
Article:
Gonzalez, J.A., Gómez, A.M., Peinado, A.M. et al. (2017) Spectral Reconstruction and Noise Model Estimation Based on a Masking Model for Noise Robust Speech Recognition. Circuits, Systems, and Signal Processing, 36. pp. 3731-3760. ISSN: 0278-081X

https://doi.org/10.1007/s00034-016-0480-7

Circuits, Systems, and Signal Processing manuscript No.
(will be inserted by the editor)

Spectral Reconstruction and Noise Model Estimation based on a Masking Model for Noise-Robust Speech Recognition

Jose A. Gonzalez · Angel M. Gómez · Antonio M. Peinado · Ning Ma · Jon Barker

Received: date / Accepted: date

Abstract An effective way to increase noise robustness in automatic speech recognition (ASR) systems is feature enhancement based on an analytical distortion model that describes the effects of noise on the speech features. One of such distortion models that has been reported to achieve a good tradeoff between accuracy and simplicity is the masking model. Under this model, speech distortion caused by environmental noise is seen as a spectral mask and, as a result, noisy speech features can be either reliable (speech is not masked by noise) or unreliable (speech is masked). In this paper we present a detailed overview of this model and its applications to noise-robust ASR. Firstly, using the masking model, we derive a spectral reconstruction technique aimed at enhancing the noisy speech features. Two problems must be solved in order to perform spectral reconstruction using the masking model: i) mask estimation, i.e. determining the reliability of the noisy features, and ii) feature imputation, i.e. estimating speech for the unreliable features. Unlike missing-data imputation techniques where the two problems are considered as independent, our technique jointly addresses them by exploiting *a priori* knowledge of the speech and noise sources in the form of a statistical model. Secondly, we propose an algorithm for estimating the noise model required by the feature enhancement technique. The proposed algorithm fits a Gaussian mixture model (GMM) to the noise by iteratively maximising the likelihood of the noisy speech signal so that noise can be estimated even during speech-dominating frames. A comprehensive set of experiments carried out on the Aurora-2 and Aurora-4 databases shows that the proposed method achieves significant improvements over the baseline system and other similar missing-data imputation techniques.

J. A. Gonzalez, N. Ma and J. Barker
Dept. of Computer Science, University of Sheffield, Sheffield, UK
E-mail: {j.gonzalez,n.ma,j.p.barker}@sheffield.ac.uk

A. M. Gómez and A. M. Peinado
Dept. of Signal Theory, Telematics and Communications, Granada, Spain
E-mail: {amgg,amp}@ugr.es

Keywords Speech recognition · noise robustness · feature compensation · noise model estimation · missing-data imputation

1 Introduction

Despite major recent advances in the field of automatic speech recognition (ASR), ASR performance is still far from that achieved by humans on the same conditions [2, 3]. One of the main reasons of the performance gap between ASR and humans is the fragility of current ASR systems to mismatches between training and testing conditions. These mismatches are due to different factors such as speaker differences (i.e. gender, age, emotion), language differences (i.e. different accents and speaking styles), and, being the topic of this paper, noise. Noise, which can refer to channel noise, reverberation, or acoustic noise, degrades ASR performance due to the distortion it causes on the speech signals. In extreme cases, e.g. at very low signal-to-noise ratio (SNR) conditions, ASR systems may become almost unusable when used in such noisy conditions.

It is therefore not surprising that noise robustness in ASR has been a very active area of research over the past three decades. We refer the reader to [25,26,48] for a comprehensive overview of this topic. In general, techniques for noise-robust ASR can be classified into two categories: feature-domain and model-domain techniques. Feature-domain techniques attempt to extract a set of features from the noisy speech signals that are less affected by noise or that better match the features used to train the system. This category can be further divided into three sub-categories: robust feature extraction techniques, which remove from the speech signals the variability irrelevant to ASR, feature normalisation techniques, in which the distribution of the testing features is normalised to match that of the training dataset, and feature compensation, where speech features are enhanced in order to compensate for the noise distortion. Model-domain techniques, on the other hand, attempt to adapt the pre-trained acoustic model to better match the environmental testing conditions. This typically involves the estimation of a transformation from an adaptation set for compensating the mismatch between the training and testing conditions and, then, applying the estimated transformation to update the acoustic model parameters.

From the above classification, one of the most effective ways to improve ASR robustness against noise is that in which the effects of noise on the speech features are explicitly modelled using an analytical distortion model. From the distortion model one can either derive a feature-domain technique to enhance the noisy features or, alternatively, the acoustic models can be adapted to the noise in order to better represent the noisy speech statistics. In both cases the challenge is to accurately estimate the characteristics of the distortion, which normally involves estimating the noise itself. Representative methods belonging to this subclass of techniques are the Wiener filter [27], vector Taylor

series (VTS) compensation [1, 31, 45], and the missing-data techniques [7, 20, 36, 37, 42].

In this paper we focus on one of such distortion models that has proved to be very effective on combating environmental noise [9, 33, 43]: the log-max model or masking model, as we will refer to it in the rest of this paper. This model was initially inspired by observations showing that the distortion caused by noise on the speech features when they are expressed in a compressed spectral domain (e.g. log-Mel features or log power spectrum) can be reasonably well approximated as a kind of spectral masking: some parts of the speech spectrum are effectively masked by noise while other parts remain unaltered.

The main objective of this work is to present an overview of the masking model and describe in detail three specific applications of it for noise-robust ASR: (i) speech feature enhancement, (ii) noise model estimation and (iii) determining the reliability of the observed noisy speech features. Firstly, we extend the work initiated by the authors in [18, 19] and present a detailed and comprehensive derivation of a feature enhancement technique based on the masking model. Unlike other feature enhancement techniques derived from the masking model (e.g. missing-data techniques), our technique has the advantage that it does not require an *a priori* segmentation of the noisy spectrum in terms of 'reliable' and 'unreliable' features, but the segmentation (a mask in the missing-data terminology) is obtained as a by-product of the spectral reconstruction process.

As we will see, the proposed technique uses prior speech and noise models for enhancing the noisy speech features. While the speech model can be easily estimated from a clean training dataset, the estimation of the noise model is more subtle. Hence, another contribution of this paper is an algorithm which estimates the statistical distribution of the environmental noise in each noisy speech signal. The distribution is represented as a Gaussian mixture model (GMM) whose parameters are iteratively updated to maximise the likelihood of the observed noisy data. The main benefit of our algorithm in comparison with other traditional approaches is that noise can be estimated even during speech segments.

Finally, another contribution of this paper is the development of a common statistical framework based on the masking model for making inferences about the speech and noise in noise-robust signal processing. This framework has enough flexibility for providing us with different statistics describing the noise effects on the speech features. For example, as will be shown later, missing-data masks, which identify the regions of the noisy speech spectrum that are degraded by noise, can be easily estimated using the proposed framework.

The rest of this paper is organised as follows. First, in Section 2, we derive the analytical expression of the masking model as an approximation to the exact distortion model between two acoustic sources (i.e. speech and additive noise) when they are expressed in the log-Mel domain. Using the masking model, a minimum mean square error (MMSE) feature enhancement technique is derived in Section 3. Then, in Section 4, we introduce the iterative algorithm for estimating the parameters of the noise model required by the

enhancement technique. Section 5 discusses the relationship between the proposed algorithms and some other similar techniques. Experimental results are given in Section 6. Finally, this paper is summarised and the main conclusions are drawn in Section 7.

2 Model of speech distortion

In this section we derive the analytical expression of the speech distortion model that will be used in the rest of the paper for speech feature enhancement and noise estimation. The model, which will be referred to as the *masking model*, can be considered as an approximation to the exact interaction function between two acoustic sources in the log-power domain or any other domain that involves a logarithmic compression of the power spectrum such as the log-Mel domain [43]. We start the derivation of the model with the standard additive noise assumption in the discrete time domain,

$$y[t] = x[t] + n[t], \tag{1}$$

where y, x, and n are the noisy speech, clean speech, and noise signals, respectively. Denoting by $Y[f]$, $X[f]$, and $N[f]$ the short-time Fourier transforms of the above signals (f is the frequency-band index), then the power spectrum of the noisy speech signal is

$$|Y[f]|^2 = |X[f]|^2 + |N[f]|^2 + 2|X[f]||N[f]|\cos\theta_f, \tag{2}$$

where $\theta_f = |\theta_f^x - \theta_f^n|$ is the difference between the phases of $X[f]$ and $N[f]$.

To simplify the derivation of the distortion model, it is common practice to assume that speech and noise are independent (i.e. $\mathbb{E}[\cos\theta_f] = 0$). It is possible, however, to account for the phase differences between both sources. This is known as *phase-sensitive model* and although it has been shown that this model is superior to its phase-insensitive counterpart (see e.g. [11, 15, 24, 46]), we will not consider it in this paper.

The power spectrum of the noisy signal is then filtered through a Mel-filterbank with D filters, each of which being characterised by its transfer function $W_f^{(i)} \geq 0$ with $\sum_f W_f^{(i)} = 1$ $(i = 1, ..., D)$. The relation between the outputs of the Mel-filterbank for the noisy, clean speech and noise signals is [11],

$$|\tilde{Y}_i| = |\tilde{X}_i| + |\tilde{N}_i|, \tag{3}$$

with $|\tilde{Y}_i| = \sum_f W_f^{(i)}|Y[f]|^2$, $|\tilde{X}_i| = \sum_f W_f^{(i)}|X[f]|^2$, and $|\tilde{N}_i| = \sum_f W_f^{(i)}|N[f]|^2$.

Let us now define the vector with the noisy log-Mel energies as $\boldsymbol{y} = (\log|\tilde{Y}_1|, \ldots, \log|\tilde{Y}_D|)$ and similarly for the clean speech and noise signals as $\boldsymbol{x}$ and $\boldsymbol{n}$, respectively. Then, these variables are related as follows

$$\boldsymbol{y} = \log(e^{\boldsymbol{x}} + e^{\boldsymbol{n}}). \tag{4}$$

This expression can be rewritten as

$$\begin{aligned}\boldsymbol{y} &= \log(e^{\max(\boldsymbol{x},\boldsymbol{n})} + e^{\min(\boldsymbol{x},\boldsymbol{n})}) \\ &= \max(\boldsymbol{x},\boldsymbol{n}) + \log\left(\mathbf{1} + e^{\min(\boldsymbol{x},\boldsymbol{n})-\max(\boldsymbol{x},\boldsymbol{n})}\right) \\ &= \max(\boldsymbol{x},\boldsymbol{n}) + \boldsymbol{\varepsilon}(\boldsymbol{x}-\boldsymbol{n}),\end{aligned} \tag{5}$$

with $\max(\boldsymbol{x},\boldsymbol{n})$ and $\min(\boldsymbol{x},\boldsymbol{n})$ being the element-wise maximum and minimum operations and

$$\boldsymbol{\varepsilon}(\boldsymbol{z}) = \log\left(\mathbf{1} + e^{-|\boldsymbol{z}|}\right). \tag{6}$$

The additive term $\boldsymbol{\varepsilon}$ in (5) can be thought of as an approximation error that depends on the absolute value of the signal-to-noise ratio (SNR) between speech and noise. Fig. 1a shows a plot of (6) for different SNR values. It can be seen that $\boldsymbol{\varepsilon}$ achieves its maximum value at 0 dB where $\varepsilon(0) = \log(2) \approx 0.69$. On the other hand, this term becomes negligible when the difference between speech and noise exceeds 20 dB. A more detailed analysis of the statistics of $\boldsymbol{\varepsilon}$ computed over the whole test set A of the Aurora-2 database [23] for all the $D = 23$ log-Mel filterbank channels is shown in Figs. 1b and 1c. In particular, Fig. 1b shows an histogram of $\boldsymbol{\varepsilon}$ estimated from all the SNR conditions in the test set A of Aurora-2. We used the clean and noisy recordings available in this database to estimate $\boldsymbol{x}$ and $\boldsymbol{n}$ required for computing $\boldsymbol{\varepsilon}(\boldsymbol{z})$. From the figure, it is clear that the error is small and mostly concentrated around zero with an exponentially-decaying probability that vanished in its maximum value $\log(2)$. Fig. 1b also shows that $\boldsymbol{\varepsilon}$ can take negative values. These negative values are due to the phase term in (2) which we ignore in this work[1]. Nevertheless, the probability of the negative error values is very small. An histogram of the relative errors $|\varepsilon(z_i)/y_i|$ $(i = 1, \ldots, D)$ is shown in Fig. 1c. Again, the relative error is mostly concentrated around zero and it very rarely exceeds $\boldsymbol{y}$ more than 10 % in magnitude.

