COMBINING FRONTEND-BASED MEMORY WITH MFCC FEATURES FOR BANDWIDTH EXTENSION OF NARROWBAND SPEECH

Amr H. Nour-Eldin and Peter Kabal

Department of Electrical & Computer Engineering
McGill University, Montréal, Québec, Canada
amr.nour-eldin@mail.mcgill.ca, peter.kabal@mcgill.ca

ABSTRACT

In this paper, we continue our previous work on improving Bandwidth Extension (BWE) of narrowband speech. We have shown that including memory into the parametrization frontend (through delta features) results in higher *highband certainty* irrespective of feature type, with MFCCs exhibiting higher correlation, in general, between both bands, reaching twice that using LSFs. By incorporating memory into the frontend of a conventional LP-based BWE system, we were able to translate the higher correlation due to memory into BWE performance improvement. Using high-resolution inverse DCT, we also achieved high quality speech reconstruction from MFCCs, thus enabling MFCC-based BWE with improved performance compared to conventional static LP-based BWE. We continue this work by incorporating the superior correlation properties of frontend memory into our MFCC-based BWE system. Log-Spectral Distortion as well as the more perceptually-correlated Itakura-based measures show that incorporating memory into our MFCC-based BWE system results in BWE performance superior to that of our dynamic LP-based BWE system.

Index Terms— Bandwidth extension, memory inclusion, high-resolution IDCT, highband certainty, mutual information

1. BACKGROUND

In traditional telephone networks, speech bandwidth is limited to the 0.3–3.4 kHz range. As a result, narrowband speech has sound quality inferior to its wideband counterpart and has reduced intelligibility especially for consonant sounds. Wideband speech reconstruction through Bandwidth Extension (BWE) attempts to regenerate the highband (3.4–7 kHz) signal lost during the filtering processes employed in traditional networks, thereby providing backward compatibility with existing networks. BWE is based on the assumption that narrowband speech correlates with the highband signal, and thus, given some a priori information about the nature of this correlation, the higher frequency speech content can be estimated given only the available narrow band. Most BWE schemes use either codebook mapping or statistical modelling to perform this estimation.

Since BWE performance closely follows the correlation available between representations of the narrow and high frequency bands, the premise of our work has been to quantify this correlation for different speech representations in order to adopt those representations with the greatest potential for BWE performance improvement. In our previous work; first introduced in [1] and later extended in [2], we made use of the concept of *highband certainty* (certainty about the high band given the narrow band); defined in [3] as the ratio of Mutual Information (MI) between the two bands to the discrete entropy of the high band, in order to quantify the correlation between speech frequency bands. Through highband certainty, we investigated the effect of including memory into the frontend on the resulting correlation (by using *delta* features in addition to the conventional static features which make no use of the considerable temporal correlation properties of speech), as well as the effect of the type of parametrization. By varying the number of static feature vectors involved in the estimation of the delta features, we have shown that frontend-based memory inclusion can increase certainty about the highband by as much as 86% for Mel-Frequency Cepstral Coefficients (MFCCs), and 207%[1] for Line Spectral Frequencies (LSFs), with no increase in dimensionality. By further incorporating frontend-based memory inclusion into an LP-based *dual-mode* BWE system based on that of [5], we were able to translate these highband certainty gains into practical BWE performance improvement. Objective analysis of the reconstructed speech quality through log-Spectral Distortion (d_{LSD}) showed that memory inclusion decreases the d_{LSD} of the extended highband speech (versus that obtained by BWE with conventional static features) by an average of $\approx 5\%$[2].

Traditionally, BWE techniques have used LP-based representations of speech spectra since reconstruction of the missing highband signal then becomes a straightforward problem given a highband excitation signal and a set of highband LP-based features. We showed, however, in [2], that for similar dimensionalities, MFCCs result in highband certainties that can reach almost twice as those resulting from the LP-based LSFs. These results agree with the findings of [6] which show MFCCs to have the highest speech class separability and second highest MI content among several speech parametrizations. In addition to their superior correlation properties compared to LP-based parameters, MFCCs have the quite important advantage of higher robustness (with the implementation of MFCC denoising techniques) to the various types of acoustic (additive) and channel (convolutional) noises. The all-pole spectral representation of LP-based parameters becomes ill-suited to noise-corrupted spectra since such spectra will suffer from zeroes introduced by noise. Therefore, LP-based parameters can not provide any robustness in practical noisy environments, resulting in degraded BWE performance, unless the narrowband speech input is pre-processed by speech enhancement algorithms. In contrast, MFCCs are derived directly from speech spectra, and hence, the effects of time-domain additive and convolutional noises on MFCCs are well-understood. Consequently,

[1] This figure, which differs from that estimated in [2], was obtained by better implementation of the LBG training algorithm in the estimation of the discrete highband entropies (shown in Table 1 of [4]).

