HMM-based MAP Prediction of Voiced and Unvoiced Formant Frequencies from Noisy MFCC Vectors

Jonathan Darch, Ben Milner

School of Computing Sciences,
University of East Anglia, Norwich, U.K.
jonathan.darch@uea.ac.uk, b.milner@uea.ac.uk

10.21437/Interspeech.2006-326

Abstract

This paper describes how formant frequencies of voiced and unvoiced speech can be predicted from mel-frequency cepstral coefficients (MFCC) vectors using maximum a posteriori (MAP) estimation within a hidden Markov model (HMM) framework. Gaussian mixture models (GMMs) are used to model the local joint density of MFCCs and formant frequencies. More localised prediction is achieved by modelling speech using voiced, unvoiced and non-speech GMMs for every state of each model of a set of HMMs. To predict formant frequencies from a MFCC vector, first a prediction of the speech class (voiced, unvoiced or non-speech) is made. Formant frequencies are predicted from voiced and unvoiced speech using a MAP estimation made using the state-specific GMMs. This 'HMM-GMM' prediction of speech class and formant frequencies was evaluated on a male 5000 word unconstrained large vocabulary speaker-independent database.

Index Terms: formants, MAP, GMM, HMM, DSR.

1. Introduction

Formants correspond to resonances in the vocal tract and may be used to recognise, synthesise, encode or enhance speech. Traditional methods for their estimation include peak-picking using either short-time spectra [1] or linear predictive coding (LPC) spectra [2] and root-finding using LPC analysis [3]. However, in a distributed speech recognition (DSR) environment, only MFCC vectors are transmitted to the remote back-end. Inverting MFCC vectors to magnitude spectra through zero-padding, inverse discrete cosine transform (DCT), exponential operation and interpolation results in spectral smoothing as much information is lost, in particular precise formant location information [4]. Because it is harder to distinguish between potential formants in such spectrally smooth spectra, traditional formant estimation techniques would be unable to accurately estimate formants.

Previous work has shown how formant frequencies associated with voiced speech can be predicted from MFCC vectors using a single GMM which models voiced speech [4]. Neither the temporal correlation of formants nor the model and state specific relationship between MFCCs and formant frequencies were considered. It has also been demonstrated that fundamental frequency can be predicted from MFCC vectors by employing model and state specific GMMs within a framework of a set of HMMs [5]. The aim of the work presented here is to extend these prediction techniques to firstly predict the speech class as voiced, unvoiced or non-speech. Secondly, for MFCC vectors predicted as voiced or unvoiced, appropriate state-specific GMMs will be used to predict formant frequencies from MFCCs of unconstrained speech using maximum a posteriori (MAP) estimation. Experimental results are presented in section 3 and conclusions are drawn in section 4.

2. HMM-GMM Prediction

Prediction of speech class and formant frequencies from MFCC vectors comprises two parts. First, the local joint density of MFCCs and formant frequencies is modelled using GMMs specific to the states of a set of HMMs. Secondly, a prediction of speech class (voiced, unvoiced or non-speech) is made. For MFCC vectors predicted as voiced or unvoiced, formant frequencies are predicted using the statistical information in the state-specific models. Model and state sequences are determined through Viterbi decoding.

2.1. Modelling of MFCCs and Formant Frequencies

Modelling the joint density of MFCC vectors and formant frequencies requires forming a set of augmented feature vectors, $\mathbf{y}$:

$$\mathbf{y}_i = [\mathbf{x}_i, \mathbf{F}_i]^T \tag{1}$$

where vector $\mathbf{x}_i = [x(0), x(1), \ldots x(12), ln(e)]$ comprises static MFCCs 0 to 12 and log energy for the i^{th} frame of speech. The formant frequency vector $\mathbf{F}_i = [F(1), F(2), F(3), F(4)]$ comprises the frequencies of the first four formants of the i^{th} frame of speech. Reference formant frequencies are obtained using LPC analysis to produce the poles of each frame of speech which form initial formant estimates. Kalman filtering is used to improve the accuracy of the resulting formant frequency estimates [6]. During non-speech periods, the formant vector, $\mathbf{F}_i$ is set to zero. Reference voicing decisions are obtained using the ETSI Aurora Extended Advanced Front End voicing classifier [7].