From the above discussion, we conclude that $\boldsymbol{\varepsilon}(\boldsymbol{z})$ can be omitted from (5) without sacrificing much accuracy. After doing this, we finally reach the following speech distortion model,

$$\boldsymbol{y} \approx \max(\boldsymbol{x},\boldsymbol{n}). \tag{7}$$

This model, which was originally proposed in [32, 47] for noise adaptation, is known in the literature as the *log-max* approximation [33, 44, 47], MIXMAX model [32, 34, 43] and, also, masking model [18, 19]. Here, we will employ the last name because the approach reminds the perceptual masking phenomena of the human auditory system. It must be pointed out that although it is an approximation in nature, it can be shown that the masking model turns to be the expected value of the exact interaction function (i.e. distortion model) for

[1] According to (2), the power spectrum of the clean speech and noise signals at a given frequency band f can exceed that of the noisy speech signal if $\cos\theta_f < 0$ and, thus, the difference $\boldsymbol{y} - \max(\boldsymbol{x},\boldsymbol{n})$ can be negative

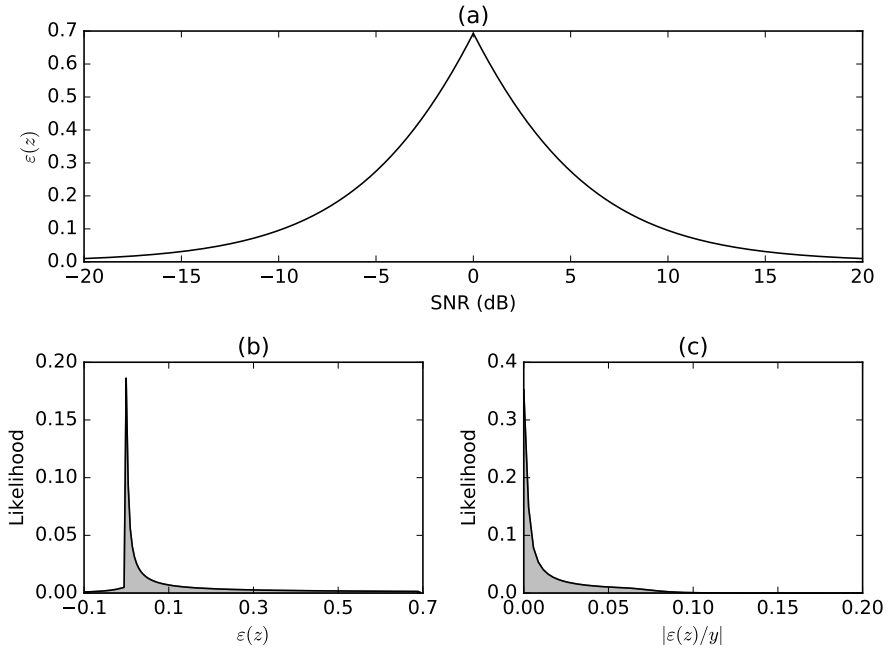

Fig. 1 Error of the log-max distortion model. (a) Plot of $\varepsilon(z)$ in (6) for different SNR values. (b) Histogram of $\varepsilon(z)$ estimated from all the utterances in test set A of the Aurora-2 database. A parametrization consisting of $D = 23$ log-Mel filterbank features is employed. (c) Histogram of relative errors also computed from the set A of Aurora-2.

two acoustic sources when the phase difference θ_f in (2) between the sources is uniformly distributed [34, 43].

According to (7), the effect of additive noise on speech simplifies to a binary masking in the log-Mel domain. Thus, the problem of speech feature compensation can be reformulated as two independent problems:

1. **Mask estimation**: this problem involves the segmentation of the noisy spectrum into masked and non-masked regions [6]. As a result, a binary mask $\boldsymbol{m}$ is usually obtained. This mask indicates, for each element y_i of the noisy spectrum, whether the element is dominated either by speech or noise, i.e.,

$$m_i = \begin{cases} 1, \text{ if } x_i > n_i \\ 0, \text{ otherwise} \end{cases} . \tag{8}$$

2. **Spectral reconstruction**: this problem involves the estimation of the clean speech features for those regions of the noisy spectrum that are masked by noise. To do so, the redundancy of speech is exploited by taking into account the correlation among the masked and non-masked speech features.

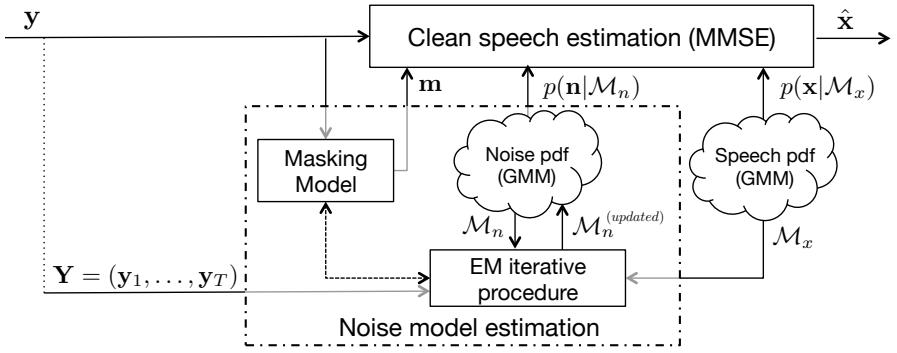

Fig. 2 Noise compensation approach proposed for ASR. An MMSE-based estimator provides clean speech estimates from noisy features using speech and noise priors and masks from the masking model. The noise model (based on GMMs) is also obtained by means of the masking model by applying an iterative EM algorithm which maximises the likelihood of the observed noisy data.

This approach based on two independent steps, mask estimation and spectral reconstruction, is the one followed by missing-data techniques [7, 16, 20, 35–37, 42]. In the next section we present an alternative, statistical approach for feature enhancement in which both problems are jointly addressed under the constraints imposed by the masking model. As we will see, our technique can be considered as a more general and robust approach which contains as particular cases the mask estimation and spectral reconstruction steps.

3 Spectral reconstruction using the masking model

The masking model derived in the last section provides us with an analytical expression that relates the (observed) noisy features with the (hidden) clean speech and noise features. This, together with statistical models for speech and noise, enables us to make inferences about the clean speech and noise sources. For speech feature enhancement we will see later that the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$ need to be estimated. Section 3.1 will address this issue. Once this distribution is estimated, it can be used to make predictions about the clean speech features and, thus, compensating for the noise distortion. The details of this estimator will be presented in Section 3.2.

It is worth mentioning here that the estimation algorithm presented in this section is similar in some aspects to other algorithms proposed in the literature for feature compensation [18, 19, 33], model decomposition [43, 47] and single-channel speaker separation [40, 41]. Nevertheless, contrary to previous work, the problem we address here is that of speech feature enhancement for noise-robust ASR under the assumption that the corrupting source (noise) is distributed according to a GMM.

Figure 2 shows a block diagram of the proposed noise-robust system comprising speech feature enhancement (Clean speech estimation) and noise model

estimation. As can be observed, GMM models are used for modelling both the distributions of speech and noise. As will be shown in Section 4, the masking model together with the inference machinery developed in this section will allow us not to only estimate the clean speech features, but also to perform noise model estimation.

3.1 Posterior of clean speech features

To compute the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$, we assume that the feature vectors $\boldsymbol{x}$ and $\boldsymbol{n}$ are i.i.d. and can be accurately modelled using GMMs[2] $\mathcal{M}_x$ and $\mathcal{M}_n$ for speech and noise, respectively. Thus,

$$p(\boldsymbol{x}|\mathcal{M}_x) = \sum_{k_x=1}^{K_x} \pi^{(k_x)} \mathcal{N}_x(\boldsymbol{x}; \boldsymbol{\mu}_x^{(k_x)}, \boldsymbol{\Sigma}_x^{(k_x)}), \tag{9}$$

$$p(\boldsymbol{n}|\mathcal{M}_n) = \sum_{k_n=1}^{K_n} \pi^{(k_n)} \mathcal{N}_n(\boldsymbol{n}; \boldsymbol{\mu}_n^{(k_n)}, \boldsymbol{\Sigma}_n^{(k_n)}), \tag{10}$$

where $\{\pi^{(k_x)}, \boldsymbol{\mu}_x^{(k_x)}, \boldsymbol{\Sigma}_x^{(k_x)}\}$ are the prior probability, mean vector, and covariance matrix of the k_x-th Gaussian distribution in the clean-speech GMM, and $\{\pi^{(k_n)}, \boldsymbol{\mu}_n^{(k_n)}, \boldsymbol{\Sigma}_n^{(k_n)}\}$ denote the parameters of the k_n-th component in the noise model. The parameters of the clean-speech GMM can be easily estimated from the clean-speech training dataset using the Expectation-Maximisation (EM) algorithm [10]. Similarly, as we will see in Section 4, an iterative procedure based on the EM algorithm can be employed to estimate the noise distribution in each utterance.

