MMSE feature reconstruction based on an occlusion model for robust ASR

José A. González*, Antonio M. Peinado, and Ángel M. Gómez

Dpt. Teoría de la Señal, Telemática y Comunicaciones,
Centro de Investigación en Tecnologías de la Información y de las Comunicaciones,
18071-Granada, Spain
{joseangl,amp,amgg}@ugr.es
http://tstc.ugr.es,http://citic.ugr.es

Abstract. This paper proposes a novel compensation technique developed in the log-spectral domain. Our proposal consists in a minimum mean square error (MMSE) estimator derived from an occlusion model [1]. According to this model, the effect of noise over speech is simplified to a binary masking, so that the noise is completely masked by the speech when the speech power dominates and the other way round when the noise is dominant. As for many MMSE-based techniques, a statistical model of clean speech is required. A Gaussian mixture model is employed here. The resulting technique has clear similarities with missing-data imputation techniques although, unlike these ones, an explicit model of noise is employed by our proposal. The experimental results show the superiority of our MMSE estimator with respect to missing-data imputation with both binary and soft masks.

Keywords: robust ASR, feature reconstruction, MMSE estimation, occlusion model.

1 Introduction

Automatic speech recognition (ASR) is currently moving toward new ubiquitous and pervasive applications where it allows an efficient and natural way for human-machine interaction. However, these scenarios may reduce the performance of ASR systems due to several reasons. Undoubtedly, an adverse acoustic environment and, in particular, environmental noise, is the main of these reasons. Thus, the robustness of ASR systems against noise is a desirable feature that must be addressed.

In order to reduce the effect of the acoustic noise over speech recognizers there exist multiple approaches, but two of them stand out from others [2]: feature compensation and model adaptation. While the first one tries to *denoise* the speech features employed for recognition, the second one modifies the acoustic model parameters to reduce the mismatch with the noisy input features. The advantages of feature compensation is that it can be developed independently from the recognition engine and, also, that it can be implemented more efficiently than adaptation.

* This work has been supported by the FPU fellowship program from the Spanish Ministry of Education and project MICINN TEC2010-18009.

This paper proposes a novel compensation technique developed in the log-spectral domain. Our proposal consists in a minimum mean square error (MMSE) estimator derived from an occlusion model [1]. According to this model, the effect of noise over speech is simplified to a binary masking, so that the noise is completely masked by the speech when the speech power dominates and the other way round when the noise is dominant. In order to model the clean speech log-spectra, we follow the classical approach based on a Gaussian mixture model (GMM). Section 2 is devoted to present and derive the proposed estimator. We will see that the application of MMSE along with the hard-decision occlusion model will yield a graceful soft-decision estimate which is a linear combination of the observed (noisy) feature vector and an estimate of the clean feature vector for the case of speech totally occluded by noise. The resulting estimator will resemble other techniques derived from a missing-data (MD) framework. The similarities and differences with these techniques are discussed is section 3. Section 4 is devoted to the experimental results. A summary of this work can be found in section 5.

2 MMSE estimation from an occlusion model

2.1 Occlusion model

We will note as $\boldsymbol{y}$ the feature vectors corresponding to the observed (noisy) log-Mel filterbank energies. Also, $\boldsymbol{x}$ and $\boldsymbol{n}$ will represent the same type of spectral features for the clean speech and the noise, respectively. The relationship between these variables is accurately represented by the following model [3],

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

where $\boldsymbol{r}$ is a residual vector that depends on the phase relationship between clean speech and noise.

Although accurate, the above distortion model does not allow an easy derivation of the MMSE estimator that we want to obtain, so some approximations must be introduced. In the case of the occlusion model (OM), the residual $\boldsymbol{r}$ is neglected and, also, the *log-max* approximation (that is, $\log(e^a + e^b) \approx \max(a, b)$) is applied to (1) (see [4] for a detailed derivation of this model). The resulting model can be finally expressed as,

$$\boldsymbol{y} \approx \max(\boldsymbol{x}, \boldsymbol{n}) \tag{2}$$

where operator **max** is applied feature by feature.

This model was first proposed in [1] and involves that some parts of the clean speech spectrogram are completely masked by noise, while others are almost unaffected (noise is masked by speech). Our proposal uses this assumption and the spectral correlations represented by the GMM to provide clean speech feature estimates.