A set of W monophone HMMs, $\mathbf{\Lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_W]$, is created such that each monophone is modelled by a HMM. Baum-Welch re-estimation is used to train the set of HMMs using each MFCC vector, $\mathbf{x}$, along with its velocity and acceleration derivatives. Associated with each state of every HMM are three GMMs. Two model the joint density of MFCC vectors and formant frequencies: one for voiced speech and the second for unvoiced speech. The third GMM which models non-speech periods only contains MFCCs, as formants are not associated with non-speech. These state-specific GMMs are created by realigning training data vectors to the HMMs using Viterbi decoding. For each training utterance, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$, comprising N MFCC vectors, model allocations, $\mathbf{m} = [m_1, m_2, \ldots, m_N]$, and state allocations, $\mathbf{q} = [q_1, q_2, \ldots, q_N]$, are obtained by forced alignment with annotations. For a given utterance, the i^{th} MFCC vector, $\mathbf{x}_i$, is asso-

ciated with a model, m_i, and state, q_i, where $m_i = 1, \ldots, W$ and $q_i = 1, \ldots, S_{m_i}$. The number of models is given by W and the number of states in model m_i is given by S_{m_i}.

Vector pools are formed to collect together feature vectors from each model, w, and state, s, for voiced, unvoiced and non-speech frames from the set of training data, Z. The pool of training feature vectors deemed voiced is given by:

$$\Omega_{s,w} = \{\mathbf{y}_i \in Z \,:\, voicing(\mathbf{x}_i) = \text{voiced},\ q_i = s,\ m_i = w\} \quad 1 \le s \le S_w, \quad 1 \le w \le W \quad (2)$$

where $voicing(\mathbf{x}_i)$ is either voiced, unvoiced or non-speech according to the reference speech class decision.

Similarly, unvoiced, $\Psi_{s,w}$, and non-speech, $\Upsilon_{s,w}$, pools are formed. For certain model and state combinations, pools will be empty, or at least sparse, due to lack of data. For example, there are no non-speech feature vectors for the centre state of model /ae/.

Model and state dependent voiced, unvoiced and non-speech GMMs ($\Phi^v_{s,w}$, $\Phi^u_{s,w}$ and $\Phi^{ns}_{s,w}$) each comprising K clusters are created from each of the vector pools using unsupervised expectation-maximisation (EM) training. For example, the state-specific voiced GMMs for state s and model w are given by:

$$\begin{aligned}\Phi^v_{s,w}(\mathbf{y}) &= \sum_{k=1}^{K} \alpha^v_{k,s,w}\ \phi^v_{k,s,w}(\mathbf{y}) \\ &= \sum_{k=1}^{K} \alpha^v_{k,s,w}\ \mathcal{N}(\mathbf{y}, \boldsymbol{\mu}^{v,y}_{k,s,w}, \boldsymbol{\Sigma}^{v,yy}_{k,s,w}) \end{aligned} \quad (3)$$

where $\alpha^v_{k,s,w}$ is the prior probability of the k^{th} cluster in the voiced GMM for state s of model w. The voiced GMM comprises K Gaussian PDFs, $\phi^v_{k,s,w}(\mathbf{y})$, given by mean vector, $\boldsymbol{\mu}^{v,y}_{k,s,w}$, and covariance matrix, $\boldsymbol{\Sigma}^{v,yy}_{k,s,w}$:

$$\boldsymbol{\mu}^{v,y}_{k,s,w} = \begin{bmatrix} \boldsymbol{\mu}^{v,x}_{k,s,w} \\ \boldsymbol{\mu}^{v,F}_{k,s,w} \end{bmatrix} \text{ and } \boldsymbol{\Sigma}^{v,yy}_{k,s,w} = \begin{bmatrix} \boldsymbol{\Sigma}^{v,xx}_{k,s,w} & \boldsymbol{\Sigma}^{v,xF}_{k,s,w} \\ \boldsymbol{\Sigma}^{v,Fx}_{k,s,w} & \boldsymbol{\Sigma}^{v,FF}_{k,s,w} \end{bmatrix} \quad (4)$$

where $\boldsymbol{\Sigma}^{v,xF}_{k,s,w}$ is the cross-covariance matrix of MFCC and formant frequency vectors for the k^{th} cluster of model w and state s. This matrix describes the local relationships between MFCCs and formant frequencies.

Model and state dependent unvoiced and non-speech GMMs ($\Phi^u_{s,w}$ and $\Phi^{ns}_{s,w}$) can be similarly described, although formants are excluded for non-speech models, leaving only the MFCC component.