Equipped with these prior models, we are ready now to make inferences about the clean speech features given the observed noisy ones. Inference involves the estimation of $p(\boldsymbol{x}|\boldsymbol{y})$, which can be expressed as

$$p(\boldsymbol{x}|\boldsymbol{y}) = \sum_{k_x=1}^{K_x} \sum_{k_n=1}^{K_n} p(\boldsymbol{x}|\boldsymbol{y}, k_x, k_n) P(k_x, k_n|\boldsymbol{y}), \tag{11}$$

where we have omitted the dependence on the models $\mathcal{M}_x$ and $\mathcal{M}_n$ to keep notation uncluttered. It can be observed that this probability requires the computation of two terms, $P(k_x, k_n|\boldsymbol{y})$ and $p(\boldsymbol{x}|\boldsymbol{y}, k_x, k_n)$. Let us first focus on the computation of $P(k_x, k_n|\boldsymbol{y})$, which can be expressed through Bayes' rule as,

$$P(k_x, k_n|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|k_x, k_n)\pi^{(k_x)}\pi^{(k_n)}}{\sum_{k'_x=1}^{K_x}\sum_{k'_n=1}^{K_n} p(\boldsymbol{y}|k'_x, k'_n)\pi^{(k'_x)}\pi^{(k'_n)}}. \tag{12}$$

[2] Besides GMMs, other generative models can also be used for modelling these distributions. In particular, spectral reconstruction can benefit from the use of more complex speech priors such as hidden Markov models (HMMs) along with language models, as it is usually done in automatic speech recognition. These priors are expected to provide more accurate estimates of the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$ and, thus, leading to better clean speech estimates.

where the likelihood $p(\boldsymbol{y}|k_x, k_n)$ is defined as the following marginal distribution,

$$\begin{aligned} p(\boldsymbol{y}|k_x, k_n) &= \iint p(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{y}|k_x, k_n) d\boldsymbol{x} d\boldsymbol{n} \\ &= \iint p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n}) p(\boldsymbol{x}|k_x) p(\boldsymbol{n}|k_n) d\boldsymbol{x} d\boldsymbol{n}, \end{aligned} \tag{13}$$

In this equation we have assumed that $\boldsymbol{y}$ is conditionally independent of Gaussians k_x and k_n given $\boldsymbol{x}$ and $\boldsymbol{n}$. As $p(\boldsymbol{x}|k_x)$ and $p(\boldsymbol{n}|k_n)$ just involve the evaluation of two Gaussian distributions, $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n})$ is the only unknown term in (13). According to the masking model in (7), each noisy feature y_i is the maximum of x_i and n_i. Therefore, $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n})$ can be expressed as the following product

$$p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n}) = \frac{1}{\mathcal{K}} \prod_{i=1}^{D} p(y_i|x_i, n_i), \tag{14}$$

where $\mathcal{K}$ is an appropriate normalization factor that assures $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n})$ integrates to one and $p(y_i|x_i, n_i)$ is defined as

$$\begin{aligned} p(y_i|x_i, n_i) &= \delta\left(y_i - \max(x_i, n_i)\right) \\ &= \delta(y_i - x_i)\mathbb{1}_{n_i \leq x_i} + \delta(y_i - n_i)\mathbb{1}_{x_i < n_i} \end{aligned} \tag{15}$$

with $\delta(\cdot)$ being the Dirac delta function and $\mathbb{1}_{\mathcal{C}}$ is an indicator function that equals to one if the condition $\mathcal{C}$ is true, otherwise it is zero.

After expanding the multiplication in (14) and grouping terms, we can rewrite (14) as,

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{n}) \propto\ & [\delta(y_1 - x_1)\delta(y_2 - x_2) \ldots \delta(y_D - x_D)\mathbb{1}_{n_1 \leq x_1}\mathbb{1}_{n_2 \leq x_2} \ldots \mathbb{1}_{n_D \leq x_D}] \\ & + [\delta(y_1 - x_1)\delta(y_2 - x_2) \ldots \delta(y_D - n_D)\mathbb{1}_{n_1 \leq x_1}\mathbb{1}_{n_2 \leq x_2} \ldots \mathbb{1}_{x_D < n_D}] \\ & + \ldots \\ & + [\delta(y_1 - n_1)\delta(y_2 - n_2) \ldots \delta(y_D - n_D)\mathbb{1}_{x_1 < n_1}\mathbb{1}_{x_2 < n_2} \ldots \mathbb{1}_{x_D < n_D}]. \end{aligned} \tag{16}$$

Each expression enclosed in brackets in the above equation represents a different segregation hypothesis for $\boldsymbol{y}$. For instance, the first expression is the hypothesis $\boldsymbol{y} = \boldsymbol{x}$, while the last one corresponds to $\boldsymbol{y} = \boldsymbol{n}$. The rest of the expressions represent hypotheses in which some elements in $\boldsymbol{y}$ are dominated by the speech and the rest by the noise.

Inference in the above model is analytically intractable since, after using (16) in (13), the likelihood $p(\boldsymbol{y}|k_x, k_n)$ results in the evaluation of 2^D double integrals. For a typical front-end consisting of $D = 23$ Mel channels, the computational cost of evaluating the integrals is clearly prohibitive. Furthermore, the integrals involve the evaluation of Gaussian cumulative density functions

(cdfs) for which no closed-form analytical solution exists in case of using distributions with full covariance matrices. To address the above two problems, we simplify the likelihood computation in (13) by assuming that the noisy features are conditionally independent given the Gaussian components k_x and k_n. Thus, instead of evaluating the 2^D possible segregation hypotheses, only 2 hypotheses are evaluated for each noisy feature: those corresponding to whether the feature is masked by noise or not. Under the independence assumption, the likelihood $p(\boldsymbol{y}|k_x, k_n)$ in (13) becomes,

$$p(\boldsymbol{y}|k_x, k_n) = \prod_{i=1}^{D} p(y_i|k_x, k_n), \tag{17}$$

with

$$p(y_i|k_x, k_n) = \iint p(y_i|x_i, n_i)p(x_i|k_x)p(n_i|k_n)dx_i dn_i. \tag{18}$$

By substituting the expression of the observation model in (15) into (18), we obtain the following likelihood function:

$$\begin{aligned}
&p(y_i|k_x, k_n) \\
&\quad = \iint p(x_i|k_x)p(n_i|k_n)\delta(y_i - x_i)\mathbb{1}_{n_i \leq x_i} dx_i dn_i + \\
&\qquad \iint p(x_i|k_x)p(n_i|k_n)\delta(y_i - n_i)\mathbb{1}_{x_i < n_i} dx_i dn_i \\
&\quad = p(y_i|k_x)\int_{-\infty}^{y_i} p(n_i|k_n)dn_i + p(y_i|k_n)\int_{-\infty}^{y_i} p(x_i|k_x)dx_i \\
&\quad = p(x_i = y_i, n_i \leq y_i|k_x, k_n) + p(n_i = y_i, x_i < y_i|k_x, k_n)
\end{aligned} \tag{19}$$

where

$$p(x_i = y_i, n_i \leq y_i|k_x, k_n) = \mathcal{N}_x\left(y_i; \mu_{x,i}^{(k_x)}, \sigma_{x,i}^{(k_x)}\right) \Phi_n\left(y_i; \mu_{n,i}^{(k_n)}, \sigma_{n,i}^{(k_n)}\right) \tag{20}$$

$$p(n_i = y_i, x_i < y_i|k_x, k_n) = \mathcal{N}_n\left(y_i; \mu_{n,i}^{(k_n)}, \sigma_{n,i}^{(k_n)}\right) \Phi_x\left(y_i; \mu_{x,i}^{(k_x)}, \sigma_{x,i}^{(k_x)}\right) \tag{21}$$

and $\mathcal{N}(\cdot; \mu, \sigma)$ and $\Phi(\cdot; \mu, \sigma)$ are, respectively, the Gaussian pdf and cdf with mean μ and standard deviation σ. We can observe that the likelihood has two terms: $p(x_i = y_i, n_i \leq y_i|k_x, k_n)$ is the probability of speech energy being dominant, while $p(n_i = y_i, x_i < y_i|k_x, k_n)$ is the probability that speech is masked by noise.

We now focus on the computation of the posterior $p(\boldsymbol{x}|\boldsymbol{y}, k_x, k_n)$ in (11). Assuming again independence among the features, this probability can be expressed as the following marginal distribution:

$$
\begin{aligned}
p(x_i|y_i,k_x,k_n) &= \int p(x_i,n_i|y_i,k_x,k_n)dn_i \\
&= \frac{\int p(y_i|x_i,n_i)p(x_i|k_x)p(n_i|k_n)dn_i}{p(y_i|k_x,k_n)} \\
&= \frac{\mathcal{N}_x\left(x_i;\mu_{x,i}^{(k_x)},\sigma_{x,i}^{(k_x)}\right)\Phi_n\left(y_i;\mu_{n,i}^{(k_n)},\sigma_{n,i}^{(k_n)}\right)\delta(x_i-y_i)}{p(y_i|k_x,k_n)} + \\
&\quad \frac{\mathcal{N}_n\left(y_i;\mu_{n,i}^{(k_n)},\sigma_{n,i}^{(k_n)}\right)\mathcal{N}_x\left(x_i;\mu_{x,i}^{(k_x)},\sigma_{x,i}^{(k_x)}\right)\mathbb{1}_{x_i<y_i}}{p(y_i|k_x,k_n)}
\end{aligned} \tag{22}
$$

To derive this equation we have proceeded as in (19), that is, $p(x_i|y_i,k_x,k_n)$ is expressed as the sum of two terms: one for the hypothesis that speech energy is dominant, and the other for the hypothesis that speech is masked by noise. We will see in the next section that these two terms may be interpreted as a speech presence probability (SPP) and a noise presence probability (NPP), respectively.

3.2 MMSE estimation

Equation (11) together with (19) and (22) form the basis of the procedure that will be used in this section to perform speech feature enhancement. This will be done using MMSE estimation as follows,

$$
\hat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}] = \int \boldsymbol{x}p(\boldsymbol{x}|\boldsymbol{y})d\boldsymbol{x}, \tag{23}
$$

that is, the estimated clean feature vector is the mean of the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$, which is given by (11). Then,

$$
\hat{\boldsymbol{x}} = \sum_{k_x=1}^{K_x}\sum_{k_n=1}^{K_n} P(k_x,k_n|\boldsymbol{y}) \underbrace{\int \boldsymbol{x}p(\boldsymbol{x}|\boldsymbol{y},k_x,k_n)d\boldsymbol{x}}_{\hat{\boldsymbol{x}}^{(k_x,k_n)}}. \tag{24}
$$

In the above equation, $P(k_x,k_n|\boldsymbol{y})$ is computed according to (12), while $\hat{\boldsymbol{x}}^{(k_x,k_n)}$ denotes the partial clean-speech estimate given the Gaussian components k_x and k_n. For computing $\hat{\boldsymbol{x}}^{(k_x,k_n)}$ we again assume that the features are independent. Then,

$$
\hat{x}_i^{(k_x,k_n)} = \int x_i p(x_i|y_i,k_x,k_n)dx_i, \tag{25}
$$

By replacing $p(x_i|y_i,k_x,k_n)$ by its value given in (22), we finally arrive at the following expression for computing the partial estimates,

$$
\hat{x}_i^{(k_x,k_n)} = w_i^{(k_x,k_n)}y_i + \left(1-w_i^{(k_x,k_n)}\right)\tilde{\mu}_{x,i}^{(k_x)}(y_i), \tag{26}
$$

where $w_i^{(k_x,k_n)}$ is the following speech presence probability

$$w_i^{(k_x,k_n)} = \frac{\mathcal{N}_x\left(y_i;\mu_{x,i}^{(k_x)},\sigma_{x,i}^{(k_x)}\right)\Phi_n\left(y_i;\mu_{n,i}^{(k_n)},\sigma_{n,i}^{(k_n)}\right)}{p(y_i|k_x,k_n)}, \tag{27}$$

and $\tilde{\mu}_{x,i}^{(k_x)}(y_i)$ is the expected value of the k_x-th Gaussian when its support is $x_i \in (\infty, y_i]$. For a general Gaussian distribution $\mathcal{N}(x;\mu,\sigma)$, the mean and variance of the so-called right-truncated distribution for $x \in (\infty, y]$ are (see e.g. [12]),

$$\tilde{\mu}(y) = \mathbb{E}[x|x \leq y, \mu, \sigma] = \mu - \sigma\rho(\overline{y}), \tag{28}$$

$$\tilde{\sigma}^2(y) = \mathrm{Var}[x|x \leq y, \mu, \sigma] = \sigma^2\left[1 - \overline{y}\rho(\overline{y}) - \rho(\overline{y})^2\right], \tag{29}$$

where $\overline{y} = (y-\mu)/\sigma$ and $\rho(\overline{y}) = \mathcal{N}(\overline{y})/\Phi(\overline{y})$ represents the quotient between the pdf and cdf of standard normal distribution.

