Robust Bandwidth Extension of Noise-corrupted Narrowband Speech

Michael L. Seltzer, Alex Acero, and Jasha Droppo

Microsoft Research
Redmond, WA 98052
{mseltzer,alexac,jdroppo}@microsoft.com

Abstract

We present a new bandwidth extension algorithm for converting narrowband telephone speech into wideband speech using a transformation in the mel cepstral domain. Unlike previous approaches, the proposed method is designed specifically for bandwidth extension of narrowband speech that has been corrupted by environmental noise. We show that by exploiting previous research in mel cepstrum feature enhancement, we can create a unified probabilistic framework under which the feature denoising and bandwidth extension processes are tightly integrated using a single shared statistical model. By doing so, we are able to both denoise the observed narrowband speech and robustly extend its bandwidth in a jointly optimal manner. A series of experiments on clean and noise-corrupted narrowband speech is performed to validate our approach.

1. Introduction

Speech transmitted over the telephony network is bandlimited to frequencies between 300-3400 Hz. While limiting speech to this bandwidth does not significantly reduce intelligibility, studies have shown that users prefer listening to wideband speech, i.e. speech with a frequency range of 50-8000 Hz [1]. As a result, there has been a significant amount of research performed recently aimed at enhancing the perceptual quality of narrowband speech by estimating and then synthesizing the missing spectral content in order to artificially extend its bandwidth, e.g. [2, 3].

Because the accuracy of the recovered spectral envelope is particularly important to its subjective quality, most bandwidth extension (BWE) research has focused on this problem. Almost all methods in the literature operate using a codebook or a statistical model, e.g. GMM or HMM, to model different sound classes of narrowband speech. Each codeword (or state) has an associated template representing the average spectral envelope of the missing frequencies for that sound class. Such methods operate by first estimating the mostly likely state or codeword for a particular speech frame, and then selecting (in a hard or soft manner) the corresponding extended frequency template.

While these methods are able to perform BWE under ideal conditions, no effort has been made to address the problem of performing such procesing in the presence of additive noise. Yet, as the number of mobile phone users continue to grow and people make calls from a variety of environments, it is essential that BWE algorithms perform robustly in noisy environments.

Performing BWE in noisy enviroments is problematic for several reasons. For example, most BWE algorithms utilize LPC-derived features, such as LPC-cepstra or LSF coefficients, to represent both the narrowband and the extended frequency spectral envelopes in the codebooks or mixture models. However, because additive noise introduces zeros in the speech spectrum, noise-corrupted speech is no longer well-represented by an all-pole model, and as a result, the BWE accuracy will degrade. In addition, because there is no way of modeling the effect of additive noise on LPC-based features directly, the noisy signal must be pre-processed using a speech enhancement algorithm, prior to BWE. However, there is no way to know if the enhancement processing is in fact optimal for the subsequent BWE processing.

In this paper, we propose a new statistical method for spectral envelope extension using mel frequency cepstral coefficients (MFCC). Because the mel cepstra are derived directly from the power spectrum and not an all-pole spectral model, the relationship between speech and noise features is well-understood. Indeed, much research in the speech recognition community has focused on removing the effects of additive noise from MFCC features. We exploit this work in order to create a new BWE algorithm for noisy speech in which narrowband envelope denoising and BWE are integrated in a unified probabilistic framework using a common statistical model. We then demonstrate how such an approach significantly outperforms the more conventional approach of enhancing the noise-corrupted speech and then performing BWE on the denoised speech in a variety of environments and SNRs.

2. MFCC-based bandwidth extension

In this section, we present a new algorithm for estimating the wideband spectral envelope from a narrowband observation using a transformation in the mel cepstral domain. Throughout this work, we assume that the narrowband speech signal has been upsampled to match the sampling rate of the desired wideband speech.

2.1. Extracting MFCCs from wideband and narrowband speech

If we define $|Z|^2$ as the power spectrum of a frame of speech derived from a Short-Time Fourier Transform, the feature extraction process for generating MFCCs can be summarized as

$$\mathbf{z} = \mathbf{C}\log(\mathbf{W}|Z|^2) \tag{1}$$

where $\mathbf{W}$ is the matrix of weighting coefficients of the mel filterbank and $\mathbf{C}$ is a DCT matrix. If the mel filters in $\mathbf{W}$ span the wideband spectrum, i.e. 50-8000 Hz, then for narrowband speech, we can use a reduced-row version of $\mathbf{W}$ which only includes the mel filters between 300-3400 Hz. Of course, this requires that a reduced-column DCT matrix be used as well.