2.2 MMSE estimation based on the occlusion model

MMSE estimation is a Bayesian tool frequently employed in feature compensation techniques. The MMSE estimate of the clean feature vector given the observed (noisy) one can be expressed as,

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

In first place, the MMSE estimator requires a clean feature model λ_X which allows the computation of the posterior needed in (3). This is usually carried out through a mixture of pdf's defined by,

$$p(\boldsymbol{x}|\lambda_X) = \sum_{k=1}^{M} P(k|\lambda_X) p(\boldsymbol{x}|k, \lambda_X) \tag{4}$$

The typical choice is a GMM where the pdf's $p(\boldsymbol{x}|k, \lambda_X) = p_X(\boldsymbol{x}|k)$ are Gaussians $\mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$ with means $\boldsymbol{\mu}_k$ and covariances $\boldsymbol{\Sigma}_k$ $(k = 1, \ldots, M)$.

The proposed MMSE estimator will also require an statistical model λ_N of noise. We will do the common assumption that the noise statistics are available at every instant. These statistics must be obtained from a previous estimation applied to the observed utterance. We will consider a single Gaussian model (for every time instant),

$$p\left(\boldsymbol{n}|\lambda_N\right) = p_N(\boldsymbol{n}) = \mathcal{N}\left(\boldsymbol{n}; \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N\right) \tag{5}$$

The posterior $p(\boldsymbol{x}|\boldsymbol{y}) \equiv p(\boldsymbol{x}|\boldsymbol{y}, \lambda_X, \lambda_N)$ required in equation (3) can be derived from (4) and (5),

$$p\left(\boldsymbol{x}|\boldsymbol{y}, \lambda_X, \lambda_N\right) = \sum_{k=1}^{M} p\left(\boldsymbol{x}|\boldsymbol{y}, k, \lambda_X, \lambda_N\right) P\left(k|\boldsymbol{y}, \lambda_X, \lambda_N\right) \tag{6}$$

so the MMSE estimate can be finally expressed as,

$$\hat{\boldsymbol{x}} = \sum_{k=1}^{M} P\left(k|\boldsymbol{y}, \lambda_X, \lambda_N\right) \underbrace{\int \boldsymbol{x} p\left(\boldsymbol{x}|\boldsymbol{y}, k, \lambda_X, \lambda_N\right) d\boldsymbol{x}}_{E[\boldsymbol{x}|\boldsymbol{y}, k, \lambda_X, \lambda_N]} \tag{7}$$

As usual in MMSE feature compensation, the above estimate requires the computation of the posterior $P(k|\boldsymbol{y}_u, \lambda_X, \lambda_N)$ and the partial estimate $E[\boldsymbol{x}|\boldsymbol{y}, k, \lambda_X, \lambda_N]$ for every Gaussian component k. In both cases, we have to solve multivariate integrals. We will see next how the OM model can help us to do this. As previously mentioned, this model keeps the maximum between $\boldsymbol{x}$ and $\boldsymbol{n}$ feature by feature. Thus, in order to ease its application to our estimation problem, we will assume statistical independence among features. That is, all Gaussians in λ_X and λ_N will be diagonal and the required integrals can be correspondingly factorized. Statistical independence between speech and noise is also assumed.

For the sake of simplicity, models λ_X and λ_N will be removed from the notation. In the case that not both but only one model applies, this will be indicated with the corresponding subscript ($p_X(\cdot)$ or $p_N(\cdot)$).

Posterior computation. In order to obtain the required posterior, we first apply the Bayes' rule,

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

Therefore, our problem becomes that of computing $p(\boldsymbol{y}|k)$. If we now apply the statistical independence assumptions, this pdf can be factorized into the product of probabilities $p(y|k)$, where y represents a given observed feature.

Thus, we focus now on the computation of,

$$p\left(y|k\right) = \iint p\left(x, n, y|k\right) dxdn \tag{9}$$

$$= \iint p(y|x, n, k) p_X(x|k) p_N(n) dxdn \tag{10}$$

where x and n denote the corresponding clean speech and noise feature, respectively. In this equation, densities $p_X(x|k)$ and $p_N(n)$ are known, but $p(y|x, n, k)$ must be determined. Since the occlusion model forces y to be the maximum of x and n, it can be expressed as,

$$p(y|x, n, k) = p(y|x, n) = \delta\left(y - \max(x, n)\right) \tag{11}$$

where $\delta(\cdot)$ is the Dirac delta function.