By substituting (26) into (24), we obtain the following final expression for the MMSE estimate of the clean speech features,

$$\begin{aligned}\hat{x}_i &= \sum_{k_x=1}^{K_x}\sum_{k_n=1}^{K_n} P(k_x,k_n|\boldsymbol{y})\left[w_i^{(k_x,k_n)}y_i + \left(1 - w_i^{(k_x,k_n)}\right)\tilde{\mu}_{x,i}^{(k_x)}(y_i)\right] \\ &= m_i y_i + \sum_{k_x=1}^{K_x}\left(P(k_x|\boldsymbol{y}) - m_i^{(k_x)}\right)\tilde{\mu}_{x,i}^{(k_x)}(y_i),\end{aligned} \tag{30}$$

with

$$m_i = \sum_{k_x=1}^{K_x}\sum_{k_n=1}^{K_n} P(k_x,k_n|\boldsymbol{y})w_i^{(k_x,k_n)}, \tag{31}$$

$$m_i^{(k_x)} = \sum_{k_n=1}^{K_n} P(k_x,k_n|\boldsymbol{y})w_i^{(k_x,k_n)}. \tag{32}$$

For convenience, we will refer to the estimator in (30) as the masking-model based spectral reconstruction (MMSR) from now on. As can be seen in (30), the MMSR estimate $\hat{x}_i$ is obtained as a weighted combination of two terms. The first term, y_i, is the estimate of the clean feature when the noise is masked by speech and, hence, the estimate is the observation itself. On the other hand, the second term in (30) corresponds to the estimate when speech is completely masked by noise. In this second case the exact level of speech energy is unknown, but the masking model enforces it to be upper bounded by the observation y_i. In this manner, the sums $\tilde{\mu}_{x,i}^{(k_x)}(y_i)$ in (30) are the means for the truncated Gaussians $k_x = 1,\ldots,K_x$ when $x_i \in (-\infty, y_i]$. An interesting aspect of the MMSR estimator is that, as a by-product of the estimation process, it automatically computes a reliability mask m_i for each element of the noisy spectrum. The elements of this mask are in the interval $m_i \in [0,1]$, thus indicating the degree in which the observation y_i is deemed to

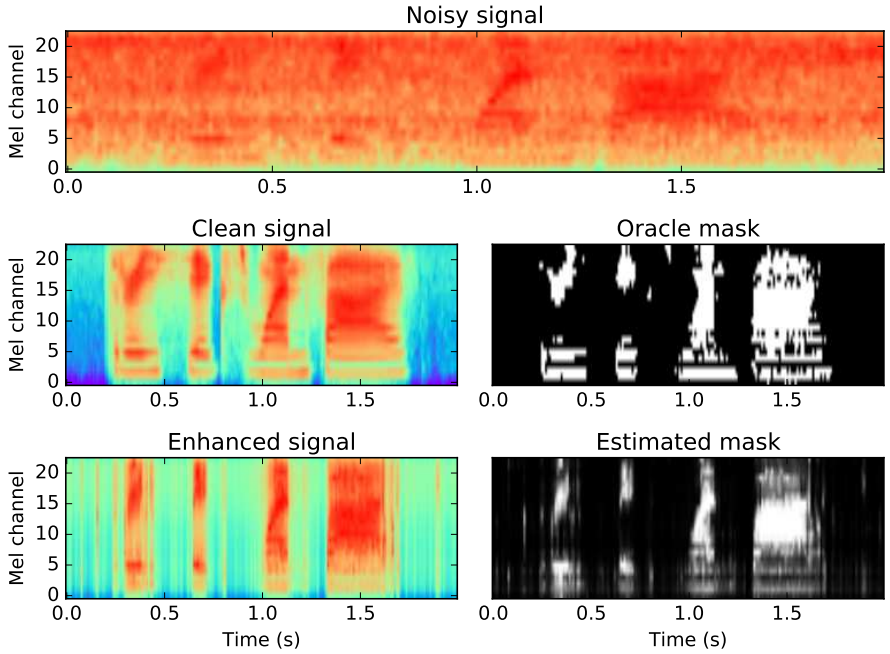

Fig. 3 Example of Log-Mel spectrograms for the utterance "three six one five" from the Aurora-2 database. *[Top panel]* Noisy speech signal distorted by car noise at 0 dB. *[Left column: top and bottom]* Original and enhanced speech signals. To obtain the enhanced signal, 256-mixture and 1-mixture GMMs are used to model speech and noise, respectively. *[Right column: top and bottom]* Oracle and estimated missing-data masks. White represents reliable regions (i.e. dominated by speech) and black unreliable regions (i.e. dominated by noise). The oracle mask is obtained from the clean and noisy signals using a 0 dB SNR threshold. The estimated soft-mask is computed using (31).

be dominated by speech or noise. As we will see in the next section, this mask will play an important role when estimating the model of the environmental noise in each utterance.

Fig. 3 shows examples of a signal reconstructed by the proposed method and the corresponding estimated soft-mask m_i in (31). In the example, the method is able to suppress the background noise while keeping those spectral regions dominated by speech. Also, the method is able to some extent to recover the speech information on those regions masked by noise by exploiting the correlations with the 'reliable' observed features and the prior information provided by the clean speech model. Finally, it is worth pointing out the similarity between the estimated soft-mask and the oracle mask computed from the clean and noisy signals.

4 Noise model estimation

The MMSR algorithm introduced in the last section requires a model of the corrupting noise for computing the corresponding speech and noise presence probabilities. Often, a voice activity detector (VAD) [38, 39] is used to detect the speech and non-speech segments in the noisy signal and, then, noise is estimated from the latter segments. Other traditional noise estimation methods are based on tracking spectral minima in each frequency band [29], MMSE-based spectral tracking [21] or comb-filtering [30]. These approaches have, however, several limitations. First, noise estimation accuracy tends to be poor at low SNRs. Second, noise estimates for the speech segments are usually unreliable, particularly for non-stationary noises, since the estimates are normally obtained through linear interpolation of the estimates obtained for the adjacent non-speech segments. Hence, we propose in this section a fully-probabilistic noise estimation procedure that works by iteratively maximising the likelihood of the observed noisy data (see Fig. 2).

Formally, the goal of the proposed algorithm is to find the set of noise model parameters $\hat{\mathcal{M}}_n$ that, together with the speech model $\mathcal{M}_x$, maximises the likelihood of the observed noisy data $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$,

$$\hat{\mathcal{M}}_n = \arg\max_{\mathcal{M}_n} p(\boldsymbol{Y}|\mathcal{M}_n, \mathcal{M}_x). \tag{33}$$

To optimise (33) we will make use of the EM algorithm [10]. Denoting the current noise model estimate by $\mathcal{M}_n$ and its updated version by $\hat{\mathcal{M}}_n$, we can write the auxiliary Q-function used in the EM algorithm as

$$\begin{aligned}\mathcal{Q}(\mathcal{M}_n, \hat{\mathcal{M}}_n) &= \sum_{t=1}^{T}\sum_{k_x=1}^{K_x}\sum_{k_n=1}^{K_n} \gamma_t^{(k_x,k_n)} \log p(\boldsymbol{y}_t, k_x, k_n) \\ &\propto \sum_{t=1}^{T}\sum_{k_x=1}^{K_x}\sum_{k_n=1}^{K_n} \gamma_t^{(k_x,k_n)} \left[\log p(\boldsymbol{y}_t|k_x, k_n) + \log \hat{\pi}_n^{(k_n)}\right], \end{aligned} \tag{34}$$

where we have used the following short notations: $\hat{\pi}_n^{(k_n)} = P(k_n|\hat{\mathcal{M}}_n)$ and $\gamma_t^{(k_x,k_n)} = P(k_x, k_n|\boldsymbol{y}_t, \mathcal{M}_n, \mathcal{M}_x)$. The latter posterior probability is given by (12) and it is computed using the speech model $\mathcal{M}_x$ and the current estimate of the noise model $\mathcal{M}_n$. It should be noted that the dependence on the speech and noise models has been omitted from the previous equation to keep the notation uncluttered.

By assuming that the elements of $\boldsymbol{y}_t$ are conditionally independent given Gaussians k_x and k_n, the auxiliary Q-function becomes

$$\mathcal{Q}(\mathcal{M}_n, \hat{\mathcal{M}}_n) = \sum_{t=1}^{T}\sum_{k_x=1}^{K_x}\sum_{k_n=1}^{K_n} \gamma_t^{(k_x,k_n)} \left[\sum_{i=1}^{D} \log p(y_{t,i}|k_x, k_n) + \log \hat{\pi}_n^{(k_n)}\right], \tag{35}$$

where $p(y_{t,i}|k_x, k_n)$ is given by (19).

To obtain the expressions for updating the noise model parameters, we set the derivatives of (35) w.r.t. the parameters equal to zero and solve. This yields the following set of equations for updating the Gaussian means $\hat{\mu}_{n,i}^{(k_n)}$, variances $\hat{\sigma}_{n,i}^{(k_n)^2}$ and mixture weights $\hat{\pi}_n^{(k_n)}$ $(k_n = 1, \ldots, K_n; i = 1, \ldots, D)$:

$$\hat{\pi}_n^{(k_n)} = \frac{1}{T} \sum_{t=1}^{T} \gamma_t^{(k_n)} \tag{36}$$

$$\hat{\mu}_{n,i}^{(k_n)} = \frac{\sum_{t=1}^{T} m_{t,i}^{(k_n)} \tilde{\mu}_{n,i}^{(k_n)}(y_{t,i}) + \left(\gamma_t^{(k_n)} - m_{t,i}^{(k_n)}\right) y_{t,i}}{\sum_{t=1}^{T} \gamma_t^{(k_n)}}, \tag{37}$$

$$\hat{\sigma}_{n,i}^{(k_n)^2} = \frac{\sum_{t=1}^{T} m_{t,i}^{(k_n)} \eta_{n,i}^{(k_n)} + \left(\gamma_t^{(k_n)} - m_{t,i}^{(k_n)}\right) \varepsilon_{n,i}^{(k_n)}}{\sum_{t=1}^{T} \gamma_t^{(k_n)}}, \tag{38}$$

where

$$\gamma_t^{(k_n)} = \sum_{k_x=1}^{K_x} \gamma_t^{(k_x,k_n)}, \tag{39}$$

$$m_{t,i}^{(k_n)} = \sum_{k_x=1}^{K_x} \gamma_t^{(k_x,k_n)} w_{t,i}^{(k_x,k_n)}, \tag{40}$$

$$\eta_{n,i}^{(k_n)} = \left[\tilde{\sigma}_{n,i}^{(k_n)^2}(y_{t,i}) + \left(\tilde{\mu}_{n,i}^{(k_n)}(y_{t,i}) - \hat{\mu}_{n,i}^{(k_n)}\right)^2\right], \tag{41}$$

$$\varepsilon_{n,i}^{(k_n)} = \left(y_{t,i} - \hat{\mu}_{n,i}^{(k_n)}\right)^2. \tag{42}$$

Similarly to what has been previously discussed for the speech estimates, the masking model imposes the constraint $n_{t,i} \in (-\infty, y_{t,i}]$ when noise is masked by speech. Therefore, $\tilde{\mu}_{n,i}^{(k_n)}(y_{t,i})$ and $\tilde{\sigma}_{n,i}^{(k_n)^2}(y_{t,i})$ in the previous equations are the mean and variance of the estimate obtained when noise is masked by speech. Both quantities are computed using (28) and (29) given the current estimate of the noise model $\mathcal{M}_n$.