2.2. Narrowband-to-wideband MFCC transformation

We now define $\mathbf{x}$ to be a narrowband MFCC feature vector and $\mathbf{z}$ to be the corresponding wideband MFCC feature vector. Our goal is to estimate $\mathbf{z}$ from $\mathbf{x}$, i.e. to compute $E[\mathbf{z}|\mathbf{x}]$.

We assume that the observed narrowband feature vectors were generated by a Gaussian mixture model (GMM). Thus, the probability distribution $p(\mathbf{x})$ can be written as

$$p(\mathbf{x}) = \sum_{s=1}^{S} p(\mathbf{x}|s)p(s) = \sum_{s=1}^{S} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)\, p(s) \quad (2)$$

where $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$ is a Gaussian distribution with mean $\boldsymbol{\mu}_s$ and covariance $\boldsymbol{\Sigma}_s$, $p(s)$ is the prior probability for state s, and S is the total number of Gaussians in the mixture. We assume that $\boldsymbol{\Sigma}_s$ is a diagonal matrix. This model is trained from narrowband training data using conventional EM.

We model the transformation from narrowband to wideband feature vectors using a piecewise-linear transformation. More specifically, we assume that

$$\mathbf{z} = \mathbf{A}_s\mathbf{x} + \mathbf{b}_s + \mathbf{e} \quad (3)$$

where $p(\mathbf{e}) = \mathcal{N}(\mathbf{e}; 0, \mathbf{I})$, and the transformation parameters $\{\mathbf{A}_s, \mathbf{b}_s\}$ are dependent on the Gaussian mixture component s.

From Eq. (3), the conditional probability of $\mathbf{z}$ is

$$p(\mathbf{z}|\mathbf{x}, s) = \mathcal{N}(\mathbf{z}; \mathbf{A}_s\mathbf{x} + \mathbf{b}_s, \mathbf{I}) = \mathcal{N}\left(\mathbf{z}; \mathbf{A}'_s\mathbf{x}', \mathbf{I}\right) \quad (4)$$

where $\mathbf{A}'_s = [\mathbf{A}_s\ \mathbf{b}_s]$, and $\mathbf{x}' = [\mathbf{x}\ 1]^T$.

The transformation parameters $\{\mathbf{A}'_1 \ldots \mathbf{A}'_S\}$ are learned using a corpus of stereo training data in which each narrowband feature vector $\mathbf{x}_t$ has a corresponding wideband feature vector $\mathbf{z}_t$. For each state s, the ML estimate of $\mathbf{A}'_s$ is given by

$$\mathbf{A}'_s = \left(\sum_{t=1}^{T} p(s|\mathbf{x}_t)\mathbf{z}_t\mathbf{x}_t'^T\right)\left(\sum_{t=1}^{T} p(s|\mathbf{x}_t)\mathbf{x}'_t\mathbf{x}_t'^T\right)^{-1} \quad (5)$$

The derivation, which is quite straightforward, is omitted here for brevity. Eq. (5) can be interpreted as a soft-decision least-squares estimate of $\mathbf{A}'_s$.

2.3. Bandwidth extension of the spectral envelope

Using Eq. (2) and Eq. (4), we can compute $E[\mathbf{z}|\mathbf{x}]$ as

$$\begin{aligned} E[\mathbf{z}|\mathbf{x}] &= \sum_{s=1}^{S} \int \mathbf{z}p(\mathbf{z}, s|\mathbf{x})d\mathbf{z} \\ &= \sum_{s=1}^{S} p(s|\mathbf{x}) \int \mathbf{z}p(\mathbf{z}|\mathbf{x}, s) \\ &= \sum_{s=1}^{S} p(s|\mathbf{x})\mathbf{A}'_s\mathbf{x}' \end{aligned} \quad (6)$$

where the state posterior probability $p(s|\mathbf{x})$ is computed from the GMM as

$$p(s|\mathbf{x}) = \frac{p(\mathbf{x}|s)p(s)}{\sum_{s'=1}^{S} p(\mathbf{x}|s')p(s')} \quad (7)$$

