Log-spectral feature reconstruction based on an occlusion model for noise robust speech recognition

José A. González, Antonio M. Peinado, Angel M. Gómez

Dpt. Signal Theory, Telematics, and Communications
University of Granada, Spain
{joseangl,amp,amgg}@ugr.es

Ning Ma

Dpt. Computer Science
University of Sheffield, UK
n.ma@dcs.shef.ac.uk

Abstract

This paper addresses the problem of feature compensation in the log-spectral domain for speech recognition in noise by recasting the speech distortion problem as an occlusion one. The usual non-linear mismatch function that represents the speech distortion due to additive noise can be reasonably well approximated by the maximum of the two mixing sources (speech and noise). Using this approximation, we propose to enhance the degraded speech features by means of a novel minimum mean square error (MMSE) estimator. The resulting technique shows clear similarities with soft-mask missing-data (MD) reconstruction, although the experimental results on both Aurora-2 and Aurora-4 databases show the effectiveness of the proposed technique in comparison with MD.

Index Terms: Feature compensation, MMSE estimation, Missing data imputation, speech recognition

1. Introduction

Automatic speech recognition (ASR) systems are currently moving from close-talk dictation tasks to mobile scenarios in which ASR is used as a more efficient and natural method to access information. In these scenarios, a number of sources of distortion such as different environmental noises, channel distortions, and room responses could affect the performance of these systems. Consequently, accomplishing noise robustness is becoming a key issue to make ASR deployable in real world conditions.

Traditionally two different approaches has been considered to minimize the mismatch produced by the noise [1]: *feature compensation*, which tries to remove the noise from the parameters representing the speech, and *model adaptation*, where the acoustic model parameters are modified to better represent the operating conditions. Feature compensation has the advantage that it can be seamlessly incorporated into existing systems as a front-end. Moreover, it is usually more efficient than model adaptation.

In this paper, a novel feature compensation technique working in the log-spectral domain is proposed. In this domain, speech distortion caused by additive noise can be considered as an occlusion problem: while some log-spectral speech features are almost unaffected by the noise, other are completely masked by it. To estimate the masked features, a minimum mean square error (MMSE) estimator using a Gaussian mixture model (GMM) to represent the distribution of clean speech features is derived in Section 2 of this work. As will be shown, the proposed estimator effectively tackles the occlusion problem by computing a linear combination of the observed feature (non-occlusion case) and a partial estimate obtained for the case of total occlusion. The analogies and differences of the proposed reconstruction technique with other similar approaches are discussed in Section 3. Experimental results for the Aurora-2 and Aurora-4 databases are reported in Section 4. Finally, conclusions can be found in Section 5.

2. Proposed reconstruction technique

Let $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ be the feature vector corresponding to the observed log-Mel filterbank energies for noisy speech. This vector is related with the unknown vectors $\boldsymbol{x}$ and $\boldsymbol{n}$, corresponding, respectively, to clean speech and additive noise log-Mel filterbank energies, through the following model [2],

$$y_i = x_i + \log\left(1 + e^{n_i - x_i}\right) + r_i, \tag{1}$$

where r_i is a residual term that depends on the phase relationship between clean speech and noise. Making the usual assumption that r_i is negligible compared to the other terms in (1), the above model can be simplified as follows,

$$y_i \approx \max(x_i, n_i), \tag{2}$$

where the *log-max* approximation has been considered to further simplify the model (i.e. $\log(e^x + e^n) \approx \max(x, n)$) [3].

We will refer to (2) as the *noise occlusion model*. According to this model, the noise distortion involves that some spectro-temporal regions of the original clean speech are completely lost, while others remain almost unaffected. In this work, we will use this fact to estimate the clean speech features masked by noise. To do so, the redundancy and sparseness of speech signals will be exploited.