On the other hand, the joint speech-noise space $\{(x, n)\}$ can be split into two subsets,

$$\begin{aligned} \mathcal{X} &= \{(x, n) | x \geq n\} \\ \mathcal{N} &= \{(x, n) | n > x\}, \end{aligned} \tag{12}$$

which yields the following expression for (10),

$$p\left(y|k\right) = \iint_{\mathcal{X}} \delta(y - x) p_X(x|k) p_N(n) dxdn + \iint_{\mathcal{N}} \delta(y - n) p_X(x|k) p_N(n) dxdn$$

Now, the integrations over variables x and n can be separated,

$$\begin{aligned} p\left(y|k\right) = &\int_{-\infty}^{\infty} p_X(x|k) \delta(y - x) dx \int_{-\infty}^{x} p_N(n) dn \\ &+ \int_{-\infty}^{\infty} p_N(n) \delta(y - n) dn \int_{-\infty}^{n} p_X(x|k) dx \end{aligned}$$

and it is finally obtained that

$$p\left(y|k\right) = p_X(y|k) C_N(y) + p_N(y) C_X(y|k) \tag{13}$$

where $C_X(y|k)$ and $C_N(y)$ are the cumulative density functions (cdf) corresponding to $p_X(y|k)$ and $p_N(y)$, respectively. Since, in our case, $p_X(y|k)$ and $p_N(y)$ are

Gaussians, these cdfs can be easily computed through the corresponding error functions as,

$$C_X(y|k) = \Phi\left(\frac{y-\mu_k}{\sigma_k}\right), \qquad C_N(y) = \Phi\left(\frac{y-\mu_N}{\sigma_N}\right). \tag{14}$$

It must be pointed out that the resulting posterior of eqn. (13) is the same as that proposed by Varga and Moore in [1] to perform speech recognition in noise. However, while Varga and Moore propose a 3-dimensional Viterbi algorithm to decode speech employing separate hidden Markov models (HMMs) for speech and noise, our proposal is oriented to feature compensation.

Partial estimate computation. Now, we must obtain the partial MMSE estimate $E[\boldsymbol{x}|\boldsymbol{y},k]$ (defined in (7)) applying the OM model. Considering the statistical independence assumptions, we must solve the following expectation,

$$E\left[x|y,k\right] = \int_{-\infty}^{\infty} xp\left(x|y,k\right)dx = \iint xp\left(x,n|y,k\right)dxdn \tag{15}$$

In this case, the pdf required for the integration is $p\left(x,n|y,k\right)$, which can be suitably developed by applying the Bayes' rule as,

$$p\left(x,n|y,k\right) = \frac{p\left(y|x,n\right)p_X\left(x|k\right)p_N\left(n\right)}{p\left(y|k\right)} \tag{16}$$

The integration can be now carried out in a similar way as performed for posterior $P(k|y)$, that is, considering $p(y|x,n) = \delta\left(y - \max(x,n)\right)$ and splitting the speech-noise space (and, therefore, the integral) into the same subsets $\mathcal{X}$ and $\mathcal{N}$ as those defined in (12). Thus, it is finally obtained that,

$$E\left[x|y,k\right] = w_k y + (1-w_k)\tilde{\mu}_k(y) \tag{17}$$

where

$$w_k = \frac{p_X(y|k)C_N(y)}{p(y|k)} \tag{18}$$

$$1 - w_k = \frac{p_N(y)C_X(y|k)}{p(y|k)} \tag{19}$$

$$\tilde{\mu}_k(y) = \int_{-\infty}^{y} x\frac{p_X(x|k)}{C_X(y|k)}dx = \mu_k - \sigma_k\frac{p_X(y|k)}{C_X(y|k)} \tag{20}$$

Discussion. The partial estimate of eqn. (17) is a linear combination of two feature estimates. The first one is the observed feature y, which can be interpreted as an estimate of the clean feature for high SNR values. The second one $\tilde{\mu}_k(y)$ can be interpreted as an estimate of the clean speech when it is completely masked by noise. In this case, we only know that the clean feature is somewhere

between $-\infty$ and y, so $\tilde{\mu}_k(y)$ is the mean value of Gaussian $p_X(x|k)$ truncated at y. Probability w_k acts as a weight which indicates how much y is affected by noise.

The final feature estimate can be expressed as,

$$\hat{x} = \left(\sum_{k=1}^{M} P(k|\boldsymbol{y}) w_k \right) y + \sum_{k=1}^{M} P(k|\boldsymbol{y})(1 - w_k)\tilde{\mu}_k(y) \tag{21}$$

which reflects again a linear combination of the observed feature and an estimate for the case of speech completely masked by noise. This former estimate is obtained as linear combination of the truncated Gaussian means.