As can be seen, the updating equations (37) and (38) for the means and variances of the noise model again involve a weighted average of two different terms: one for the case when noise is masked by speech and vice versa. The weights of the average are $m_{t,i}^{(k_n)}$ and $(\gamma_t^{(k_n)} - m_{t,i}^{(k_n)})$ that play the role of a missing-data mask and a complementary mask, respectively, for the Gaussian component k_n. In particular, as can be seen from (40), $m_{t,i}^{(k_n)}$ is the proportion of the evidence of $y_{t,i}$ being masked by speech that can be explained by the k_n-th component.

Equations (36)-(38) form the basis of the iterative procedure for fitting a GMM to the noise distribution in each utterance. In each iteration the parameters of the GMM estimated in the previous iteration, $\mathcal{M}_n$, are used to compute the sufficient statistics required for updating those parameters, thus

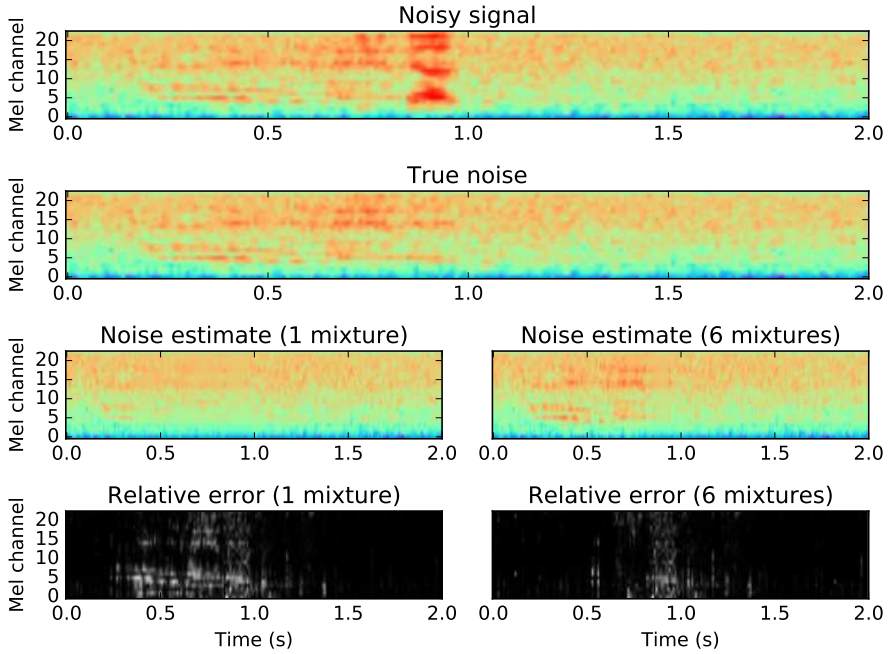

Fig. 4 Example noise estimates computed by the MMSR technique in (30) using the GMMs obtained by the proposed noise estimation algorithm. *[Top]* Sentence "He doesn't" from the Aurora-4 database distorted by street noise at 8 dB. *[2nd row]* True noise spectrogram computed from the clean and noisy signal available in the database. *[3rd row]* Noise estimates obtained using 1-mixture (left) and 6-mixture (right) noise models. *[Bottom]* Relative errors of the estimates w.r.t. the true noise signal.

yielding the updated model $\hat{\mathcal{M}}_n$. In this work, the parameters of the initial GMM are found by fitting a GMM to the first and last frames of the utterance (i.e. we assume that these segments correspond to silence). Finally, the equations (36)-(38) are applied until a certain stopping criterion is met (e.g. a number of iterations is reached).

To illustrate the proposed algorithm, Fig. 4 shows example Log-Mel spectrograms of the noise estimates obtained using 1-mixture and 6-mixture GMMs. To obtain the noise estimates from the noise models, a similar procedure to that described in Section 3.2 for computing the speech estimates is used. That is, we use the MMSR technique in (30) for computing the noise estimates, but now the models $\mathcal{M}_x$ and $\mathcal{M}_n$ play opposite roles. From the comparison with the true noise spectrum, it can be seen that more accurate noise estimates are obtained using the 6-mixture GMM because it offers more flexibility for modelling less stationary noises (e.g. from seconds 0.5 to 1.0). In the example, an average root mean square error (RMSE) of 0.093 is achieved with the single mixture GMM while 0.090 is achieved with the 6-mixture GMM.

5 Comparison with other missing-data techniques

The MMSR and noise model estimation techniques presented in the previous sections share some similarities with other techniques developed within the missing data (MD) paradigm to noise-robust ASR. In this section we briefly review several well-known MD techniques and highlight the similarities and differences with our proposals.

Missing-data techniques reformulate the problem of enhancing noisy speech as a missing data problem [7, 35]. This alternative formulation appears naturally as a result of expressing the spectral features in a compressed domain and adopting the masking model in (7) for modelling the effects of noise in speech. Contrary to MMSR, MD techniques tend to make very little assumptions about the corrupting noise. Thus, instead of estimating the noise in each utterance as we do in here, MD techniques assume that a mask is available *a priori* identifying the reliable and unreliable time-frequency bins of the noisy spectrum. The masks can be binary, but soft masks are generally preferred since they are known to provide better reconstruction performance [5]. It must be pointed out, however, that although MD techniques make no assumptions about the noise, in practice the missing data masks are usually obtained from noise estimates. Thus, in a way or other, both approaches, MMSR and MD techniques, require the noise to be estimated. In this sense, we see the joint conception of the noise-robustness problem developed in this paper as an advantage compared to traditional MD techniques.

There are two alternative MD approaches to perform speech recognition in the presence of missing data. The first approach is known as the *marginalisation* approach and, in brief, it basically involves modifying the computation of the observation probabilities in the recogniser to take into account the missing information [7,8]. The second approach, known as *imputation*, involves "filling in" the missing information in the noisy spectrum before speech recognition actually happens [16, 20, 36, 37, 42]. For MD imputation techniques, the estimate of the missing speech features is obtained as follows (see [17,36] for more details),

$$\hat{x}_i = m_i y_i + (1 - m_i) \sum_{k_x=1}^{K_x} P(k_x|\boldsymbol{y}) \tilde{\mu}_{x,i}^{(k_x)}(y_i), \tag{43}$$

where m_i represents the value of the missing-data mask (either binary or soft) for the i-th element of the noisy spectrum.

We can see that there is a clear parallelism between the MD imputation technique in (43) and the MMSR algorithm in (30). First, both techniques involve a linear combination of the observed feature y_i (case of speech masking noise) and an speech estimate for the case of noise masking speech. Second, the weights of the linear combination depend on the reliability of the observation captured by the missing-data mask m_i. Nevertheless, a notable advantage of MMSR compared to the MD techniques is that it requires no prior information about the reliability of the elements of the noisy spectrum, as the soft-mask m_i appears naturally as a by-product of the estimation process. In fact, as we

will see later on Section 6, the soft masks obtained by MMSR in (31) can be directly used to perform MD imputation.

Another interesting MD approach for performing speech recognition in the presence of other interfering sources is the speech fragment decoder (SFD) of [4]. Unlike the above mentioned marginalisation method, the SFD technique carries out both mask estimation and speech recognition at the same time by searching for the optimal segregation mask and HMM state sequence given a set of time-frequency fragments identified prior to the decoding stage. These fragments correspond to patches in the noisy spectrum that are dominated by the energy of an acoustic source [28]. Thus, the SFD approach determines the most likely set of speech fragments among all the possible combinations of source fragments by exploiting knowledge of the speech source provided by the speech models in the recogniser. We can see that the way the SFD proceeds is somehow similar to our MMSR proposal. However, there are some differences between both approaches. First, SFD is an extended decoding algorithm in the presence of other interfering acoustic sources, while MMSR is a feature compensation technique. Second, the way missing-data masks are estimated in both approaches differs. In SFD, mask estimation is obtained as a by-product of the extended search among all the possible fragments. In MMSR, the source models (i.e. speech and noise models) are used to obtain the most likely segmentation of the observed noisy spectrum. Finally, the requirements of both techniques are different: SFD requires a clean speech model and an *a priori* segmentation of the noisy spectrum in terms of source fragments, while our proposal only requires models for the speech and noise sources.

6 Experimental results

To evaluate the proposed methods, we employed two metrics in this paper. Firstly, we computed the root mean square error (RMSE) between the estimated enhanced speech signals and the corresponding clean ones. Similarly, for noise estimation, the RMSE measure was computed between the estimated noise log-Mel spectrum and the true noise estimated from clean and noisy speech signals. Since lower RMSE values might not necessarily imply better ASR performance, we also conducted a second evaluation using speech recognition experiments on noisy speech data.

For both evaluations we used the Aurora-2 [23] and Aurora-4 [22] 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 different test sets (set A, B, and C) are defined for testing. Every set is 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 in set C are also filtered using a different channel response. Because in this work we only address the distortion caused by additive noise, we only evaluated our techniques on sets A and B. Aurora-4 on the other hand is a medium-large vocabulary

database which is based on the Wall Street Journal (WSJ0) 5000-word recognition task. Fourteen 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. We only evaluated our techniques on the sets T-01 to T-07 with no convolutive distortion.

In this work the acoustic features used by the recogniser were extracted by the ETSI standard front-end [13], which consisted of 12 Mel-frequency cepstral coefficients (MFCCs) along with the 0th order coefficient and their respective velocity and acceleration parameters. Spectral reconstruction, however, was implemented in the log-Mel domain. Thus, the 23 outputs of the log-Mel filterbank were first processed by the spectral reconstruction technique before the discrete cosine transform (DCT) was applied to the enhanced features to obtain the final MFCC parameters. Cepstral mean normalisation (CMN) was applied as a final step in the feature extraction pipeline to improve the robustness of the system to channel mismatches.

The acoustic models of the recogniser were trained on clean speech using the baseline scripts provided with each database. In particular, left to right continuous density HMMs with 16 states and 3 Gaussians per state were used in Aurora-2 to model each digit. Silences and short pauses were modelled 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 were used. The language model used in Aurora-4 is the standard bigram for the WSJ0 task.

Besides MMSR, the MD imputation (MDI) technique described in Section 5 was also considered for comparison purposes. MDI was evaluated using oracle binary masks (Oracle), which allow us to determine the reconstruction performance using ideal knowledge of noise masking, and three types of estimated masks: estimated binary masks (Binary), soft masks computed by the MMSR technique in (31) (Soft MMSR), and soft masks obtained by applying a sigmoid compression to SNR estimates, as proposed in [5] (Soft Sigmoid). In all cases except the Soft MMSR masks, the masks were derived from the SNR values estimated for each time-frequency element of the noisy spectrum. For the Oracle masks, the true noise was used to compute the SNR values and, then, a 7 dB threshold was employed to binarise the values in order to obtain the final oracle mask. For the Binary and Soft Sigmoid masks, the noise estimates described below were employed to estimate the SNR for each time-frequency element. Then, the SNR values were thresholded (Binary masks) or compressed using a sigmoid function (Soft Sigmoid). In both cases the parameters used to estimate the masks from the SNR values (i.e. binary threshold and sigmoid function parameters) were empirically optimised for each database using a development set.