In order to compensate for the effects of noise, the MMSE criterion is adopted in this work. Thus, the MMSE estimate of the clean feature vector can be computed as,

$$\hat{\boldsymbol{x}} = E[\boldsymbol{x}|\boldsymbol{y}] = \int \boldsymbol{x} p\left(\boldsymbol{x}|\boldsymbol{y}\right) d\boldsymbol{x}. \tag{3}$$

As a first step to obtain the posterior distribution in (3), we assume that the clean speech feature distribution can be modeled using a GMM λ_X as follows,

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

where $\boldsymbol{\mu}_X^k$ and $\boldsymbol{\Sigma}_X^k$ are the mean vector and covariance matrix of the kth component in the GMM.

We also consider that, for every time instant, the noise spectra can be estimated. Note that this is the usual assumption

made by both feature compensation and speech enhancement techniques. Moreover, it is assumed that a (diagonal) covariance matrix is also available for every estimate, so that the uncertainty of noise estimation can be accounted for. Hence, the noise is assumed to be Gaussian distributed as follows,

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

where λ_N is the noise model and the time dependency is omitted for the sake of simplicity.

Using (4) and (5), the posterior $p(\boldsymbol{x}|\boldsymbol{y}) \equiv p(\boldsymbol{x}|\boldsymbol{y}, \lambda_X, \lambda_N)$ in (3) can be obtained by marginalizing over the Gaussian components of the clean speech GMM as follows,

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

Applying (6) to (3), the MMSE estimate becomes

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

As can be seen, the MMSE estimate requires the computation of the posterior $P(k|\boldsymbol{y}, \lambda_X, \lambda_N)$ and the partial estimate $E[\boldsymbol{x}|\boldsymbol{y}, k, \lambda_X, \lambda_N]$ for every Gaussian component k. Let us first consider the computation of the posterior probability. Using Bayes' rule, we obtain:

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

where speech and noise are assumed to be statistically independent. Furthermore, by assuming independence among features, $p(\boldsymbol{y}|k, \lambda_X, \lambda_N)$ in (8) can be expressed as

$$p\left(\boldsymbol{y}|k, \lambda_X, \lambda_N\right) = \prod_{i=1}^{n} p\left(y_i|k, \lambda_X, \lambda_N\right). \tag{9}$$

The value of $p(y_i|k, \lambda_X, \lambda_N)$ can be obtained by marginalizing $p(x_i, n_i, y_i|k, \lambda_X, \lambda_N)$ over those values of x_i and n_i that satisfy the occlusion model in (2), i.e. $\max(x_i, n_i) = y_i$. Thus, this probability can be obtained as,

$$p\left(y_i|k, \lambda_X, \lambda_N\right) = \iint p\left(x_i, n_i, y_i|k, \lambda_X, \lambda_N\right) dx_i dn_i. \tag{10}$$

Assuming again independence between speech and noise, the probability distribution in (10) can be factorized as the product of the three following terms,

$$p\left(x_i, n_i, y_i|k, \lambda_X, \lambda_N\right) = p\left(y_i|x_i, n_i\right) p\left(x_i|k, \lambda_X\right) p\left(n_i|\lambda_N\right) \tag{11}$$

where we have considered that y_i is statistically independent of the models λ_X and λ_N provided that x_i and n_i are known.

Taking into account the noise occlusion model in (2), $p(y_i|x_i, n_i)$ is given by,

$$\begin{aligned} p\left(y_i|x_i, n_i\right) &= \delta\left(y_i - \max\left(x_i, n_i\right)\right) \\ &= \delta\left(y_i - x_i\right) \mathbb{1}_{n_i \leq x_i} + \delta\left(y_i - n_i\right) \mathbb{1}_{x_i < n_i} \end{aligned} \tag{12}$$

with $\delta(x - a)$ being the Dirac delta function translated to a, and $\mathbb{1}_{\mathcal{C}}$ is the usual indicator function being equal to 1 for those values that satisfy the condition $\mathcal{C}$, and 0 otherwise.