3 Comparison with related MD techniques

The OM model has already been employed for feature compensation in previous works. This section is devoted to the comparison between these previous techniques and our proposal. In particular, we will focus on missing-data (MD) imputation techniques where the OM model is employed for spectral reconstruction [5–7].

The starting point of the MD techniques is a binary mask representing the reliability if the observed features. This mask has the same size as the input utterance spectrogram and each pixel m in it indicates whether the corresponding feature y is reliable ($m = 1$) or not ($m = 0$). Considering the OM model, $m = 1$ means $x \geq n$ while $m = 0$ means that the corresponding feature is completely occluded by noise, that is, $n > x$. Then, the conditional probability of eqn. (11) can be compactly written now as,

$$p\left(y|x,n\right) = m\delta(y - x) + (1 - m)\delta(y - n) \tag{22}$$

so that $p(y|k)$, which is required to compute $P(k|\boldsymbol{y})$ in (8), can be obtained as,

$$p\left(y|k\right) = m p_X(y|k) C_N(y) + (1 - m) p_N(y) C_X(y|k) \tag{23}$$

The expected value of eqn. (7) can be derived in a similar way,

$$E\left[x|y,k\right] = my + (1 - m)\tilde{\mu}_k(y) \tag{24}$$

The feature estimate can be finally obtained as,

$$\hat{x} = my + (1 - m) \sum_{k=1}^{M} P(k|\boldsymbol{y})\tilde{\mu}_k(y) \tag{25}$$

Let us compare now the MMSE estimator defined by eqns. (23)-(25) with that obtained in the previous section (eqns. (13), (17) and (21)). It is clear that the MD approach introduces a hard decision (reliable / not reliable) which is a consequence of the binary mask. On the other hand, our proposal avoids this and introduces a soft decision by considering the probability of feature occlusion. As

a result, it can be expected that the proposed technique will be more robust to errors in noise estimation since, in the case of MD, this may lead to erroneous masks and, therefore, to incorrect feature reliability classification.

The perjudicial effect that erroneous masks have over recognition performance is usually mitigated through soft masks [8, 9] where $m \in [0, 1]$. This involves that $m \in [0, 1]$, that is, a continuous degree of reliability from 0 (fully unreliable) until 1 (fully reliable). Formulae (22)-(25) can be kept for this case and can be also found in [8].

We can observe that when soft masks are applied in the MD approach, the resulting estimator has clear similarities with the proposed one, although it must be noticed that our OM-based technique does not require any *a priori* knowledge about the feature reliability (that is, the mask). In particular, comparing the MD and OM estimators of equations (21) and (25), we could consider that our proposal provides a method to estimate the soft mask values as,

$$m = \sum_{k=1}^{M} P\left(k|\boldsymbol{y}\right) w_k \tag{26}$$

This last equation allows a direct comparison of (21) and (25). In both equations we have an estimate $\boldsymbol{y}$ (with weight m) for the case of speech-masking-noise which is linearly combined with an estimate for the case of noise-masking-speech. Although the terms on $\boldsymbol{y}$ are equivalent in both estimators, the comparison also reveals that the noise-masking-speech terms are clearly different.

Fig. 1 shows examples of log-Mel spectrograms for clean and noisy versions of the utterance *eight six zero one one six two* extracted from the Aurora-2 database [10]. Subway noise at 0dB was artificially added to the clean utterance in order to obtain the noisy one. Noise was estimated through linear interpolation of initial noise estimates obtained from the first and last frames of the utterance. It can be seen that the proposed technique effectively compensate for the noise degradation (enhanced speech plot) and also it is able to estimate feature reliability (estimated mask plot).