The spectral reconstruction techniques were initially evaluated using estimated noise rather than using our proposed algorithm of Section 4. In this case, noise was estimated as follows. For each frame, a noise estimate was obtained by linear interpolation of two initial noise estimates computed independently by averaging the N first and N last frames of each utterance ($N = 20$ for Aurora-2 and $N = 40$ for Aurora-4). The noise estimates were then post-processed to ensure they do not exceed the magnitude of the observed noisy speech, as this would violate the masking model. For those techniques that require the noise covariance (e.g. MMSR), a fixed, diagonal-covariance matrix was estimated also from the N first and last frames. Thus, when using noise estimates in MMSR, the noise model corresponds to a single, time-dependent Gaussian whose mean at each frame is the noise estimate for the frame.

For spectral reconstruction a 256-component GMM with diagonal covariance matrices was used in all the cases as prior speech model. The GMM was estimated using the EM algorithm from the same clean training dataset used for training the acoustic models of the recogniser.

6.1 Performance of the spectral reconstruction methods

Tables 1 and 2 show the average RMSE values obtained by the feature enhancement techniques on the Aurora-2 and Aurora-4 databases, respectively. For Aurora-2, the results are given for each SNR value and are computed over test sets A and B. Also, the overall average (Avg.) between 0 dB and 20 dB is also shown, as it is common practice for Aurora-2. For Aurora-4, the results for test sets T-01 to T-07 and the average RMSE value over all sets are reported. For comparison purposes, the RMSE results directly computed from the noisy signals with no compensation are also shown (Baseline).

It is clear from both tables that all the spectral reconstruction methods significantly improve the quality of the noisy signals, particularly at low SNR levels (e.g. 0 and -5 dB in Table 1). It can also be observed that the average RMSE results obtained by these methods are significantly lower on Aurora-4 than on Aurora-2, owing this to the lower average SNR on Aurora-2 compared to Aurora-4. As expected, the best results (lower RMSE values) are obtained by MDI-Oracle, which uses oracle masks. Although oracle masks are not usually available in real-word conditions, it is interesting to analyse the results of this technique since they are indicative of the upper bound performance that can be expected from the enhancement techniques derived from the masking model. For example, it can be seen in Table 1 that the performance of this technique consistently decrease between the clean and -5 dB conditions. In the latter condition, it is more difficult to estimate accurately the clean speech energy in the spectral regions masked by noise because there is less reliable evidence (i.e. less reliable speech features) for missing-data imputation.

When estimated masks are used, MDI with Binary masks is significantly worse than the rest of the methods (paired t-test with $p < 0.05$). The reason could be that this method is less robust to noise estimation errors due to the

Method		Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg.
Baseline		0.00	1.01	1.32	1.67	2.06	2.50	2.97	1.71
MDI	Oracle	0.00	0.37	0.48	0.61	0.75	0.91	1.10	0.62
	Binary	0.19	0.79	0.97	1.19	1.51	2.01	2.38	1.29
	Soft MMSR	0.06	0.58	0.75	0.93	1.15	1.45	1.83	0.97
	Soft Sigmoid	0.12	0.59	0.75	0.94	1.16	1.43	1.76	0.97
MMSR		0.06	0.58	0.74	0.92	1.12	1.38	1.69	**0.95**

Table 1 RMSE values obtained by the proposed MMSR technique and other similar feature enhancement methods on the Aurora-2 database.

Method		T-01	T-02	T-03	T-04	T-05	T-06	T-07	Avg.
Baseline		0.00	1.14	1.51	1.50	1.62	1.41	1.64	1.26
MDI	Oracle	0.00	0.38	0.53	0.52	0.57	0.47	0.57	0.43
	Binary	0.07	0.62	1.05	1.08	1.07	0.97	1.03	0.84
	Soft MMSR	0.12	0.57	0.95	1.01	0.94	0.93	0.91	0.78
	Soft Sigmoid	0.13	0.59	0.91	0.96	0.96	0.86	0.93	**0.76**
MMSR		0.08	0.58	0.95	1.01	0.98	0.90	0.93	0.78

Table 2 RMSE values obtained by the proposed MMSR technique and other similar feature enhancement methods on the Aurora-4 database.

hard decisions made when computing the binary masks from SNR estimates. Nevertheless, important gains are observed for MDI-Binary over the baseline, particularly at low and medium SNRs. There is no significant differences (at the 95 % confidence level) between both types of soft masks (Soft MMSR and Soft Sigmoid) on Aurora-2. On Aurora-4, on the other hand, MDI with Soft Sigmoid masks achieves slightly better results than MDI with Soft MMSR masks due to the sigmoid function parameters being empirically optimised for this database using adaptation sets. However, the MMSR technique has the advantage of requiring no such parameter tuning. Likewise, our MMSR technique is significantly better ($p < 0.05$) than the rest of the techniques except MDI-Oracle on the Aurora-2 database, being the differences particularly noticeable at the medium-low SNR levels. On Aurora-4, however, MDI with Soft Sigmoid masks is slightly superior to MMSR due to, again, the sigmoid function parameters being empirically optimised for Aurora-4.

We also conducted a series of speech recognition experiments on noisy data as a complementary evaluation for the spectral reconstruction techniques. The average word accuracy results (WAcc) are given in Table 3 for Aurora-2 and in Table 4 for Aurora-4. For both databases the relative improvement (R.I.) with respect to the baseline system is also provided. For comparison purposes, the recognition results obtained by the ETSI advanced front-end (ETSI AFE) [14], which is especially designed for noise robustness, are also shown. One of the first things we can observe is that despite the RMSE values shown in Tables 1 and 2 are better for Aurora-4, the recognition accuracies are significantly higher in Aurora-2 than in Aurora-4. This is not surprising given that the speech task in Aurora-4 is much more difficult than in Aurora-2: medium-large vocabulary vs. connected-digit recognition.

Method		Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg.	R.I.
Baseline		99.10	97.40	92.84	76.18	43.44	22.87	12.94	66.54	–
ETSI AFE		99.24	98.24	96.87	93.38	84.43	60.98	27.25	86.78	30.42
MDI	Oracle	99.10	99.04	98.84	98.17	96.39	90.73	74.06	96.63	45.22
	Binary	98.88	97.60	95.48	90.66	78.83	54.75	24.24	83.46	25.43
	Soft MMSR	98.89	98.04	96.74	92.81	81.92	57.96	27.39	85.49	28.48
	Soft Sigmoid	98.87	98.16	96.77	92.82	81.54	58.08	27.70	85.47	28.45
MMSR		98.88	98.22	97.06	93.72	84.09	61.54	28.56	**86.93**	30.64

Table 3 Performance results in terms of WAcc (%) of the MMSR method on the Aurora-2 database and comparison with other similar compensation techniques.

Again, the best results on both databases are obtained by MDI-Oracle. Although not realistic, it is remarkable that the performance achieved by this method on noisy data is close to clean conditions for medium-high SNRs. For the rest of the techniques, the same performance pattern as for the RMSE results is observed. MDI-Binary is the worst method on both databases while MMSR is the best method on Aurora-2 and, on Aurora-4, MDI with Soft sigmoid masks is the best method but followed very closely by MMSR. For Aurora-4 it is especially noteworthy the performance gap between MDI-Binary and the rest of the techniques. For example, the relative improvement of MMSR over MDI-Binary is 10.90%. It is therefore interesting to analyze why MDI-Binary is much more fragile than the rest of the methods in low SNR conditions. In those conditions it is frequent that the estimated masks contains errors: either some unreliable features are labelled as reliable or viceversa. In the first case, the unreliable features are not treated by the imputation method. The second case is even more problematic, since the reliable features will be labelled as unreliable and will be replaced by the imputation method, hence, further degrading the observed signal. Therefore, in the light of the results, it is clear that a better strategy in those cases is to adopt a soft-decision approach, such as those of MDI with soft masks or our method MMSR. In particular, the approach adopted by our proposal MMSR, where a noise distribution is assumed for the noise instead of just using noise estimated, seems to be more suited for this high noise conditions.

Finally, another interesting result from Tables 3 and 4 is the good performance achieved by MMSR in comparison with ETSI AFE, despite that this front-end includes several complex noise-reduction blocks (e.g. blind equalisation) not implemented in our technique.

6.2 Performance of the noise model estimation algorithm

In this section we evaluate the performance of the noise model estimation algorithm proposed in Section 4. In the first evaluation, we computed the RMSE values between the noise estimates derived from the GMMs obtained by this algorithm and the true noise signals computed from the clean and noisy recordings in the speech databases. To obtain the noise estimates, we employed

Method		T-01	T-02	T-03	T-04	T-05	T-06	T-07	Avg.	R.I.
Baseline		87.69	75.30	53.24	53.15	46.80	56.36	45.38	59.70	–
ETSI AFE		88.25	81.41	69.14	64.80	67.44	66.34	68.78	72.31	21.12
MDI	Oracle	87.69	86.74	84.46	84.44	83.19	85.90	82.38	84.97	42.33
	Binary	86.96	80.78	58.47	52.74	59.63	56.14	61.42	65.16	9.15
	Soft MMSR	87.52	83.65	66.62	63.78	63.48	69.19	65.31	71.36	19.53
	Soft Sigmoid	87.22	83.95	69.76	65.31	67.01	69.42	68.19	**72.98**	22.24
MMSR		87.54	83.28	69.23	64.49	64.88	70.63	66.93	72.43	21.32

Table 4 Performance results in terms of WAcc (%) of the MMSR method on the Aurora-4 database and comparison with other similar compensation techniques.

Noise model		Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg.
Interpolated		3.82	0.40	0.39	0.37	0.36	0.35	0.33	0.37
GMM	1 Gauss.	3.18	0.35	0.33	0.31	0.29	0.27	0.24	0.31
	2 Gauss.	3.52	0.35	0.32	0.30	0.28	0.26	0.22	**0.30**
	4 Gauss.	3.62	0.36	0.33	0.31	0.29	0.26	0.22	0.31
	6 Gauss.	3.63	0.36	0.34	0.32	0.30	0.27	0.22	0.32
	8 Gauss.	3.62	0.37	0.35	0.34	0.31	0.28	0.22	0.33

Table 5 Evaluation of the performance of different noise estimation methods on the Aurora-2 database. RMSE values computed between the estimated noise spectrum and the true noise spectrum, both expressed in the log-Mel domain, are reported.

the MMSR technique in (30), but instead of using it for speech enhancement as we have been doing so far, we applied it to predict the noise energy in the spectral regions masked by speech.

The RMSE results on the Aurora-2 database are shown in Table 5. The results are presented as a function of the number of Gaussians in the noise model: 1, 2, 4, 6, and 8 components. Also, for comparison purposes, the RMSE values for the noise estimates obtained by linear interpolation are also shown (Interpolated). From the results it is clear that the noise estimates obtained by the proposed algorithm are significantly better (with a 95 % confidence level) than those obtained by the interpolation method. Interestingly, the RMSE values are better for the lower SNR levels than for the higher SNRs. The reason of this apparent contradiction is that the true noise level for the higher SNRs (i.e. clean or 20 dB) is zero or close to zero, while the estimate of our method correspond to the energy level of the silences, which is not completely zero. As can be seen, the best overall results are achieved when 2-mixture GMMs are used. From the results, it seems that using more than 2 mixture components degrades the performance because the GMMs are poorly trained due to their high number of parameters. On the other hand, single-Gaussian GMMs are unable to properly model non-stationary noises.