Finally, using (11) and (12) into (10), results in the observation probability shown in (13) (next page), where $\Phi(\cdot)$ is the Gaussian cumulative distribution function (CDF). As can be seen, the resulting equation is the same as that proposed by Varga and Moore in [3] to perform speech recognition in noise. Nevertheless, while Varga and Moore propose a 3-dimensional Viterbi algorithm to decode speech over separate hidden Markov models (HMMs) for speech and noise, a feature compensation technique is proposed here.

Once the derivation of the posterior probability in (7) is completed, we will tackle the computation of the expectation term in the MMSE estimate. Assuming again independence among features, this term corresponds to the following integral,

$$E\left[x_i|y_i, k, \lambda_X, \lambda_N\right] = \int x_i p\left(x_i|y_i, k, \lambda_X, \lambda_N\right) dx_i, \tag{14}$$

where the probability can be obtained through marginalization over the hidden noise variable n_i:

$$p\left(x_i|y_i, k, \lambda_X, \lambda_N\right) = \int p\left(x_i, n_i|y_i, k, \lambda_X, \lambda_N\right) dn_i. \tag{15}$$

Applying Bayes' rule, $p\left(x_i, n_i|y_i, k, \lambda_X, \lambda_N\right)$ can be expressed as,

$$p\left(x_i, n_i|y_i, k, \lambda_X, \lambda_N\right) = \frac{p\left(y_i|x_i, n_i\right) p\left(x_i|k, \lambda_X\right) p\left(n_i|\lambda_N\right)}{p\left(y_i|k, \lambda_X, \lambda_N\right)} \tag{16}$$

where $p(y_i|x_i, n_i)$ is given by (12) and $p(y_i|k, \lambda_X, \lambda_N)$ by (13).

Then, using (15), (16) and (12), the expected value in (14) becomes that in (17) (next page), where we introduced the weights w_i^k being defined as,

$$w_i^k = \frac{p\left(y_i|k, \lambda_X\right) \int_{-\infty}^{y_i} p\left(n_i|\lambda_N\right) dn_i}{p\left(y_i|k, \lambda_X, \lambda_N\right)} \tag{18}$$

and $\tilde{\mu}_{X,i}^k$ is the mean of a right-truncated Gaussian distribution taking values within the interval $(-\infty, y_i]$. This value can be obtained as [4],

$$\begin{aligned} \tilde{\mu}_{X,i}^k &= \frac{1}{\Phi_X\left(\frac{y_i - \mu_{X,i}^k}{\sigma_{X,i}^k}\right)} \int_{-\infty}^{y_i} x_i p\left(x_i|k, \lambda_X\right) dx_i \\ &= \mu_{X,i}^k - \sigma_{X,i}^k \frac{\mathcal{N}_X\left(\frac{y_i - \mu_{X,i}^k}{\sigma_{X,i}^k}\right)}{\Phi_X\left(\frac{y_i - \mu_{X,i}^k}{\sigma_{X,i}^k}\right)}. \end{aligned} \tag{19}$$

The resulting estimator in (17) has a clear interpretation as a linear combination of two terms. The observation in the first term, y_i, corresponds to the clean speech estimate for the case of undistorted speech. The second term, $\tilde{\mu}_{X,i}^k$, is the estimate when the speech is completely masked by noise. In this situation, the estimate computed for each Gaussian is the expected value between $-\infty$ and the upper bound constraint imposed by the observation y_i. Both terms y_i and $\tilde{\mu}_{X,i}^k$ are weighted by the probabilities w_i^k and $1 - w_i^k$, corresponding, respectively, to the probability of speech being unaffected by noise and the probability of total occlusion.