Once the wideband cepstral vector $\hat{\mathbf{z}} = E[\mathbf{z}|\mathbf{x}]$ has been estimated, we can perform a series of inverse transformations to generate the corresponding wideband spectral envelope. If we define S_Z to be the smooth spectral envelope corresponding to a power spectrum $|Z|^2$, then we can generate an estimate the wideband envelope as

$$\hat{S}_Z = \mathbf{W}^\dagger \exp(\mathbf{C}^\dagger\hat{\mathbf{z}}) \quad (8)$$

where $\mathbf{W}^\dagger$ and $\mathbf{C}^\dagger$ are the pseudoinverses of $\mathbf{W}$ and $\mathbf{C}$, respectively. The segment or segments of the spectrum missing from the original narrowband speech can then be extracted from $\hat{S}_Z$.

Note that while Eq. (8) generates the entire wideband spectral envelope, we can easily extract the spectral envelope for only those frequencies we require by selecting appropriate partitions of $\mathbf{C}^\dagger$ and $\mathbf{W}^\dagger$.

3. Bandwidth extension of noise-corrupted observations

In this section, we describe how the BWE algorithm described in the previous section can be extended to perform BWE on noise-corrupted speech.

3.1. An overview of MFCC feature enhancement

If we assume that the speech and noise are uncorrelated, then the noise and speech are additive in power spectrum domain as

$$|Y|^2 = |X|^2 + |N|^2 + \epsilon \quad (9)$$

where ϵ is a zero-mean random variable that models the contribution of the cross-terms. If we assume ϵ can be neglected, it can be shown through algebraic manipulation [4] that in the MFCC domain, this relationship becomes

$$\begin{aligned} \mathbf{y} &= \mathbf{x} + \mathbf{C}\log(1 + \exp(\mathbf{C}^\dagger(\mathbf{n} - \mathbf{x}))) \\ &= \mathbf{x} + \mathbf{f}(\mathbf{x}, \mathbf{n}) \end{aligned} \quad (10)$$

where $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{n}$ are the MFCC vectors for the clean speech, noise-corrupted speech, and the noise, respectively. Thus, the noisy cepstrum $\mathbf{y}$ is related to the clean cepstrum $\mathbf{x}$ by an additive bias $\mathbf{f}$ which is a non-linear function of the speech and noise cepstra.

The relationship defined in Eq. (10) has been widely used in the speech recognition community as the basis of a family of feature enhancement (FE) algorithms, e.g. [4, 5]. While the details of these techniques vary, the majority operate under an EM framework utilizing a prior speech model, e.g. a GMM or HMM, and an iterative Vector Taylor Series (VTS) approximation algorithm. At each iteration, a VTS expansion is used to linearize $\mathbf{f}$, and then compute the ML estimate of the posterior distributions of the hidden variables $\mathbf{x}$ and $\mathbf{n}$. The means of these distributions are then used as the VTS expansion point for the next iteration. The process is repeated until convergence. At this point, a MMSE estimate of the clean cepstral vector is generated as

$$E[\mathbf{x}|\mathbf{y}] = \sum_{s=1}^{S} p(s|\mathbf{y}) \int \mathbf{x}p(\mathbf{x}|\mathbf{y}, s)d\mathbf{x} \quad (11)$$

where $p(s|\mathbf{y})$ is the state posterior distribution, and $p(\mathbf{x}|\mathbf{y}, s)$ is a Gaussian that represents the state-conditional posterior distribution of the clean cepstra. While the details of these algorithms are beyond the scope of this paper, we will demonstrate how the distributions in Eq. (11) can be used to tightly integrate any of these algorithms with the BWE algorithm described in the previous section. The will enable us to perform robust bandwidth extension on noise-corrupted speech.

3.2. Integrating BWE with feature enhancement

We would like to estimate the clean wideband cepstral vector $\mathbf{z}$ from a noisy narrowband vector $\mathbf{y}$. We will show how using the same GMM for *both* feature enhancement and bandwidth extension enables us to do so. We refer to this method as Feature-Enhanced BWE (FE-BWE). Using the GMM in Eq. (2), we can express the MMSE estimate as

$$E[\mathbf{z}|\mathbf{y}] = \sum_{s=1}^{S} \int \mathbf{z} \left(\int p(\mathbf{z}, \mathbf{x}, s|\mathbf{y}) d\mathbf{x} \right) d\mathbf{z} \quad (12)$$