Fig. 5 shows the RMSE values for the GMM-based noise estimates on the different noise types in Aurora-2. Again, 2-mixture GMMs achieve the best results for most of the noise types. Nevertheless, the 1-mixture model is significantly better than the other models on the most stationary noises (car and exhibition). For less stationary noises (e.g. subway or street), however,

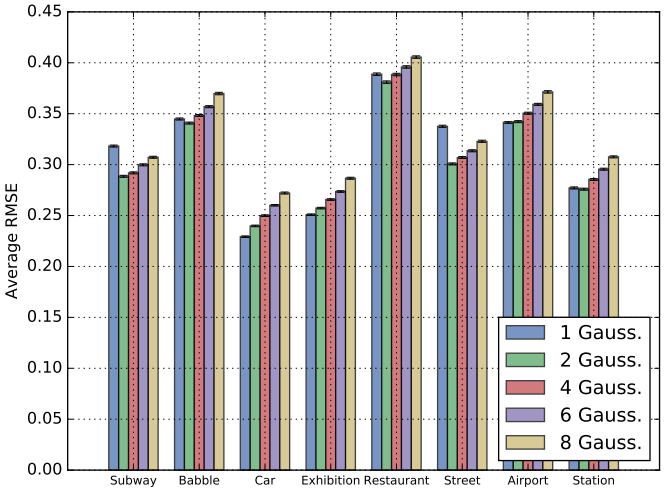

Fig. 5 Average RMSE results between 0 dB and 20 dB for the GMM-based noise estimates on the different noise types in the Aurora-2 database. Error bars are plotted at the 95% confidence level.

Method		T-01	T-02	T-03	T-04	T-05	T-06	T-07	Avg.
Interpolated		4.21	0.32	0.41	0.47	0.36	0.46	0.34	0.94
GMM	1 Gauss.	4.00	0.27	0.31	0.38	0.30	0.36	0.27	**0.84**
	2 Gauss.	4.17	0.28	0.30	0.37	0.27	0.36	0.25	0.86
	4 Gauss.	4.20	0.29	0.31	0.37	0.27	0.36	0.26	0.87
	6 Gauss.	4.20	0.29	0.32	0.37	0.28	0.36	0.27	0.87
	8 Gauss.	4.19	0.29	0.32	0.37	0.28	0.36	0.27	0.87
	10 Gauss.	4.19	0.30	0.32	0.37	0.28	0.36	0.27	0.87

Table 6 Evaluation of the performance of different noise estimation methods in terms of noise RMSE values on the Aurora-4 database.

the performance of the single-Gaussian models is significantly lower than that achieved by the 2-mixture GMMs.

The RMSE values obtained by the noise estimation methods on Aurora-4 are presented in Table 6. As in Aurora-2, the GMM-based method yields significantly better estimates than the interpolation-based method on all the test sets. Surprisingly, the best results on this database are achieved by using single-mixture GMMs. Fig. 6 shows the detailed results per noise type, providing more insight about this result. As can be observed, the single-mixture model only outperforms the other models on the car noise (T-02) and the clean condition (T-01). For the rest of the noise types, 2-mixture models achieve significantly better results than 1-mixture models, especially on the less stationary noises such as restaurant, street or train station. Because the clean condition is not usually problematic for speech enhancement purposes, we can neglect it when computing the average RMSE results. Thus, the average RMSE values for test sets T-02 to T-07 for GMMs with 1 to 10 Gaussians are, respectively, 0.32, 0.31, 0.31, 0.32, 0.32, and 0.32. Now it is clear that 2-mixture and

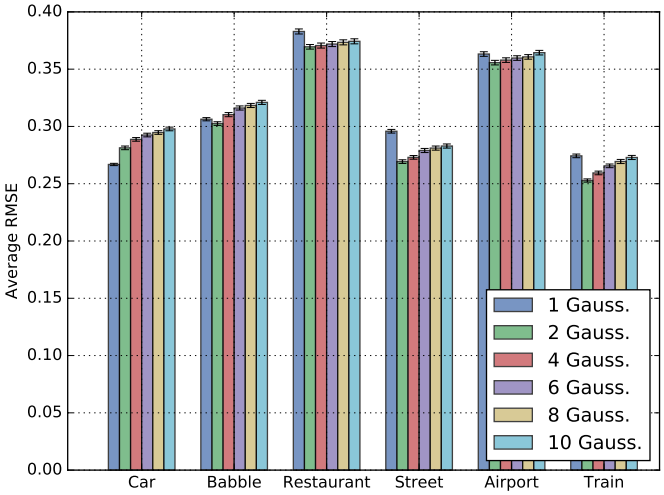

Fig. 6 RMSE values for the GMM-based noise estimates on the different noise types in the Aurora-4 database. Confidence intervals are plotted at the 95% level.

Noise model	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg.
Interpolated	0.06	0.58	0.74	0.92	1.12	1.38	1.69	0.95
GMM 1 Gauss.	0.03	0.57	0.73	0.91	1.12	1.40	1.73	0.95
2 Gauss.	0.06	0.56	0.71	0.88	1.11	1.41	1.74	**0.93**
4 Gauss.	0.06	0.56	0.72	0.89	1.13	1.46	1.79	0.95
6 Gauss.	0.06	0.57	0.73	0.91	1.17	1.51	1.84	0.98
8 Gauss.	0.06	0.58	0.74	0.94	1.21	1.57	1.88	1.01

Table 7 Objective evaluation of the proposed MMSR technique for speech feature enhancement in terms of speech RMSE values on the Aurora-2 database.

Method	T-01	T-02	T-03	T-04	T-05	T-06	T-07	Avg.
Interpolated	0.08	0.58	0.95	1.01	0.98	0.90	0.93	0.78
GMM 1 Gauss.	0.06	0.56	0.86	0.95	0.94	0.83	0.90	0.73
2 Gauss.	0.08	0.56	0.84	0.93	0.88	0.81	0.86	0.71
4 Gauss.	0.09	0.56	0.85	0.92	0.88	0.80	0.86	**0.71**
6 Gauss.	0.09	0.56	0.85	0.92	0.89	0.80	0.87	0.71
8 Gauss.	0.09	0.57	0.86	0.92	0.89	0.80	0.88	0.72
10 Gauss.	0.09	0.57	0.86	0.92	0.89	0.81	0.89	0.72

Table 8 Objective evaluation of the proposed MMSR technique for speech feature enhancement in terms of speech RMSE values on the Aurora-4 database.

4-mixture GMMs are, on average, better than the rest of the models on the noisy conditions. Disappointingly, we can see that increasing the number of Gaussians in the models do not necessarily produce better results. It might be that the GMMs with many components are not robustly trained due to their high number of parameters. Also, it could be that the speech model imposes only weak constraints during the EM algorithm and, as a result, the noise model ends up modelling some parts of the speech spectrum.

Noise model	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg.
Interpolated	98.88	98.22	97.06	93.72	84.09	61.54	28.56	86.93
GMM 1 Gauss.	99.10	98.42	97.09	93.37	82.81	58.07	25.12	85.95
2 Gauss.	99.08	98.29	97.17	94.06	84.91	61.48	27.75	**87.18**
4 Gauss.	99.03	98.27	97.02	94.02	84.64	61.79	28.70	87.15
6 Gauss.	99.07	98.38	96.98	93.96	84.13	60.03	27.86	86.70
8 Gauss.	99.04	98.32	97.13	93.92	83.80	58.55	26.86	86.34

Table 9 Comparison of the MMSR performance (in terms of WAcc) on Aurora-2 using either interpolated noise estimates or GMM noise models estimated with the iterative algorithm proposed in Section 4.

Method	T-01	T-02	T-03	T-04	T-05	T-06	T-07	Avg.
Interpolated	87.54	83.28	69.23	64.49	64.88	70.63	66.93	72.43
GMM 1 Gauss.	87.60	83.22	68.52	64.21	62.86	70.47	64.34	71.60
2 Gauss.	87.58	83.36	68.78	64.34	64.75	70.54	67.21	72.37
4 Gauss.	87.67	83.22	69.62	63.95	65.78	70.48	68.41	**72.73**
6 Gauss.	87.54	83.45	69.23	64.23	65.25	69.33	67.98	72.43
8 Gauss.	87.52	82.57	69.08	64.45	65.20	68.60	66.84	72.04
10 Gauss.	87.60	82.65	67.79	64.58	65.01	68.69	66.19	71.79

Table 10 Comparison of the MMSR performance (in terms of WAcc) on Aurora-4 using either interpolated noise estimates or GMM noise models estimated with the proposed iterative algorithm.

Next, we evaluated the performance of the proposed algorithm for noise model estimation on speech feature enhancement. To do so, we computed the RMSE values between the speech estimates obtained by the MMSR technique and the corresponding clean signals. To obtain the speech estimates, the MMSR technique directly used the noise GMMs estimated by the algorithm instead of the noise estimates derived from the GMMs. The RMSE values on the Aurora-2 and Aurora-4 databases are summarized in Tables 7 and 8, respectively. It is worth noting that the Interpolated results are the same than those reported in Tables 1 and 2 for the MMSR technique. As can be seen, the MMSR technique greatly benefits from using GMMs for modelling the noise distribution. Thus, the results for the GMM-based noise models are significantly better than those for the Interpolated noise estimates except when an excessive number of Gaussian components are employed (i.e. 6 and 8 Gaussians in Aurora-2). On the Aurora-4 database, on the other hand, the GMM-based enhancement method always outperforms the results achieved by the Interpolated system. Despite the worst results on noise estimation were obtained for the clean condition (Clean in Aurora-2 and T-01 in Aurora-4), it can be seen that this does not affect spectral reconstruction performance and the RMSE values for this condition are always close to zero.

Finally, we also carried out a series of speech recognition experiments using the signals enhanced by MMSR. The word accuracy results are shown in Tables 9 and 10 for Aurora-2 and Aurora-4, respectively. As expected, significant improvements over both the baseline results reported in Tables 1 and 2 and the MMSR technique using Interpolated noise are achieved when MMSR

employs GMMs for modelling the noise distribution. As can be observed, the results obtained for Aurora-2 are in consonance with the RMSE values shown in Table 7, where 2-mixture GMMs for noise modelling yield the best performance. Although something similar happens in Aurora-4 (the best results are achieved by using GMMs with 4 mixtures), it is worth to note that improvements in speech recognition are smaller than those reported Table 8 for speech enhancement. This can be justified by the fact ASR is already able to deal with certain amount of noise (mismatch) in the speech signals.

7 Summary and Discussion

In order to make ASR systems usable in real-life conditions, they must be equipped with noise-resistant mechanisms to protect them from the degradation caused by the environmental noise that is often present in these conditions. In this paper we have introduced one of such mechanisms based on an analytical model which describes how speech is distorted by noise in the feature domain. Under this model, which has been referred to as the masking model, noise corruption translates to an effective masking of the speech spectrum. Thus, two different problems must be solved to compensate for the noise distortion: mask estimation, which involves the segmentation of the noisy spectrum into reliable and unreliable regions, and spectral reconstruction, which involves estimating speech in the unreliable regions.

We have shown that the above two problems can be jointly addressed using the proposed MMSE-based spectral reconstruction algorithm derived from the masking model. Unlike other similar techniques, the proposed algorithm needs no prior segmentation of the noisy spectrum (i.e. a mask). Instead, the segmentation that best explains the noisy spectrum is automatically estimated on the basis of available prior models for the speech and noise features. In other words, the masks are estimated using a top-down approach under the constraints imposed by the *a priori* speech and noise models.