For each state, s, of every model, w, a prior voiced probability, $P(v|s,w)$, is defined as the proportion of voiced vectors in that model and state:

$$P(v|s,w) = \frac{N_{\Omega_{s,w}}}{N_{\Omega_{s,w}} + N_{\Psi_{s,w}} + N_{\Upsilon_{s,w}}} \quad 1 \le s \le S_w, \quad 1 \le w \le W \quad (5)$$

where $N_{\Omega_{s,w}}$ is the number of voiced vectors, $N_{\Psi_{s,w}}$ the number of unvoiced vectors and $N_{\Upsilon_{s,w}}$ the number of non-speech vectors in state s of model w. Similarly, prior unvoiced and non-speech probabilities, $P(u|s,w)$ and $P(ns|s,w)$, are calculated for each state, s, of every model, w, such that $P(v|s,w) + P(u|s,w) + P(ns|s,w) = 1$. As an example, for the centre state of the phoneme /ae/, $P(v|s,w) = 0.932$, $P(u|s,w) = 0.068$ and $P(ns|s,w) = 0$.

2.2. Prediction of Voicing Class and Formant Frequencies

For an input MFCC vector stream, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_1, \ldots, \mathbf{x}_N]$, Viterbi decoding is used to obtain a model sequence, $\mathbf{m} = [m_1, m_2, \ldots, m_N]$, and state sequence, $\mathbf{q} = [q_1, q_2, \ldots, q_N]$.

Predictions of formant frequencies are made from voiced and unvoiced speech, so first the voicing of an input MFCC vector, $\mathbf{x}_i$, must be determined. The probability of a MFCC vector, $\mathbf{x}_i$, coming from voiced speech in state q_i of model m_i is calculated as:

$$P\left(v|\mathbf{x}_i, q_i, m_i\right) = \frac{P(v|q_i, m_i)\ p\left(\mathbf{x}_i|v, q_i, m_i\right)}{p\left(\mathbf{x}_i|q_i, m_i\right)} \quad (6)$$

where $P(v|q_i, m_i)$ is the prior probability of the voiced class (from equation 5), $p(\mathbf{x}_i|q_i, m_i)$ is the prior probability of vector $\mathbf{x}_i$, and $p(\mathbf{x}_i|v, q_i, m_i)$ is given by the corresponding marginalised GMM $\Phi^{v,x}_{q_i,m_i}$:

$$\begin{aligned} p(\mathbf{x}_i|v, q_i.m_i) &= \Phi^{v,x}_{q_i,m_i}(\mathbf{x}_i) = \sum_{k=1}^{K} \alpha^v_{k,q_i.m_i} \phi^{v,x}_{k,q_i,m_i}(\mathbf{x}_i) \\ &= \sum_{k=1}^{K} \alpha^v_{k,q_i.m_i} p(\mathbf{x}_i|\phi^{v,x}_{k,q_i,m_i}) \end{aligned} \quad (7)$$

where $p\left(\mathbf{x}_i|\phi^{v,x}_{k,q_i,m_i}\right)$ is the marginal distribution of the MFCC vector for the k^{th} cluster of the voiced GMM in state q_i and model m_i.

The probabilities of the input MFCC vector, $\mathbf{x}_i$, coming from unvoiced speech and non-speech are also calculated in a similar way using equations 6 and 7. An input MFCC vector is deemed to be from voiced speech, unvoiced speech or non-speech depending on the largest overall speech classification probability.

If a given input MFCC vector is predicted to be from voiced or unvoiced speech, formant frequencies are then predicted. For prediction from voiced MFCC vectors, the maximum a posteriori (MAP) [8] estimation of the i^{th} vector of formant frequencies, $\hat{\mathbf{F}}_i$, from $\mathbf{x}_i$ is given by:

$$\hat{\mathbf{F}}_i = \arg\max_{\mathbf{F}_i} \left\{ p\left(\mathbf{F}_i|\mathbf{x}_i, \Phi^v_{k,q_i,m_i}\right) \right\} \quad (8)$$