$$p\left(y_i|k,\lambda_X,\lambda_N\right) = \iint p\left(x_i|k,\lambda_X\right) p\left(n_i|\lambda_N\right) \delta\left(y_i - x_i\right) \mathbb{1}_{n_i \leq x_i} dx_i dn_i + \iint p\left(x_i|k,\lambda_X\right) p\left(n_i|\lambda_N\right) \delta\left(y_i - n_i\right) \mathbb{1}_{x_i < n_i} dx_i dn_i$$
$$= p\left(y_i|k,\lambda_X\right) \int_{-\infty}^{y_i} p\left(n_i|\lambda_N\right) dn_i + p\left(y_i|\lambda_N\right) \int_{-\infty}^{y_i} p\left(x_i|k,\lambda_X\right) dx_i$$
$$= \mathcal{N}_X\left(y_i; \mu_{X,i}^k, \sigma_{X,i}^k\right) \Phi_N\left(\frac{y_i - \mu_{N,i}}{\sigma_{N,i}}\right) + \mathcal{N}_N\left(y_i; \mu_{N,i}, \sigma_{N,i}\right) \Phi_X\left(\frac{y_i - \mu_{X,i}^k}{\sigma_{X,i}^k}\right) \quad (13)$$

$$E\left[x_i|y_i,k,\lambda_X,\lambda_N\right] = \frac{1}{p\left(y_i|k,\lambda_X,\lambda_N\right)} \left[y_i p\left(y_i|k,\lambda_X\right) \int_{-\infty}^{y_i} p\left(n_i|\lambda_N\right) dn_i + p\left(y_i|\lambda_N\right) \int_{-\infty}^{y_i} x_i p\left(x_i|k,\lambda_X\right) dx_i\right]$$
$$= w_i^k y_i + (1 - w_i^k)\tilde{\mu}_{X,i}^k \quad (17)$$

3. Comparative discussion

We can find in the literature several other feature compensation techniques based on the noise occlusion model in (2) that are similar to the proposed log-spectral reconstruction method. In this section, we will analyze the relationship between these techniques and our proposal. In particular, we will focus on the missing-data approach to spectral reconstruction [5, 6, 7].

Missing-data techniques assume that knowledge about the feature reliability is available *a priori* through a binary mask $\boldsymbol{m}$. In this mask the undistorted clean speech features (reliable features) are represented by $m_i = 1$, while the occluded features (unreliable or missing features) are represented by $m_i = 0$. Using this information, the conditional probability in (12) for the missing-data techniques is,

$$p\left(y_i|x_i,n_i\right) = m_i \delta\left(y_i - x_i\right) \mathbb{1}_{n_i \leq x_i} + (1 - m_i)\delta\left(y_i - n_i\right) \mathbb{1}_{x_i < n_i}. \quad (20)$$

We define $\boldsymbol{s}_r$ and $\boldsymbol{s}_u$ as the sets containing the frequency indexes corresponding to reliable and unreliable features, respectively. Then, substituting (10), (11) and (20) into (9), the observation probability $p(\boldsymbol{y}|k,\lambda_X,\lambda_N)$ employed by the missing-data techniques can be obtained as,

$$p\left(\boldsymbol{y}|k,\lambda_X,\lambda_N\right) = \prod_{i \in \boldsymbol{s}_r} p\left(y_i|k,\lambda_X\right) \int_{-\infty}^{y_i} p\left(n_i|\lambda_N\right) dn_i \times \prod_{j \in \boldsymbol{s}_u} p\left(y_j|\lambda_N\right) \int_{-\infty}^{y_j} p\left(x_j|k,\lambda_X\right) dx_j$$
$$= \gamma \prod_{i \in \boldsymbol{s}_r} p\left(y_i|k,\lambda_X\right) \prod_{j \in \boldsymbol{s}_u} \int_{-\infty}^{y_j} p\left(x_j|k,\lambda_X\right) dx_j \quad (21)$$

with

$$\gamma = \prod_{i \in \boldsymbol{s}_r} \int_{-\infty}^{y_i} p\left(n_i|\lambda_N\right) dn_i \prod_{j \in \boldsymbol{s}_u} p\left(y_j|\lambda_N\right). \quad (22)$$

As can be noted, γ can be considered as a constant value since it depends only on the noise model λ_N. Thus, it does not affect to the computation of $P(k|\boldsymbol{y},\lambda_X,\lambda_N)$ in (8) and, hence, it can be discarded from (21).

