70.

[24] J. F. Gemmeke, H. Van Hamme, B. Cranen, and L. Boves, "Compressive sensing for missing data imputation in noise robust speech recognition," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 272–287, Apr. 2010.

[25] F. Faubel, H. Raja, J. McDonough, and D. Klakow, "Particle filter based soft-mask estimation for missing feature reconstruction," in *Proc. IWAENC*, 2008.

[26] F. Faubel, J. McDonough, and D. Klakow, "Bounded conditional mean imputation with Gaussian mixture models: A reconstruction approach to partly occluded features," in *Proc. ICASSP*, 2009, pp. 3869–3872.

[27] J. A. González, A. M. Peinado, A. M. Gómez, N. Ma, and J. Barker, "Combining missing-data reconstruction and uncertainty decoding for robust speech recognition," in *Proc. ICASSP*, 2012, pp. 4693–4696.

[28] 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.

[29] J. A. González, A. M. Peinado, A. M. Gomez, and J. L. Carmona, "Efficient MMSE estimation and uncertainty processing for multienvironment robust speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 5, pp. 1206–1220, July 2011.

[30] 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.

[31] B. J. Borgström and A. Alwan, "HMM-based reconstruction of unreliable spectrographic data for noise robust speech recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 6, pp. 1612–1623, Aug. 2010.

[32] L. Deng, J. Droppo, and A. Acero, "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 3, pp. 218–233, May 2004.

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

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

[35] 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.

[36] S. T. Roweis, "One microphone source separation," in *Proc. Neuroal Information Processing Systems*, 2001, pp. 793–799.

[37] ——, "Factorial models and refiltering for speech separation and denoising," in *Proc. EUROSPEECH*, 2003, pp. 1009–1012.

[38] S. J. Rennie, J. R. Hershey, and P. A. Olsen, "Single-channel multitalker speech recognition," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 66–80, Nov. 2010.
[39] A. Genz, "Numerical computation of multivariate normal probabilities," *J. Comp. Graph. Stat.*, vol. 1, no. 2, pp. 141–149, Jun. 1992.
[40] F. Seide and P. Zhao, "On using missing-feature theory with cepstral features – Approximations to the multivariate integral," in *Proc. Interspeech*, 2010, pp. 2094–2097.
[41] J. A. González, A. M. Peinado, A. M. Gómez, and N. Ma, "Log-spectral feature reconstruction based on an occlusion model for noise robust speech recognition," in *Proc. Interspeech*, 2012.
[42] H. G. 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.
[43] I. T. Jolliffe, *Principal Component Analysis*. Springer-Verlag New York, 2002.
[44] 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.
[45] H. G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends of large vocabulary task," STQ AURORA DSR Working Group, Tech. Rep., 2002.
[46] *ETSI ES 201 108 - Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms*, Std., 2000.
[47] T. H. Crystal and A. S. House, "Segmental durations in connected-speech signals: Current results," *J. Acoust. Soc. Am.*, vol. 83, no. 4, pp. 1553–1573, Apr. 1988.
[48] *Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algoritms*, ETSI 202 050 v1.1.4 Std., Nov. 2005.
[49] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous univariate distributions*. Wiley, 1994, vol. 1.

José A. Gónzalez received the B.Sc. degree in computer 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 Dpt. of Signal Theory, Networking and Communications, UGR, under a research grant from the Spanish Ministry of Education. In 2010 and 2011, he was at the Speech and Hearing Research Group, University of Shefeld, U.K., as a Visiting Research Scientist, studying the missing-data and speech fragment decoding approaches for robust speech recognition.
His research interests include noise-robust speech recognition, computational auditory scene analysis, and machine learning.

Antonio M. Peinado (M95SM05) received the M.S. and the Ph.D. degrees in Physics from the University of Granada, Granada, Spain, in 1987 and 1994, respectively.
Since 1988, he has been working at the University of Granada, where he has led several research projects related to signal processing and communications. In 1989, he was a Consultant at the Speech Research Department, AT&T Bell Labs. He earned the positions of Associate Professor (1996) and Full Professor (2010) in the Department of Signal Theory, Networking and Communications, University of Granada, and is currently director of the research group on Signal Processing, Multimedia Transmission and Speech/Audio Processing (SigMAT) at the same university.
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 international journals, conferences and project proposals. His current research interests are focused on robust speech recognition and transmission, robust video transmission, and ultrasound signal processing.

Ning Ma received the B.Sc. degree from the South China University of Technology in 2002, and the M.Sc. and Ph.D. degrees from the University of Sheffield in 2003 and 2008 respectively.
He has been a visiting research scientist at the University of Washington, Seattle, WA, studying the graphical models for robust conversational speech recognition. He is currently a research fellow at the MRC Institute of Hearing Research, working on auditory scene analysis with cochlear implants. Before that he was a research associate in the Department of Computer Science, University of Sheffield. His research interests include computational auditory scene analysis, hearing impairment, noise-robust automatic speech recognition, and speech perception in noise.

Angel M. Gómez received the M.Sc. and Ph.D. degrees in computing science from the University of Granada, Granada, Spain, in 2001 and 2006, respectively. In 2002, he joined the Research Group on Signals, Networking, and Communications (GSTC) of the University of Granada. During 2004 and 2010, he was visiting with the Speech, Language, and Music group from the University of East Anglia, U.K. and the Speech Processing Laboratory from the University of Griffith, Australia, respectively. Since 2006, he has been an Assistant Professor at the Department of Signal Theory, Networking, and Communications of the University of Granada. His research interests are in robust speech recognition and coding, and signal processing. Dr. Gómez has served as a reviewer for several international journals and conferences.

Jon Barker received the B.A. degree in electrical and information sciences from the University of Cambridge, Cambridge, U.K., in 1991 and the Ph.D. degree in computer science from the University of Sheffield, Sheffield, U.K., in 1998.
He has worked as a Researcher at GIPSA-lab, Grenoble, France, studying audiovisual speech perception and has spent time as a Visiting Research Scientist at IDIAP (Martigny, Switzerland), ICSI (Berkeley, CA) and the Columbia University. He is currently a Senior Lecturer in Computer Science at the University of Sheffield. His research interests include human speech processing, robust automatic speech recognition and machine listening. He has authored or coauthored over 60 papers in these areas.