[2] The d_{LSD} results reported in [2] suffer from gain mismatches between the reference wideband test waveforms and those obtained by BWE. By normalizing the BWE-reconstructed waveforms based on the energy in the original 0.3–3.0 kHz range, we obtain the more accurate average d_{LSD} decrease above.

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

there has been ample research on removing the effects of noise from MFCCs, particularly in the field of automatic speech recognition (ASR). As a result, MFCC denoising techniques that are successful with more types of noise and at more adverse conditions than speech enhancement pre-processing is, have been developed. This has led MFCCs to become ubiquitous in ASR. To cite but only two such techniques; the Vector Taylor Series approach of [7] and the Cepstral Mean Normalization technique of [8] compensate for time-domain additive and convolutional noise, respectively.

Despite their advantages, the difficulty of synthesizing speech from MFCCs has restricted their use to fields that do not require inverting MFCC vectors back into the original time-domain speech signals, e.g., ASR. This difficulty arises from the non-invertibility of several steps employed in MFCC generation; using the magnitude of the complex spectrum, mel-scale filterbank *binning* and higher-order cepstral coefficient truncation. In [4], however, we showed that high-quality highband speech reconstruction from MFCCs is feasible using a simple cepstral-domain interpolation scheme based on high-resolution inverse Discrete Cosine Transform (IDCT) [9]. Through this scheme, we were able to exploit the advantages of MFCCs to implement an MFCC-based BWE system that not only can perform better than conventional LP-based BWE in clean environments (due to MFCCs' superior correlation properties), but which can also perform more robustly in noisy environments (with MFCC denoising). Indeed, evaluating the performance of our static MFCC-based BWE scheme in [4] by Itakura-based measures has shown an improvement of up to 14.3%[3] compared to static LP-based BWE.

In the work presented here, we attempt to replicate the performance gains obtained by incorporating frontend-based memory into LP-based BWE, with that based on MFCCs. Results show that, indeed, memory inclusion improves MFCC-based BWE performance to the same extent it improves that using LSFs; i.e., dynamic MFCC-based BWE outperforms dynamic LP-based BWE by, more or less, the same degree static MFCC-based BWE outperforms that based on LSFs. Thus, by combining both frontend-based memory inclusion and MFCC features, we were able to exploit the superior correlation properties of each to reach an average cumulative improvement of $\approx 7.5\%$, as measured by LSD. Measured with the more subjectively-correlated COSH and symmetrized gain-optimized Itakura distortion measures [4], the improvement is $\approx 43.7\%$ and 53.9%, respectively.

2. MEMORY INCLUSION

2.1. Delta features

As described in [1], we include memory directly in the representation of spectral envelopes by means of delta coefficients, which are appended to (or replace part of) the MFCC/LSF static features[4]. Delta coefficients are obtained from static vectors by a first-order regression (time-derivative) implemented by calculating linearly weighted differences between neighbouring static vectors per

$$\delta_t = \frac{\sum_{l=1}^{L} l \cdot (c_{t+l} - c_{t-l})}{2\sum_{l=1}^{L} l^2},$$

where δ_t is a delta coefficient at frame t, $c_{t\pm l}$ is the corresponding static feature at frame $t \pm l$, and L specifies the number of neighbouring static frames (on each side of the tth frame) to consider.

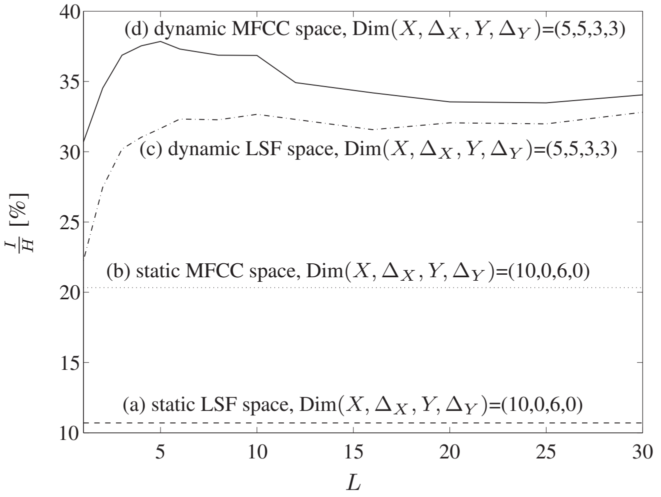

Fig. 1. Highband certainty, $\frac{I}{H}$, versus span of memory, L, for static and dynamic (static+delta) MFCC/LSF feature spaces with narrow and high band dimensionalities of 10 and 6, respectively.