4 Experimental results

In order to test our proposal and other reference techniques, we have employed the Aurora-2 [10] (connected digits) and Aurora-4 [11] (sentences from WSJ) databases and experimental frameworks. Aurora-2 has 3 test sets: A, B and C. Sets A and B consist of speech artifically contaminated by 4 different types of additive noise in each case (set A: the same noises as in training; set B: different from the training ones), and at 7 different SNRs (-5 to 20 dB, plus clean condition). Set C uses only two types of additive noise and also introduces channel distortion. Aurora-4 is a large vocabulary database with 14 test sets. The first seven sets (T-01 to T-07) artificially add six different noise types (T-01 is the clean condition) with SNR values between 5 dB and 15 dB. The last seven sets are obtained in the same way, but the utterances have been recorded with microphones different than those of training. For both databases, the acoustic

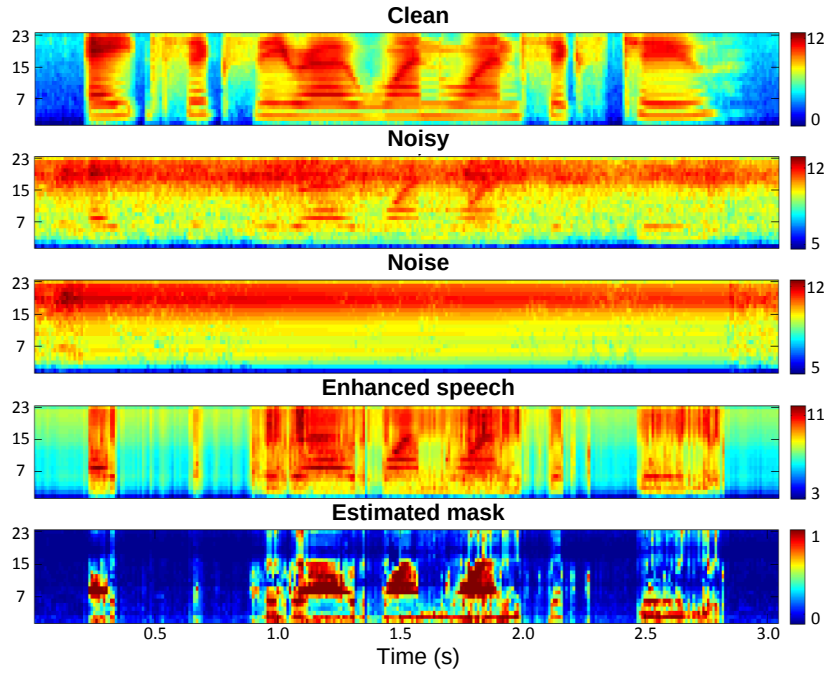

Fig. 1. Example of speech reconstruction and mask estimation (eqn. (26)) from the proposed OM- MMSE-based estimator.

models are trained with the usual scripts provided with the databases and using only clean speech.

The final feature vector employed for recognition consist of 13 Mel-frequency cepstral coefficients (MFCCs) ($C0$ is included instead of log-Energy) enlarged with Δ and $\Delta\Delta$ coefficients. Feature compensation is carried out over the 23 log-outputs of the Mel filterbank, which are DCT-transformed to obtain MFCCs. Also, cepstral mean normalization (CMN) is applied in order to mitigate channel distortions.

The clean spectral features were modeled with a 256-component GMM with diagonal covariance matrices, which has been trained through the expectation-maximization algorithm over the corresponding training set. The required noise estimates are obtained as follows: the first and last T frames ($T = 20$ for Aurora-2, $T = 35$ for Aurora-4) of every utterance have been averaged and the estimates for the intermediate frames are obtained through linear interpolation between the former ones. The noise model at every frame is completed with a covariance matrix fixed for all frames and computed from the first and last frames.

In order to assess our proposal in comparison with other techniques, the MD estimators described in the previous section has been also evaluated. Both, binary and soft masks have been considered. The binary masks are simply obtained by SNR thresholding at 0 dB. Soft masks are obtained by two different

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg.
Baseline	99.11	97.29	92.55	75.56	42.82	22.69	12.92	66.18
Oracle	99.11	99.01	98.74	97.84	95.72	89.64	73.79	96.19
BMD	98.88	97.45	95.32	90.01	78.47	54.99	25.55	83.25
SMD1	98.90	98.04	96.51	92.15	80.62	56.70	26.89	84.80
SMD2	98.91	97.91	96.32	91.74	79.77	55.30	26.20	84.21
SRO	98.91	98.08	96.69	92.77	82.18	58.76	27.21	85.70

Table 1. Word accuracy results (%) for Aurora-2 at different SNRs.

methods: a relaxation of the binary mask through a sigmoid function (its center and slope parameters has been optimized for a validation subset) and the mask defined by equation (26) and based on the proposed models (OM as well as clean speech and noise models). This last mask allows a direct comparison between MD imputation and MMSE estimation (both based on the occlusion model).