Formant frequency predictions from each cluster are weighted by the posterior probability, $h_{k,q_i,m_i}(\mathbf{x}_i)$, of the i^{th} MFCC vector $\mathbf{x}_i$, belonging to the k^{th} cluster:

$$\hat{\mathbf{F}}_i = \sum_{k=1}^{K} h_{k,q_i,m_i}(\mathbf{x}_i) \Big\{ \boldsymbol{\mu}^{v,F}_{k,q_i,m_i} + \boldsymbol{\Sigma}^{v,Fx}_{k,q_i,m_i} \left(\boldsymbol{\Sigma}^{v,xx}_{k,q_i,m_i}\right)^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}^{v,x}_{k,q_i,m_i}\right) \Big\} \quad (9)$$

The posterior probability, $h_{k,q_i,m_i}(\mathbf{x}_i)$, is given by:

$$h_{k,q_i,m_i}(\mathbf{x}_i) = \frac{\alpha^v_{k,q_i,m_i}\ p\left(\mathbf{x}_i|\phi^{v,x}_{k,q_i,m_i}\right)}{\sum_{k=1}^{K} \alpha^v_{k,q_i,m_i}\ p\left(\mathbf{x}_i|\phi^{v,x}_{k,q_i,m_i}\right)} \quad (10)$$

For formant frequency prediction from MFCC vectors deemed to be unvoiced, the set of unvoiced GMMs, $\Phi^u_{s,w}$, is used in equations 8 to 10.

A five point median filter is used to smooth each formant track by removing discontinuities. Segments of speech and non-speech are also forced to have a minimum duration of 30ms.

3. Experimental Results

The accuracy of speech class and formant frequency prediction is measured by comparison with reference formants obtained using LPC analysis followed by Kalman filtering. Evaluation measures are defined and used to investigate the accuracy of voicing class and formant frequency prediction. Results are presented for varying numbers of clusters in each GMM, followed by results for decreasing signal to noise ratio (SNR).

3.1. Database and Evaluation Measures

Male speech from a subset of the speaker-independent 5000 word WSJCAM0 large vocabulary database [9] was used to evaluate speech class and formant frequency prediction. A set of 1080 utterances (780,734 vectors) from 54 speakers was used for training and 765 utterances (517,573 vectors) from 10 different speakers were used for testing. MFCCs were extracted from the 8kHz sampled speech using 25ms frames with a 10ms shift in accordance to the ETSI Aurora standard [10].

Two error measures are used to evaluate prediction accuracy. Speech classification error, E^c, indicates the percentage of frames where incorrect voicing prediction either leads to formant frequency predictions being made from non-speech or no predictions made during speech:

$$E^c = \frac{N_{v|ns} + N_{u|ns} + N_{ns|v} + N_{ns|u}}{N_{total}} \times 100\% \qquad (11)$$

where $N_{v|ns}$ and $N_{u|ns}$ are the number of non-speech frames classified as voiced and unvoiced, $N_{ns|v}$ and $N_{ns|u}$ are the number of voiced and unvoiced frames classified as non-speech respectively and N_{total} is the total number of frames in the test data. For frames classed as voiced or unvoiced, separate percentage formant frequency errors are calculated as the mean formant frequency prediction error across all four formants:

$$E^p = \frac{1}{4}\sum_{j=1}^{4}\frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{\mathbf{F}}_i(j) - \mathbf{F}_i(j)}{\mathbf{F}_i(j)}\right| \times 100\% \qquad (12)$$

where j denotes formant number and N is the number of frames for which the speech class was predicted as voiced or unvoiced.

For certain state and model combinations, there were insufficient data to successfully create GMMs with more than one cluster. For example, it was not possible to create voiced GMMs with more than one cluster for the unvoiced phoneme /s/ as only a few voiced frames were recognised as /s/ by the recogniser at the boundary between /s/ and voiced phonemes. Therefore the number of clusters, K, in the results refers to the requested number of clusters for a particular GMM. During prediction, if a GMM with the requested number of clusters was not available, the GMM with the next highest number of clusters was used.