Notice that rather than relying on a point estimate of the narrowband clean cepstral vector $\hat{\mathbf{x}}$, we marginalize over all values of $\mathbf{x}$. This will make the solution much more robust to estimation errors. Using Bayes' rule, we can rewrite Eq. (12) as

$$E[\mathbf{z}|\mathbf{y}] = \sum_{s=1}^{S} p(s|\mathbf{y}) \int \mathbf{z} \left(\int p(\mathbf{z}, \mathbf{x}|\mathbf{y}, s) d\mathbf{x} \right) d\mathbf{z} \quad (13)$$

$$= \sum_{s=1}^{S} p(s|\mathbf{y}) \int \mathbf{z} p(\mathbf{z}|\mathbf{y}, s) d\mathbf{z} \quad (14)$$

All that remains is to compute the state posterior distribution $p(s|\mathbf{y})$ and the state-conditional distribution $p(\mathbf{z}|\mathbf{y}, s)$ in Eq. (14). Because both feature enhancement and bandwidth extension are performed with the same GMM, it is clear from Eq. (11) that $p(s|\mathbf{y})$ can be obtained directly from the output of the feature enhancement algorithm. To estimate the parameters of $p(\mathbf{z}|\mathbf{y}, s)$, we first rewrite the joint posterior probability of $\mathbf{z}$ and $\mathbf{x}$ in Eq. (13) as

$$p(\mathbf{z}, \mathbf{x}|\mathbf{y}, s) = p(\mathbf{z}|\mathbf{x}, \mathbf{y}, s) p(\mathbf{x}|\mathbf{y}, s) \quad (15)$$

$$= p(\mathbf{x}|\mathbf{z}, \mathbf{y}, s) p(\mathbf{z}|\mathbf{y}, s) \quad (16)$$

The first term on the right side of Eq. (15) can be simplified to $p(\mathbf{z}|\mathbf{x}, s)$ because given $\mathbf{x}$, $\mathbf{y}$ provides no additional information about $\mathbf{z}$. Thus, this is simply the wideband cepstra conditional probability distribution in Eq. (4). The second term $p(\mathbf{x}|\mathbf{y}, s)$ is the state-conditional posterior distribution generated by the feature enhancement algorithm, as shown in Eq. (11). Since both of these terms are Gaussian, the two expressions in Eq. (16) must also be Gaussian. We can therefore find their parameters through algebraic manipulation. Thus, it can be shown that if the posterior distribution of $\mathbf{x}$ in Eq. (11) is expressed as

$$p(\mathbf{x}|\mathbf{y}, s) = \mathcal{N}(\mathbf{x}; \boldsymbol{\nu}_s, \boldsymbol{\Phi}_s) \quad (17)$$

then $p(\mathbf{z}|\mathbf{y}, s)$ can be expressed as

$$p(\mathbf{z}|\mathbf{y}, s) = \mathcal{N}\left(\mathbf{z}; \mathbf{A}'_s \boldsymbol{\nu}'_s, \mathbf{A}_s^T \boldsymbol{\Phi}_s \mathbf{A}_s + \mathbf{I}\right) \quad (18)$$

Finally, by substituting Eq. (18) into Eq. (14), we can now write the final expression for the expected value as

$$\hat{\mathbf{z}} = E[\mathbf{z}|\mathbf{y}] = \sum_{s=1}^{S} p(s|\mathbf{y}) \mathbf{A}'_s \boldsymbol{\nu}'_s \quad (19)$$

The extended spectral envelope $\hat{S}_Z$ can then be generated from $\hat{\mathbf{z}}$ as described in Section 2.3.

3.3. Noise reduction of the narrowband envelope

As discussed in Section 2.3, $\hat{\mathbf{z}}$ represents the entire spectral envelope, not just the extended frequency segment. As a result, it can be used to denoise the observed narrowband speech in the following manner. From $\mathbf{y}$ and $\hat{\mathbf{z}}$, we can generate the the noisy and estimated clean spectral envelopes, S_Y and $\hat{S}_Z$, respectively, using Eq. (8). The narrowband frequencies can then be extracted from these envelopes to create a Wiener filter to denoise the narrowband speech. This filter can be expressed as

$$H = \hat{S}_Z / S_Y \quad (20)$$

We can now apply this filter to enhance the power spectrum of the noisy narrowband speech as follows.

$$|\hat{Z}|^2 = H|Y|^2 \quad (21)$$

4. Generating a wideband waveform

Once the wideband speech envelope has been estimated, we can generate the wideband waveform using a conventional BWE approach. We used the method proposed in [2]. The denoised and extended power spectrum is converted to the LPC domain. The denoised narrowband signal is then passed through the LPC filter to obtain the narrowband excitation signal which is then modulated, high-pass filtered, and then combined with the original excitation. This now-wideband excitation is then used to drive the LPC synthesis filter to generate the wideband speech. Other methods of generating bandwidth extended speech from a given extended spectral envelope estimate may also be used.