2.2. Highband certainty

As stated above, highband certainty is defined as the ratio of mutual information, $I(\cdot\,;\cdot)$, to discrete highband entropy, $H(\cdot)$. Representing the static MFCC/LSF vectors of the narrow and high bands by X and Y, respectively, with Δ_X and Δ_Y representing the corresponding delta coefficient vectors, the highband certainty obtained with static and dynamic frontends can then be written as $\frac{I(X;Y)}{H(Y)}$ and $\frac{I(X,\Delta_X;Y,\Delta_Y)}{H(Y,\Delta_Y)}$, respectively. Using Gaussian mixture models and vector quantization of the highband feature vectors to estimate I and H, respectively[5], we obtain the highband certainty results illustrated in Fig. 1 for varying widths, L, of the time window used to calculate delta features[6]. Fig. 1 shows the superiority of MFCCs in retaining information content mutual to both bands (with both static and dynamic frontends). It also shows the considerable highband certainty gains achieved by memory inclusion (86% and 207% for MFCCs and LSFs, resp.), yet with no increase in frontend dimensionality. The gains peak for $3 \lesssim L \lesssim 13$ ($60 \lesssim t \lesssim 260$ ms), which includes the $200 \lesssim t \lesssim 250$ ms syllabic range. These results, thus, agree with the modulation spectra findings of [10] which show speech information content to be highest at the syllabic rate of 4–5 Hz.

In [2] and [4], we evaluated BWE performance using frontends (a), (b), and (c) in Fig. 1. Our results confirmed that, indeed, BWE performance follows the highband certainty results for these frontends. Thus, in the work presented here, we attempt to further improve BWE performance by making use of the superior correlation properties of frontend (d).

3. SPEECH RECONSTRUCTION FROM MFCC FEATURES

3.1. MFCC parametrization

Our MFCC parametrization for the narrowband (0–4 kHz) and highband (4–8 kHz) signals (obtained by filtering the wideband speech to be used in BWE GMM training), is detailed in [4]. Namely, the steps

[3] As described in Footnote 2, the results of our static MFCC-based BWE vary slightly from those reported in [4].

[4] LSF and MFCC parameterizations are detailed in [2] and [4], respectively.

[5] Refer to [2] for complete details on: (a) the estimation of I and H, and (b), the training and testing data sets used.

[6] The static LSF results of Fig. 1 differ slightly from those of Fig. 1 in [2] due to the changes in VQ implementation mentioned in Footnote 1.

involved are: (1) pre-emphasis, (2) windowing, (3) magnitude spectrum calculation, (4) mel-scale filterbank binning, (5) log operation, and (6) type-III DCT per

$$c_n = \sqrt{\frac{2}{N}} \sum_{k=0}^{N-1} (\log Y_k) \cos\left(\frac{(2k+1)n\pi}{2N}\right),$$

where c_n is the nth MFCC ($0 \leq n \leq N-1$), N is the number of mel-scale filters ($N = 6$), and Y_k (or X_k) is the kth highband (or narrowband) mel-scale filter energy. At $\text{Dim}(X,\Delta_X,Y,\Delta_Y) = (10,0,6,0)$, we showed in [4] that correlation between the two MFCC-parameterized frequency bands is almost twice that when using LSFs. Accordingly, we used these dimensionalities for our static MFCC-based BWE scheme to emphasize the performance improvement using MFCCs versus LSFs. To demonstrate the effect of including memory whilst preserving dimensionality, we replace the five higher-order narrowband MFCCs and the three higher-order highband MFCCs by the delta coefficients of the remaining five lower-order narrowband MFCCs and the three lower-order highband MFCCs, respectively; i.e., $\text{Dim}(X,\Delta_X,Y,\Delta_Y) = (5,5,3,3)$. This also allows us to compare BWE performance with the highband certainty results of Fig. 1.