Proceeding in the same manner as for the derivation of (21), it is easy to see that the expectation in (7) obtained by the missing-data approach is given by,

$$E\left[x_i|y_i,k,\lambda_X,\lambda_N\right] = \begin{cases} y_i & m_i = 1 \\ \tilde{\mu}_{X,i}^k & m_i = 0 \end{cases} \quad (23)$$

By comparing the estimation formulae of the missing-data approach (eqns. (21) and (23)) with those obtained for the proposed reconstruction (eqns. (9), (13), and (17)), an important difference can be observed. While the use of a binary mask in the missing-data approach involves a hard decision, i.e. the features are considered either reliable or completely masked, a soft decision is implemented in our approach by exploiting the probabilities of feature occlusion. Hence, it is expected that our technique will be more resilient to errors in noise estimation or, alternatively, errors in the estimation of the missing-data masks, than the missing-data approach.

To overcome the limitations of the binary masks, the use of soft masks has been also considered for missing-feature reconstruction [8, 9]. Instead of performing a binary classification of the features according to their reliability, a confidence value $m_i \in [0,1]$ is now assigned to every feature. Then, using soft masks, $p\left(y_i|k,\lambda_X,\lambda_N\right)$ in (9) is computed as follows [8],

$$p\left(y_i|k,\lambda_X,\lambda_N\right) = m_i p\left(y_i|k,\lambda_X\right) \int_{-\infty}^{y_i} p\left(n_i|\lambda_N\right) dn_i + (1 - m_i)\, p\left(y_i|\lambda_N\right) \int_{-\infty}^{y_i} p\left(x_i|k,\lambda_X\right) dx_i \quad (24)$$

and the estimate for each Gaussian component is,

$$E\left[x_i|y_i,k,\lambda_X,\lambda_N\right] = m_i y_i + (1 - m_i)\,\tilde{\mu}_{X,i}^k. \quad (25)$$

As can be seen, the resulting missing-data technique using soft masks turns out to be very similar to the proposed reconstruction technique. Nevertheless, the proposed technique requires no *a priori* knowledge about the feature reliability. Indeed, our technique can be alternatively considered as a robust technique for soft mask estimation, in which the confidence values for every feature are computed as,

$$m_i = \sum_{k=1}^{M} P\left(k|\boldsymbol{y},\lambda_X,\lambda_N\right) w_i^k \quad (26)$$

with w_i^k being computed according to (18).

4. Experimental results

The proposed technique was evaluated on Aurora-2 [10] and Aurora-4 [11] databases. Aurora-2 consists of utterances of English connected digits distorted by noise. Three tests sets are

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg.	R.I.%
Baseline	99.11	97.29	92.55	75.56	42.82	22.69	12.92	66.18	-
Oracle	99.11	99.01	98.74	97.84	95.72	89.64	73.79	96.19	45.34
BMD	98.88	97.45	95.32	90.01	78.47	54.99	25.55	83.25	25.79
SMD	98.91	97.91	96.32	91.74	79.77	55.30	26.20	84.21	27.24
SRO	98.91	98.08	96.69	92.77	82.18	58.76	27.21	85.70	29.49

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

	T-01	T-02	T-03	T-04	T-05	T-06	T-07	T-08	T-09	T-10	T-11	T-12	T-13	T-14	Avg.	R.I.%
Baseline	87.69	75.30	53.24	53.15	46.80	56.36	45.38	77.04	64.24	45.30	42.07	36.15	47.43	36.67	54.77	-
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	45.61
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	13.36
SMD	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	23.52
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	25.72

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

defined: Set A, Set B, and Set C. Set A and Set B employs eight different types of additive noise at 7 signal-to-noise ratio (SNR) values. Set C employs only two types of additive noise and also considers a different linear filtering distortion. For the Aurora-4 large vocabulary database, 14 test sets are defined. In the first seven sets (T-01 to T-07), six different noise types are considered (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 are recorded with different microphones than the one used for recording the training set. For both databases, the acoustic models are trained with the usual scripts provided with the databases using clean speech.