5. Experimental results

To test the effectiveness of the proposed BWE algorithm, we performed a series of experiments on clean and noise-corrupted narrowband speech. To train the GMM and the narrowband-to-wideband transformation parameters, we utilized the training set from the WSJ0 corpus [6], which consists of 12 hours of wideband clean speech from 84 different speakers. To create a parallel narrowband corpus, the speech was downsampled to 8 kHz, filtered according to the G.712 telephony channel specification, and then upsampled back to 16 kHz. Wideband features were created by extracting 64-dim log mel spectral vectors from the power spectrum, and then converting these to 19-dim cepstra, including c0. The narrowband cepstral vectors were created by extracting 45-dim log mel spectra (using mel filters 4-48) and then converting these to 13-dim cepstra, including c0. We used feature vectors with higher dimensionality than typically used in feature extraction for speech recognition in order to retain more detail in the spectral envelope. The narrowband cepstra were used to train a GMM with 256 densities using conventional EM. The transformation parameters were then trained according to Eq. (5).

We first evaluated the proposed BWE algorithm on clean narrowband speech. A test corpus of telephone speech was created from 112 utterances selected at random from the WSJ0 test set. There were 8 speakers (5 male, 3 female) with 14 utterances each. These utterances were converted to narrowband speech in the manner described above. For each utterance, BWE was performed to estimate the spectral envelope for both the low and high frequency regions missing from telephone speech.

To evaluate the performance of the algorithm, we computed the mean RMS log spectral distortion (RMS-LSD) of the smooth power spectral envelopes between the original wideband speech S_Z and the bandwidth-extended speech $\hat{S}_Z$ over the extended frequencies. RMS-LSD is defined as

$$D = \sqrt{\frac{1}{\omega_2 - \omega_1} \int_{\omega_1}^{\omega_2} \left| 10 \log_{10} |S_Z / \hat{S}_Z| \right|^2 d\omega} \quad (22)$$

where $\{\omega_1, \omega_2\}$ represents the frequency range of interest. The results are shown in Table 1. Although differences in test data make direct comparisons difficult, the performance of the proposed MFCC-based BWE algorithm on clean speech is comparable to that of other state-of-the-art algorithms in the literature. The table also shows that applying FE-BWE on clean speech results in a negligible change in performance.

RMS Log Spectral Distortion (dB)	Low Freq (50-300 Hz)	High Freq (3.4-8 kHz)
BWE	3.39	7.30
FE-BWE	3.49	7.41

Table 1: Spectral distortion of the extended spectral envelopes obtained using the proposed BWE and FE-BWE algorithms on clean narrowband speech.

To evaluate the FE-BWE algorithm on noisy speech, the narrowband test set was mixed with samples of noise from the Aurora 2 corpus [7]. The noises represented 8 environments: airport, babble, car, exhibition hall, restaurant, street, subway, and train. For each environment, test sets were created at SNRs between 0 and 25 dB.

We compared three different methods of BWE on noisy speech. First, BWE was performed directly on the noisy speech. In the second case, the speech was first enhanced using conventional Wiener filtering, followed by BWE. This is equivalent to performing BWE on a point estimate of the clean speech. Finally, we performed the proposed FE-BWE algorithm. We used the Zero Variance Model (ZVM) feature enhancement algorithm [5] to generate the required posterior distributions. These distributions were then used in Eq. (19) to estimate the clean wideband cepstral vector and the resulting spectral envelope. The segment of the envelope spanning the narrowband frequencies was used to denoise the narrowband envelope and the portion that lay outside this region was used as the estimate of the extended envelope. The resulting performance, averaged over all noise environments, is shown shown as a function of SNR in Figures 1 and 2. Figure 1 shows the spectral distortion of the denoised narrowband envelope while Figure 2 shows that of the high frequency extended envelope.