One of the problems in evaluating formant frequency estimation techniques is the lack of large speech corpora containing baseline formants such as those from hand-labelled spectrograms. In order to provide further evidence of the effectiveness of HMM-GMM formant frequency from MFCC vectors, formant frequencies are also predicted from the means of the GMMs to provide a worst case scenario of prediction. Formant frequency prediction from the means of the GMMs makes no use of input MFCC vectors, so is based entirely on prior statistical knowledge. For an input MFCC predicted as voiced, predicting formant frequencies from the means of the GMMs is performed by setting $\mathbf{x}_i$ to be equal to the mean of the MFCCs for the k^{th} cluster of the appropriate state-specific voiced GMM, $\boldsymbol{\mu}_{k,}^{v,x}$, in equation 9 and by setting $p(\mathbf{x}_i|\phi_k^{v,x}) = 1$. This reduces equation 9 to:

$$\hat{\mathbf{F}}_i = \sum_{k=1}^{K} \alpha_{k,q_i,m_i}^{v}\, \boldsymbol{\mu}_{k,q_i,m_i}^{v,F} \qquad (13)$$

The procedure is the same for MFCCs predicted as unvoiced, except that unvoiced GMMs are used and $p(\mathbf{x}_i|\phi_k^{u,x})$ is set to 1.

3.2. Prediction Accuracy using Clean Speech

First, prediction accuracy was investigated as the number of clusters in each state-specific GMM in the HMM framework was varied. Table 1 shows the speech classification confusion matrix. About 20% of frames deemed to be non-speech according to the reference speech classification are wrongly predicted as unvoiced. This is mainly due to recognition errors, typically at monophone boundaries, from the recogniser used to obtain model and state sequences. The recogniser used a simple unconstrained grammar and has a monophone recognition accuracy of 56%. There is little confusion between non-speech and voiced speech because of the large difference in energy level between these segements of speech.

		Predicted		
		non-speech	unvoiced	voiced
Correct	non-speech	0.7894	0.1980	0.0126
	unvoiced	0.0074	0.8455	0.1471
	voiced	0.0016	0.1072	0.8911

Table 1: Confusion matrix for speech class prediction

Figure 1a presents speech class prediction error, E^c, as the number of clusters in the state-specific GMMs is varied. The number of clusters in each state-specific GMM has very little effect on speech class prediction error, which remains at 5.16%. This indicates that speech classification, which is largely based on energy, is described by a relatively simple distribution, easily modelled with one cluster for each state-specific GMM.

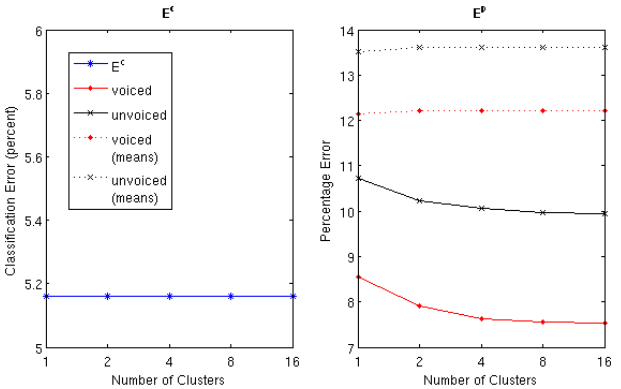

Figure 1: a) speech classification error (E^c) and b) mean percentage error (E^p) with increasing numbers of clusters

Figure 1b shows formant frequency prediction error measures divided according to whether MFCC vectors were classified as voiced or unvoiced. The combined voiced and unvoiced formant frequency prediction error results (not plotted) would be the weighted combination of the voiced and unvoiced plots. Both voiced and unvoiced formant frequency prediction errors decrease

with increasing numbers of clusters due to the improved modelling of the MFCC and formant frequency distributions. The greatest improvement in formant frequency prediction accuracy occurs when the number of clusters is increased from one to two. Formant frequency prediction is more accurate for frames predicted as voiced than those predicted as unvoiced because of the clearer formant structure in voiced speech.

Comparing MAP prediction of formant frequencies with prediction given as the means of the GMMs (indicated by the dotted lines in figure 1b) shows that prediction from just the means of the GMMs is less accurate, as expected. This confirms that there is sufficient information described by the MFCC vectors to allow the prediction of formant frequencies.

3.3. Prediction Accuracy using Noisy Speech

In this section, speech class and formant frequency prediction errors are presented for prediction in the presence of noise. 'Exhibition hall' noise, extracted from the ETSI Aurora database [10] was added to clean speech at SNRs from 20dB to −5dB. This noise was chosen as the contaminant because it contains competing speakers and noises with narrow high-energy frequency bands which resemble formants. Prediction in noise was carried out using unmatched condition testing: the system was trained using clean speech, but tested with noisy speech, causing a mismatch between the clean speech models and noisy test data. Figure 2a shows that as the SNR decreases, speech class prediction error increases due to the variability introduced by the noise and decreasing state sequence accuracy obtained through Viterbi decoding. The unconstrained monophone recognition accuracy of the Viterbi decoding is 56% using clean speech, but falls to 17% at −5dB.