3.2. Highband speech synthesis

Using the trained GMMs, the dynamic (static+delta) 6-dimensional highband MFCC vectors are estimated with maximum likelihood from the available dynamic 10-dimensional MFCC-parameterized narrow band. As indicated by the highband certainty results of Fig. 1, the GMMs trained on dynamic vectors will result in highband MFCC vectors with likelihoods higher than those resulting from the GMMs trained on static-only data. However, only the 3 static components of the highband vectors can now be used to reconstruct highband speech (compared to all 6 components in the static BWE case). Highband speech reconstruction from these 3 static components follows in a manner similar to that from the static-only 6-dimensinal MFCC vectors of [4]. Since simple IDCT would result in only 6 log-energies (with zero-padding the 3 static components to a total dimensionality of 6, in order to match the number of mel filters used in MFCC generation); insufficient to recreate the power spectrum, we use high-resolution IDCT instead, given by

$$\log \hat{Y}_{k'} = \sqrt{\frac{2}{N}} \sum_{n=0}^{N-1} c_n \cos\left(\frac{(2k'+1)n\pi}{2iN}\right),$$

where $0 \leq k' \leq iN-1$, $N = 6$, and i is an interpolation factor. This has the effect of performing interpolation between the mel filters' centre frequencies but in the cepstral domain, using the DCT basis cosine functions as the interpolating functions. The result is a total number of iN log-energies in the 4–8 kHz range. As shown in [4], a resolution of 1 mel in that range translates into $N = 6$ and $i = 100$, resulting in a fine 600-log-energy sample representation.

By exponentiation and mel-to-linear conversion, we obtain highband power spectra, which, using IFFT and the Levinson-Durbin recursion, can be further converted to LPCs. To generate the highband excitation signal, we use the narrowband signal equalized in the 3.4–4 kHz band. As shown in [5], Gaussian noise modulation by the 3–4 kHz signal envelope (containing strong pitch harmonics) results in a superior excitation signal, which, combined with the reconstructed highband LPCs, leads to excellent highband signal reconstruction through LP synthesis.

4. BWE RESULTS AND ANALYSIS

Through minor modifications, we incorporate frontend-based memory inclusion into our static dual-mode MFCC-based BWE system, detailed in [4]. For comparison, we also use the static and dynamic versions of our LSF-based BWE system of [2].

We evaluate BWE performance in the missing 4–7 kHz band by LSD (dB), given by

$$d^2_{\text{LSD}} = \frac{1}{\pi}\int_{\omega_l}^{\omega_h} \left(20\log_{10}\frac{g}{|Y(e^{j\omega})|} - 20\log_{10}\frac{\hat{g}}{|\hat{Y}(e^{j\omega})|}\right)^2 \mathrm{d}\omega,$$

where ω_l and ω_h are the cutoff frequencies of the missing high band, g and $Y(e^{j\omega})$ are the highband gain and frequency spectrum of the original wideband signal, respectively, while $\hat{g}$ and $\hat{Y}(e^{j\omega})$ are those of the GMM-estimated reconstructed signal.

In addition to LSD, which is widely used to evaluate BWE performance (and hence, allowing direct comparison of our results with those of previous works), we obtain more subjectively-correlated results by also using two Itakura-based distortion measures to evaluate our BWE performance. We argued in [4] that the the Itakura-Saito distortion [11] is more appropriate than LSD for evaluating the spectral reconstruction performance of BWE schemes in general; while LSD ignores the perceptual importance of some aspects of the LP speech spectrum representation (since it weights differences in formants and valleys equally), the Itakura-Saito distortion has more perceptual relevance in that it weights differences in LP spectra more heavily for peaks (which generally occur at formant locations) than for valleys. Indeed, it has been shown in [12] that the gain-optimized variant; the Itakura distortion [11], has a higher correlation of 0.73 with the subjective Diagnostic Acceptability Measure (versus 0.63 for LSD). Thus, we also evaluate BWE performance using: (a) the symmetrized Itakura-Saito distortion; the COSH measure, given by

$$d_{\text{COSH}} = \frac{1}{2}\left[d_{\text{IS}}\!\left(\tfrac{g^2}{|Y|^2}, \tfrac{\hat{g}^2}{|\hat{Y}|^2}\right) + d_{\text{IS}}\!\left(\tfrac{\hat{g}^2}{|\hat{Y}|^2}, \tfrac{g^2}{|Y|^2}\right)\right],$$