The speech features employed by the recognizer are 13 Mel-frequency cepstral coefficients (MFCCs) (C0 is used instead of the log energy) along with their delta and delta-delta coefficients. Spectral reconstruction is applied to the 23 log-Mel filterbank channels. After reconstruction, the discrete cosine transform (DCT) is used to obtain the final cepstral parameters. Cepstral mean normalization (CMN) is applied in all cases to increase the robustness against channel mismatches.

Spectral features are modeled using a GMM with 256 components and diagonal covariances. Training is carried out by means of the EM algorithm on the same clean dataset as for acoustic model training. Noise estimates are obtained for every time instant through linear interpolation of initial noise estimates computed by averaging the first and last frames of each utterance (20 frames for Aurora-2 and 35 frames for Aurora-4). A fixed time-invariant diagonal covariance is assumed for all the noise estimates. This covariance is also computed from the first and last frames of the utterance.

For comparative purposes, both binary-mask and soft-mask missing-data approaches described in Section 3 are also tested. The binary masks are obtained from the aforementioned noise estimates using a fixed SNR threshold of 0 dB for both databases. The soft masks are obtained from the noise estimates using (26).

Table 1 shows the word accuracy results (WAcc) for the Aurora-2 database. This table compares the baseline system (MFCC features plus CMN) with four reconstruction techniques: the missing-data reconstruction technique using perfect knowledge about the feature reliability (Oracle), the same technique using estimated binary masks (BMD), the soft-mask missing-data approach (SMD), and the proposed spectral reconstruction technique based on the occlusion model (SRO). The results from the three test sets are averaged for each SNR. In addition, the result from an overall average between 0 dB and 20 dB (Avg.) and the relative improvement (R.I.) regarding the baseline are also shown for every technique.

Spectral reconstruction with perfect knowledge about the feature reliability (Oracle) yields the best results. Hence, this can be considered as an upper bound for the performance of the techniques derived from the occlusion model in (2). When noise is estimated, the performance of these techniques suffer a degradation. Nevertheless, SRO presents a better robustness than BMD and SMD to noise estimation errors. For BMD, this difference can can be explained by the use of binary masks. Thus, in case of mask estimation errors, unreliable features could be identified as being reliable and vice-versa. In the first case, unreliable features will be kept untreated. In the second case, the reliable features labeled as unreliable will be replaced and, hence, a greater error will be obtained. In SMD, the use of both soft masks and a noise distribution as shown in (24) is somehow redundant, resulting in a poorer performance.

Table 2 shows the results for the Aurora-4 database. Again, SRO outperforms BMD and SMD, yielding average relative improvements of 10.90 % and 1.77 % regarding both techniques, respectively.

5. Conclusions

In this work, a novel technique for compensating log-spectral features distorted by additive noise has been proposed. This technique is based on a simplification of the noise distortion model that only considers features as reliable or completely masked by noise. Experimental results show the effectiveness of our proposal in compensating the noise effects.

6. Acknowledgements

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

7. References

[1] X. Huang, A. Acero, and H. Hon, "Spoken language processing: A guide to theory, algorithm, and system development", *Prentice Hall*, 2001.

[2] 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.

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

[4] N. L. Johnson, "Continuous univariate distributions", Wiley, vol. 1, 1994.

[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. 4693–4696, 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 ASR*, 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.