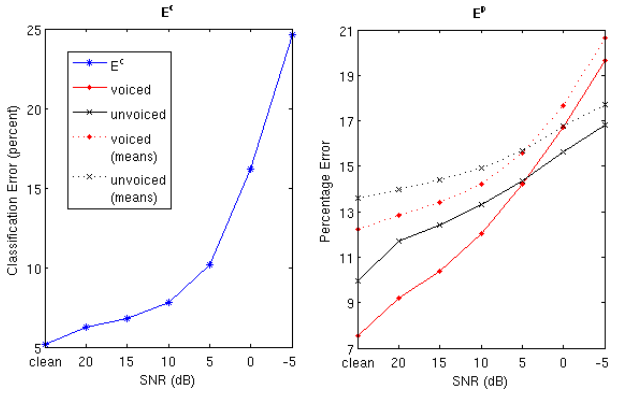

Figure 2: a) speech classification error (E^c) and b) mean percentage error (E^p) with decreasing SNR

Figure 2b shows that at SNRs greater than about 5dB, the percentage formant frequency prediction error, E^p, is lower for frames predicted as voiced, rather than unvoiced speech. E^p increases more rapidly with increasing noise for frames predicted as voiced. This is partly attributed to the increasing number of voiced frames incorrectly predicted as unvoiced which count as unvoiced frames for formant frequency prediction error measures, despite being from voiced speech, where formants are more clearly defined. In figure 2b, as noise increases, prediction error from the means of the GMMs also increases due to the greater variability of the noisy MFCC vectors which corrupts the state sequence and hence state-specific mean GMM values. Even at −5dB, MAP prediction of formant frequencies is more accurate than prediction from the means of the GMMs. The figure shows that the plots for MAP prediction and prediction from the means begin to converge as the SNR decreases. It is expected that at some SNR below −5dB, MAP prediction of formant frequencies will be no better than prediction from the means of the GMMs as noise dominates the signal.

4. Conclusions

This work has shown how it is possible to predict speech class and formant frequencies (for voiced and unvoiced speech) from MFCC vectors. This allows formants to be obtained in a distributed speech recognition environment, where the time-domain waveform is unavailable and the spectrum obtained through inverting MFCC vectors is too crude for traditional formant estimation techniques.

5. Acknowledgements

The authors gratefully acknowledge the support of the UK Engineering and Physical Sciences Research Council (EPSRC).

6. References

[1] R. W. Schafer and L. R. Rabiner, "System for automatic formant analysis of voiced speech," *Journal of the Acoustical Society of America*, vol. 47, no. 2, pp. 634–648, Feb. 1970.

[2] S. McCandless, "An algorithm for automatic formant extraction using linear prediction spectra," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 22, no. 2, pp. 135–141, Apr. 1974.

[3] B.S. Atal and S.L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *Journal of the Acoustical Society of America*, vol. 50, no. 2, pp. 637–655, Aug. 1971.

[4] J. Darch, B. Milner, and S. Vaseghi, "Formant frequency prediction from MFCC vectors in noisy environments," in *Eurospeech*, Lisbon, Portugal, Sept. 2005.

[5] B. Milner, X. Shao, and J. Darch, "Fundamental frequency and voicing prediction from MFCCs for speech reconstruction from unconstrained speech," in *Eurospeech*, Lisbon, Portugal, Sept. 2005.

[6] Q. Yan, E. Zavarehei, S. Vaseghi, and D. Rentzos, "A formant tracking LP model for speech processing in car/train noise," in *ICSLP*, Jeju, Korea, Oct. 2004.

[7] A. Sorin and T. Ramabadran, "Extended Advanced Front End Algorithm Description, Version 1.1," Tech. Rep. ES 202 212, ETSI STQ-Aurora DSR Working Group, Apr. 2003.

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

[9] J. Fransen, D. Pye, T. Robinson, P. Woodland, and S. Young, "WSJCAM0 corpus and recording description," Tech. Rep. CUED/F-INFENG/TR.192, Cambridge University Engineering Department, Sept. 1994.

[10] D. Pearce and H.-G. Hirsch, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ICSLP*, Beijing, China, Oct. 2000.