The word accuracy results for Aurora-2 are shown in table 1. The baseline corresponds to MFCCs with CMN. Three MD imputation techniques with three different types of masks are considered: masks obtained from the actual noise (Oracle), binary masks (BMD) and soft-masks (SMD1 and SMD2 for sigmoid-based and model-based masks, respectively). The oracle results can be considered an upper bound of the MD techniques (since the feature reliability is perfectly known). Our spectral reconstruction based on the OM model will be denoted as SRO. The results correspond to the average score over sets A, B and C for every SNR. Also, the average (Avg.) for SNRs from 0 to 20 dB is shown.

As it could be expected, the Oracle experiment achieves the best performance, but SRO provides the best results with estimated noise. Therefore, since the different techniques can be all considered variants of MD imputation which mainly differ in the way the mask values are computed (as explained in the previous section), we can say that SRO is more robust against noise estimation errors and, therefore, to mask errors. In this regard, the worst behavior corresponds to BMD. In this case, when a mask error ocurrs, an unreliable feature can be classified as reliable and the other way round. In the first case, the observed unreliable feature is kept. The second case is even worse, since it involves that reliable feature are treated as unreliable, being degraded by the estimation processing. In the case of MD with soft masks, the use of a mask looks *redundant* with the estimation based on a noise model as it is evident in equations (22) and (23), since the estimator can obtain its own mask (eqn. (26)). SMD1 yields results slightly better than SMD2 since SMD1 involves an optimization of the sigmoid parameters while the masks in SMD2 are completely extracted from statistical and OM models.

The results for Aurora-4 can be found in table 2. The proposed SRO technique outperforms again the MD imputation techniques with binary or soft masks, with relative improvements of 10.90 % and 1.77 %, respectively.

	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.
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
Oracle	87.69	86.74	84.46	84.44	83.19	85.90	82.38	79.13	77.86	74.03	73.45	70.48	75.04	71.77	79.75
BMD	86.96	80.78	58.47	52.74	59.63	56.14	61.42	79.39	74.13	54.83	46.76	50.55	51.26	56.17	62.09
SMD2	87.52	83.65	66.62	63.78	63.48	69.19	65.31	81.00	75.64	60.98	55.02	54.89	62.39	57.74	67.66
SRO	87.54	83.28	69.23	64.49	64.88	70.63	66.93	80.52	76.48	63.53	55.67	56.62	63.87	60.38	68.86

Table 2. Word accuracy results (%) for the different test sets of Aurora-4.

5 Conclusions

In this work we have proposed a technique for the MMSE estimation of log-spectral features corrupted by additive noise. The starting point is a simplification of a general noise distortion model through the *log-max* approximation, which yields the so-called occlusion model. This modeling involves that either the speech feature dominates the noise or, on the contrary, the speech is completely masked by noise. The resulting estimator has clear similarities with some MD imputation techniques. Indeed, it can be considered an MD technique or equivalently, a way of computing soft masks for MD imputation. Our experimental results have shown the superiority of our proposal over the reference MD techniques.

References

1. A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise", in *Proc. ICASSP*, pp. 845–848, Apr. 1990.
2. X. Huang, A. Acero, and H. Hon, "Spoken language processing: A guide to theory, algorithm, and system development", *Prentice Hall*, 2001.
3. 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.
4. A. M. Reddy, and B. Raj, "Soft Mask Methods for Single-Channel Speaker Separation", *IEEE Trans. Audio Speech and Language Process.*, vol. 15, no. 6, pp. 1766-1776, Aug. 2007.
5. M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable data", *Speech Comm.*, vol. 34, no. 3, pp. 267–285, June 2001.
6. B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition", *Speech Comm.*, vol. 48, no. 4, pp. 275–296, 2004.
7. 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*, pp. 4693-96, Mar. 2012.
8. B. Raj and R. Singh, "Reconstructing spectral vectors with uncertain spectrographic masks for robust speech recognition", in *Proc. ASRU*, pp. 275–296, pp. 65–70, 2005.
9. F. Faubel, H. Raja, J. McDonough, and D. Klakow, "Particle filter based soft-mask estimation for missing-feature reconstruction", in *Proc. IWAENC*, 2008.
10. H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluations of the speech recognition systems under noisy conditions", in *ISCA ITRW ASR2000*, Paris, France, 2000.
11. H. G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends of large vocabulary task", Tech. Rep., STQ AURORA DSR Working Group, 2002.