As the figures show, enhancing the speech prior to performing BWE results in minimal improvement over BWE processing on the noisy speech directly. On the other hand, the proposed FE-BWE algorithm results in significantly less distortion in both the observed narrowband spectral envelope and the extended high frequency envelope. Similar performance was seen in the extension of the low frequencies as well.

6. Conclusion

In this paper, we have presented a new algorithm for robust bandwidth extension of narrowband speech that has been corrupted by additive noise. The bandwidth extension algorithm utilizes a GMM trained on narrowband speech and a state-conditional affine transformation in the MFCC domain to transform the narrowband spectral envelope into a wideband spectral envelope. The accuracy of the proposed algorithm on clean narrowband speech is comparable to that of other state-of-the-art BWE algorithms. Unlike other methods, however, we showed how this algorithm can be tightly integrated with an MFCC-based feature enhancement algorithm using a common speech model and a unified statistical framework. Through a series of experiments on noise-corrupted narrowband speech, we showed that the proposed feature-enhanced bandwidth extension (FE-BWE) algorithm significantly outperforms a more conventional enhance-then-extend approach. A true noise-robust BWE algorithm must also robustly handle the effects of additive noise in the excitation signal. This is the focus of our future work.

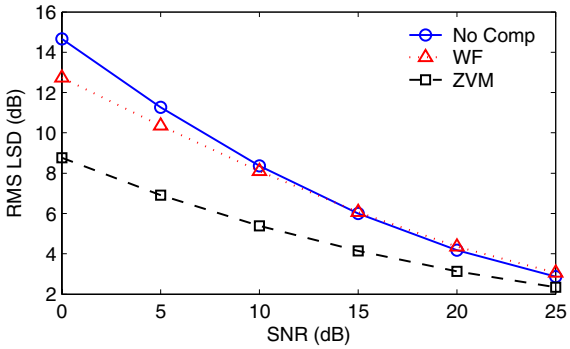

Figure 1: Spectral distortion of the observed narrowband (300-3400 Hz) envelope vs. SNR obtained with no compensation, Wiener filtering and ZVM feature enhancement.

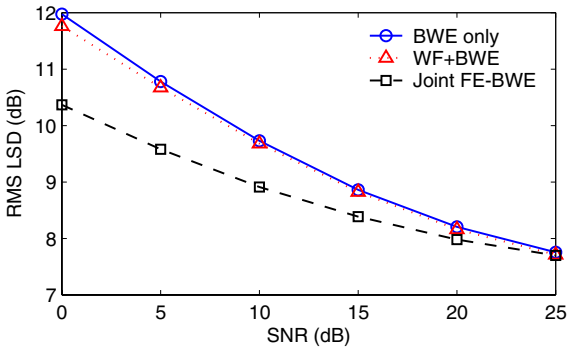

Figure 2: Spectral distortion of the extended high frequency (3.4-8 kHz) envelope vs. SNR obtained from BWE without preprocessing, Wiener filtering then BWE, and the proposed joint FE-BWE algorithm.

7. References

[1] ITU, "Paired comparison test of wideband and narrowband telephony," Tech. Rep. COM 12-9-E, ITU, Mar. 1993.

[2] P. Jax and P. Vary, "On artificial bandwidth extension of telephone speech," *Signal Processing*, vol. 83, no. 8, pp. 1707–1719, Aug. 2003.

[3] Y. Qian and P. Kabal, "Combining equalization and estimation for bandwidth extension of narrowband speech," in *Proc. ICASSP*, Montreal, May 2004.

[4] P. J. Moreno, B. Raj, and R. M. Stern, "A Vector Taylor Series approach to environment-independent speech recognition," in *Proc. ICASSP*, Atlanta, GA, May 1996.

[5] J. Droppo, L. Deng, and A. Acero, "A comparison of three non-linear observation models for noisy speech features," in *Proc. Eurospeech*, Geneva, Sept. 2003.

[6] D. B. Paul and J. M. Baker, "The design of the Wall Street Journal-based CSR corpus," in *Proc. ARPA Speech and Nat. Lang. Workshop*, Harriman, NY, Feb. 1992, pp. 357–362.

[7] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *ISCA ITRW ASR2000*, Paris, France, Sept. 2000.