where

$$d_{\text{IS}}\!\left(\tfrac{g^2}{|Y|^2}, \tfrac{\hat{g}^2}{|\hat{Y}|^2}\right) = \frac{1}{2\pi}\int_{\omega_l}^{\omega_h}\left[\frac{g^2/|Y|^2}{\hat{g}^2/|\hat{Y}|^2} - \log\frac{g^2/|Y|^2}{\hat{g}^2/|\hat{Y}|^2} - 1\right]\mathrm{d}\omega,$$

and (b), the symmetrized gain-optimized Itakura distortion, given by

$$d_{\text{I}} = \frac{1}{2}\left[d_{\text{It}}\!\left(\tfrac{g^2}{|Y|^2}, \tfrac{\hat{g}^2}{|\hat{Y}|^2}\right) + d_{\text{It}}\!\left(\tfrac{\hat{g}^2}{|\hat{Y}|^2}, \tfrac{g^2}{|Y|^2}\right)\right],$$

where

$$d_{\text{It}}\!\left(\tfrac{g^2}{|Y|^2}, \tfrac{\hat{g}^2}{|\hat{Y}|^2}\right) \triangleq \min_{\hat{g}>0} d_{\text{IS}}\!\left(\tfrac{g^2}{|Y|^2}, \tfrac{\hat{g}^2}{|\hat{Y}|^2}\right) = \log\left(\frac{\hat{\boldsymbol{y}}^T \boldsymbol{R}_Y \hat{\boldsymbol{y}}}{g^2}\right),$$

using the reconstructed LPC vector, $\hat{\boldsymbol{y}}^T$, and the Toeplitz autocorrelation matrix, $\boldsymbol{R}_Y$, of the original signal LP model. Since d_{I} is gain-independent while d_{COSH} is not, not only do we obtain more subjectively correlated results by employing both distortion measures, but we also gain a means by which to evaluate performance of highband gain-related and spectral shape-related reconstruction, separately.

Fig. 2 shows BWE performance evaluated using the three distortion measures. While illustrating the gains obtained by incorporating memory into our MFCC-based BWE system, Fig. 2, in effect, also summarizes our previous works in [2] and [4]. Comparing the bottom subplots with the top ones (i.e., dynamic frontends versus static ones) shows the considerable gains in BWE performance obtained by memory inclusion in general. These gains are tabulated in Table 1. The benefits of exploiting MFCCs' superior correlation

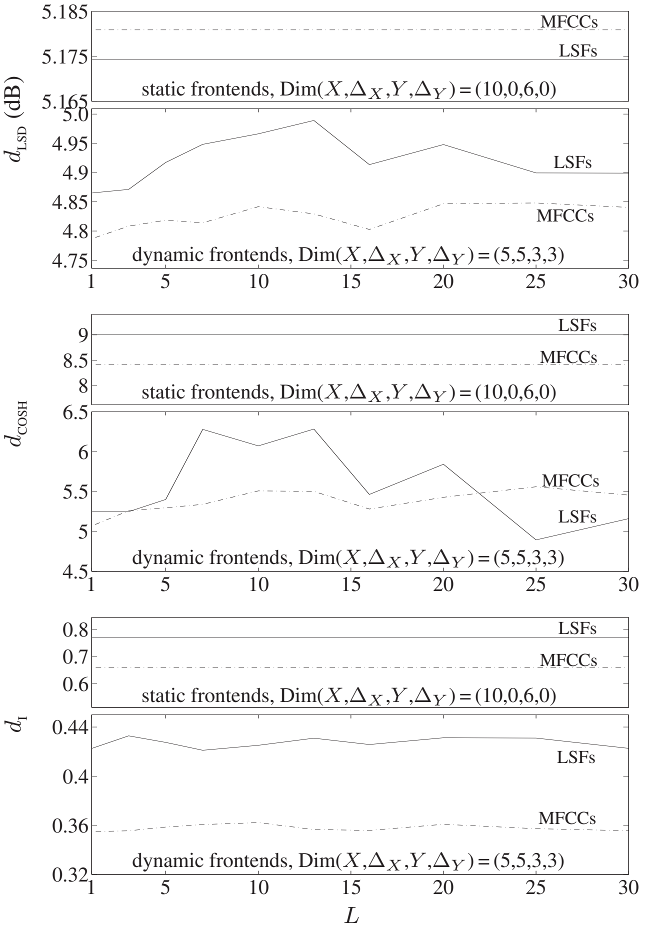

Fig. 2. BWE performance with memory inclusion (bottom subplots) versus performance with no memory inclusion (top subplots), for frontends with MFCCs (dotted lines) and LSFs (solid lines).

properties are also shown by comparing the dotted lines in Fig. 2 to the corresponding solid ones, with the gains tabulated in Table 2 as well. Tabulated gains in the dynamic cases are estimated over L.