To estimate the noise models required by the spectral reconstruction technique, we have also proposed a novel algorithm based on the EM method. The proposed algorithm finds the noise model parameters (represented as a GMM) by iteratively optimising the likelihood of the observed noisy data under the constraints imposed by the clean speech model. This ensures the algorithm convergence and, hence, that the parameters of the noise model properly represent the noise distribution in the utterance.

The proposed techniques were evaluated on the Aurora-2 digit recognition task and on the Aurora-4 large vocabulary task. In both cases, the proposed feature enhancement technique achieves significant relative improvements of 29.49% (for Aurora-2) and 25.72% (for Aurora-4) over the baseline. Furthermore, the results also show that the proposed MMSR method outperforms comparable MD imputation techniques in most cases. In particular, it is shown that the proposed method tends to be more robust than MD imputation to errors in noise estimation, especially at medium and low SNR levels. Regard-

ing the performance of the proposed noise model estimation algorithm, the results show that significant improvements on speech enhancement and ASR are achieved by this method over using simple noise estimates. It is also found that increasing the number of Gaussian components in the noise model is particularly beneficial when modelling non-stationary noises.

This work has several interesting directions for future research. First, the perceptual interpretation of the masking model suggests that it could be enhanced by incorporating auditory features. Second, future work will aim to reduce the performance gap between our proposal and MDI using oracle masks. In this regard, MMSR could be extended to also combat the degradation caused by convolutive noise. Third, instead of using GMMs as prior speech models, the proposed algorithms could be extended to exploit the HMMs used by the recogniser as this would provide with additional temporal constraints for further improvement. Another topic to be investigated is the automatic determination of the number of mixture components when estimating the noise model. Finally, inspired by the speech fragment decoding approach in [4], an additional source information that could be exploited by the proposed algorithms for further improvement are the source constraints derived from the noisy signal itself. These constraints translate to bottom-up processes for segmenting the noisy signal into time-frequency fragments dominated by the energy of a particular sound source. Information that could be used to guide the fragment generation process is the pitch of the different sources and the common onsets or offsets. These bottom-up constraints, combined with the top-down constrains imposed by the speech and noise models, could be helpful for estimating the missing-data masks.

Acknowledgements

This work was supported by the Spanish MINECO (Ministerio de Economía y Competitividad)/FEDER Project TEC2013-46690-P.

References

1. Acero, A., Deng, L., Kristjansson, T., Zhang, J.: HMM adaptation using vector Taylor series for noisy speech recognition. In: Proc. ICSLP, pp. 229–232 (2000)
2. Baker, J.M., Deng, L., Glass, J., Khudanpur, S., Lee, C.H., Morgan, N., O'Shaughnessy, D.: Research developments and directions in speech recognition and understanding, Part 1. IEEE Signal Process. Mag. **26**(3), 75–80 (2009)
3. Baker, J.M., Deng, L., Khudanpur, S., Lee, C.H., Glass, J., Morgan, N., O'Shaughnessy, D.: Updated MINDS report on speech recognition and understanding, Part 2. IEEE Signal Process. Mag. **26**(4), 78–85 (2009)
4. Barker, J., Cooke, M., Ellis, D.P.W.: Decoding speech in the presence of other sources. Speech Commun. **45**(1), 5–25 (2005)
5. Barker, J., Josifovski, L., Cooke, M.P., Green, P.D.: Soft decisions in missing data techniques for robust automatic speech recognition. In: Proc. ICSLP (2000)
6. Cerisara, C., Demange, S., Haton, J.P.: On noise masking for automatic missing data speech recognition: A survey and discussion. Comput. Speech Lang. **21**(3), 443–457 (2007)

7. Cooke, M., Green, P.D., Josifovski, L., Vizinho, A.: Robust automatic speech recognition with missing and unreliable acoustic data. Speech Commun. **34**(3), 267–285 (2001)
8. Cooke, M., Morris, A., Green, P.D.: Missing data techniques for robust speech recognition. In: Proc. ICASSP, pp. 863–866 (1997)
9. Delcroix, M., Kinoshita, K., Nakatani, T., Araki, S., Ogawa, A., Hori, T., Watanabe, S., Fujimoto, M., Yoshioka, T., Oba, T., et al.: Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation. In: Proc. 1st Int. Workshop on Machine Listening in Multisource Environments (CHiME), pp. 12–17 (2011)
10. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society **39**(1), 1–38 (1977)
11. Deng, L., Droppo, J., Acero, A.: Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise. IEEE Trans. Speech Audio Process. **12**(2), 133–143 (2004)
12. Dhrymes, P.J.: Moments of truncated (normal) distributions (2005)
13. ETSI: ETSI ES 201 108 - distributed speech recognition; front-end feature extraction algorithm; compression algorithms (2003)
14. ETSI: ETSI ES 202 050 - distributed speech recognition; advanced front-end feature extraction algorithm; compression algorithms (2007)
15. Faubel, F., McDonough, J., Klakow, D.: A phase-averaged model for the relationship between noisy speech, clean speech and noise in the log-Mel domain. In: Proc. Interspeech, pp. 553–556 (2008)
16. Faubel, F., McDonough, J., Klakow, D.: Bounded conditional mean imputation with Gaussian mixture models: A reconstruction approach to partly occluded features. In: Proc. ICASSP, pp. 3869–3872 (2009)
17. Faubel, F., Raja, H., McDonough, J., Klakow, D.: Particle filter based soft-mask estimation for missing feature reconstruction. In: Proc. IWAENC (2008)
18. González, J.A., Peinado, A.M., Gómez, A.M.: MMSE feature reconstruction based on an occlusion model for robust ASR. In: Advances in speech and language technologies for Iberian languages - IberSPEECH 2012, Communications in Computer and Information Science, pp. 217–226. Springer (2012)
19. González, J.A., Peinado, A.M., Gómez, A.M., Ma, N.: Log-spectral feature reconstruction based on an occlusion model for noise robust speech recognition. In: Proc. Interspeech, pp. 2630–2633 (2012)
20. González, J.A., Peinado, A.M., Ma, N., Gómez, A.M., Barker, J.: MMSE-based missing-feature reconstruction with temporal modeling for robust speech recognition. IEEE Trans. Audio Speech Lang. Process. **21**(3), 624–635 (2013)
21. Hendriks, R.C., Heusdens, R., Jensen, J.: MMSE based noise PSD tracking with low complexity. In: Proc. ICASSP, pp. 4266–4269 (2010)
22. Hirsch, H.G.: Experimental framework for the performance evaluation of speech recognition front-ends of large vocabulary task. Tech. rep., STQ AURORA DSR Working Group (2002)
23. Hirsch, H.G., Pearce, D.: The Aurora experimental framework for the performance evaluations of speech recognitions systems under noise conditions. In: Proc. ISCA ITRW ASR 2000, pp. 181–188 (2000)
24. Leutnant, V., Haeb-Umbach, R.: An analytic derivation of a phase-sensitive observation model for noise robust speech recognition. In: Proc. Interspeech, pp. 2395–2398 (2009)
25. Li, J., Deng, L., Gong, Y., Haeb-Umbach, R.: An overview of noise-robust automatic speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(4), 745–777 (2014)
26. Li, J., Deng, L., Haeb-Umbach, R., Gong, Y.: Robust Automatic Speech Recognition: A Bridge to Practical Applications. Academic Press (2015)
27. Loizou, P.C.: Speech enhancement: Theory and practice. CRC (2007)
28. Ma, N., Green, P., Barker, J., Coy, A.: Exploiting correlogram structure for robust speech recognition with multiple speech sources. Speech Commun. **49**(12), 874–891 (2007)
29. Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**(5), 504–512 (2001)

30. Morales-Cordovilla, J.A., Ma, N., Sánchez, V.E., Carmona, J.L., Peinado, A.M., Barker, J.: A pitch based noise estimation technique for robust speech recognition with missing data. In: Proc. ICASSP, pp. 4808–4811 (2011)
31. Moreno, P.J.: Speech recognition in noisy environments. Ph.D. thesis, Carnegie Mellon University (1996)
32. Nádas, A., Nahamoo, D., Picheny, M.A.: Speech recognition using noise-adaptive prototypes. IEEE Trans. Acoust., Speech, Signal Process. **37**(10), 1495–1503 (1989)
33. Nakatani, T., Yoshioka, T., Araki, S., Delcroix, M., Fujimoto, M.: Logmax observation model with MFCC-based spectral prior for reduction of highly nonstationary ambient noise. In: Proc. ICASSP, pp. 4029–4032 (2012)
34. Radfar, M.H., Banihashemi, A.H., Dansereau, R.M., Sayadiyan, A.: Nonlinear minimum mean square error estimator for mixture-maximisation approximation. Electron. Lett. **42**(12), 724–725 (2006)
35. Raj, B., Seltzer, M.L., Stern, R.M.: Reconstruction of missing features for robust speech recognition. Speech Commun. **48**(4), 275–296 (2004)
36. Raj, B., Singh, R.: Reconstructing spectral vectors with uncertain spectrographic masks for robust speech recognition. In: Proc. ASRU, pp. 65–70 (2005)
37. Raj, B., Stern, R.M.: Missing-feature approaches in speech recognition. IEEE Signal Process. Mag. **22**(5), 101–116 (2005)
38. Ramírez, J., Górriz, J.M., Segura, J.C.: Voice activity detection. fundamentals and speech recognition system robustness. INTECH Open Access Publisher (2007)
39. Ramírez, J., Segura, J.C., Benítez, C., De La Torre, A., Rubio, A.: Efficient voice activity detection algorithms using long-term speech information. Speech Commun. **42**(3), 271–287 (2004)
40. Reddy, A.M., Raj, B.: Soft mask estimation for single channel speaker separation. In: Workshop on Statistical and Perceptual Audio Processing SAPA (2004)
41. Reddy, A.M., Raj, B.: Soft mask methods for single-channel speaker separation. IEEE Trans. Audio Speech Lang. Process. **15**(6), 1766–1776 (2007)
42. Remes, U., Nankaku, Y., Tokuda, K.: GMM-based missing-feature reconstruction on multi-frame windows. In: Proc. Interspeech, pp. 1665–1668 (2011)
43. Rennie, S.J., Hershey, J.R., Olsen, P.A.: Single-channel multitalker speech recognition. IEEE Signal Process. Mag. **27**(6), 66–80 (2010)
44. Roweis, S.T.: Factorial models and refiltering for speech separation and denoising. In: Proc. Eurospeech, pp. 1009–1012 (2003)
45. Segura, J.C., de la Torre, A., Benítez, M.C., Peinado, A.M.: Model-based compensation of the additive noise for continuous speech recognition. Experiments using the Aurora II database and tasks. In: Proc. Eurospeech, pp. 221–224 (2001)
46. Stouten, V., Van Hamme, H., Wambacq, P.: Effect of phase-sensitive environment model and higher order VTS on noisy speech feature enhancement. In: Proc. ICASSP, vol. 1, pp. 433–436 (2005)
47. Varga, A.P., Moore, R.K.: Hidden Markov model decomposition of speech and noise. In: Proc. ICASSP, pp. 845–848 (1990)
48. Virtanen, T., Singh, R., Raj, B. (eds.): Techniques for noise robustness in automatic speech recognition. Wiley-Blackwell (2012)