Thus, by combining both frontend memory and MFCCs features in a conventional LSF-based BWE system, we achieve an average cumulative improvement of 7.5% in terms of d_{LSD}. The more subjectively-correlated d_{COSH} indicates that the improvement is, in fact, quite higher; 43.7%. Considering spectral-shape reconstruction alone (by eliminating the effect of gain-related differences through d_{I}) shows an even higher improvement of 53.9%, indicating that the trained GMMs were more successful in reconstructing the shapes of highband spectra than in reconstructing highband gains. This latter finding is true for most d_{I} results in Tables 1 and 2 above.

Finally, we note that these BWE performance results generally follow the trends of highband certainty in Fig. 1; e.g., adding frontend memory to static LSF-based BWE leads to a performance improvement higher than that obtained by using static MFCCs rather than LSFs, conforming with the higher highband certainty of frontend (c) in Fig. 1 compared to frontend (b). While the actual BWE performance gains of Tables 1 and 2 are minor relative to the considerable highband certainty gains of Fig. 1, we should also note that the latter are, in fact, upper bounds on achievable performance improvements; highband certainty only matches GMM performance, and hence, does not take into account all other components in a real BWE system (e.g., errors in generating the excitation signal, errors in reconstructing speech from MFCCs, etc.). Thus, the ideal BWE system is that which can translate the full potential of temporal memory and parametrization into matching performance gains.

Table 1. Average (maximum) BWE performance improvement obtained by frontend memory inclusion.

	d_{LSD} (dB)	d_{COSH}	d_{I}
LSFs	4.9% (6.0%)	38.0% (45.7%)	44.6% (45.3%)
MFCCs	6.9% (7.6%)	36.2% (39.7%)	45.8% (46.2%)

Table 2. Average (maximum) BWE performance improvement obtained by employing MFCCs rather than LSFs.

	d_{LSD} (dB)	d_{COSH}	d_{I}
static	-0.1%	6.6%	14.3%
dynamic	2.0% (3.2%)	3.3% (15.0%)	16.2% (17.9%)

5. REFERENCES

[1] A. H. Nour-Eldin, T. Z. Shabestary and P. Kabal, "The effect of memory inclusion on mutual information between speech frequency bands", in *Proc. ICASSP*, pp. III-53–56, 2006.

[2] A. H. Nour-Eldin and P. Kabal, "Objective analysis of the effect of memory inclusion on bandwidth extension of narrowband speech", in *Proc. InterSpeech*, pp. 2489–2492, 2007.

[3] M. Nilsson, H. Gustafsson, S. V. Andersen and W. B. Kleijn, "Gaussian mixture model based mutual information estimation between frequency bands in speech", in *Proc. ICASSP*, pp. I-525–528, 2002.

[4] A. H. Nour-Eldin and P. Kabal, "Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech", in *Proc. InterSpeech*, pp. 53–56, 2008.

[5] Y. Qian and P. Kabal, "Combining equalization and estimation for bandwidth extension of narrowband speech", in *Proc. ICASSP*, pp. I-713–716, 2004.

[6] P. Jax and P. Vary, "Feature selection for improved bandwidth extension of speech signals", in *Proc. ICASSP*, pp. I-697–700, 2004.

[7] P. J. Moreno, B. Raj, and R. M. Stern, "A Vector Taylor Series approach to environment-independent speech recognition", in *Proc. ICASSP*, pp. II-733–736, 1996.

[8] B. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *J. Acoust. Soc. Am.*, vol. 55, no. 6, pp. 1304–1312, 1974.

[9] T. Ramabadran, J. Meunier, M. Jasiuk and B. Kushner, "Enhancing distributed speech recognition with back-end speech reconstruction", in *Proc. EuroSpeech*, pp. 1859-1862, 2001.

[10] S. Greenberg and B. E. D. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech", in *Proc. ICASSP*, pp. III-1647–1650, 1997.

[11] R. M. Gray, A. Buzo, A. H. Gray, Jr. and Y. Matsuyama, "Distortion measures for speech processing", *IEEE Trans. ASSP*, vol. 28, no. 4, pp. 367–376, 1980.

[12] S. R. Quackenbush, T. P. Barnwell III and M. A. Clements, *Objective Measures of Speech Quality*, Englewood Cliffs, NJ: Prentice-Hall, 1988.