Feature Compensation Techniques for ASR on Band-Limited Speech

Nicolás Morales, *Member, IEEE*, Doroteo Torre Toledano, *Member, IEEE*, John H. L. Hansen, *Fellow, IEEE*, and Javier Garrido, *Member, IEEE*

***Abstract*—Band-limited speech (speech for which parts of the spectrum are completely lost) is a major cause for accuracy degradation of automatic speech recognition (ASR) systems particularly when acoustic models have been trained with data with a different spectral range. In this paper, we present an extensive study of the problem of ASR of band-limited speech with full-bandwidth acoustic models. Our focus is mainly on band-limited feature compensation, covering even the case of time-varying band-limiting distortions, but we also compare this approach to more common model-side techniques (adaptation and retraining) and explore the combination of feature-based and model-side approaches. The feature compensation algorithms proposed are organized in a unified framework supported by a novel mathematical model of the impact of such distortions on Mel-frequency cepstral coefficient (MFCC) features. A crucial and novel contribution is the analysis made of the relative correlation of different elements in the MFCC feature vector for the cases of full-bandwidth and limited-bandwidth speech, which justifies an important modification in the feature compensation scheme. Furthermore, an intensive experimental analysis is provided. Experiments are conducted on real telephone channels, as well as artificial low-pass and band-pass filters applied over TIMIT data, and results are given for different experimental constraints and variations of the feature compensation method. Results for other well-known robustness approaches, such as cepstral mean normalization (CMN), model retraining, and model adaptation are also given for comparison. ASR performance with our approach is similar or even better than model adaptation, and we argue that in particular cases such as rapidly varying distortions, or limited computational or memory resources, feature compensation is more convenient. Furthermore, we show that feature-side and model-side approaches may be combined, outperforming any of those approaches alone.**

***Index Terms*—Automatic speech recognition (ASR), feature compensation, restricted communications bandwidth channels, robustness.**

Manuscript received January 19, 2008; revised November 17, 2008. First published April 15, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Simon King.

N. Morales was with the HCTLab-Universidad Autónoma de Madrid, Escuela Politécnica Superior, 28049 Madrid, Spain. He is now with Nuance Communications GmbH, 52052 Aachen, Germany (e-mail: nicolas.morales@nuance.com).

D. T. Toledano is with the ATVS-Universidad Autónoma de Madrid, Escuela Politécnica Superior, 28049 Madrid, Spain (e-mail: doroteo.torre@uam.es).

J. H. L. Hansen is with the Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, Richardson, TX 75083 USA (e-mail: john.hansen@utdallas.edu).

J. Garrido is with the HCTLab-Universidad Autónoma de Madrid, Escuela Politécnica Superior, 28049 Madrid, Spain (e-mail: javier.garrido@uam.es).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2008.2012321

I. Introduction

MISMATCH between training and test conditions is a serious cause for accuracy loss in automatic speech recognition (ASR) systems [37]. Among the variety of possible causes for mismatch in real systems, we study in this work those introduced by the channel, and more specifically band-limiting distortions that completely remove parts of the spectrum.

Band-limiting distortions exist in historical recordings due to the limited capabilities of recording devices and storage units [16], [17], [34]. Also, telephone-transmitted signals are usually bandpass filtered with cutoff frequencies 300 Hz and 3.4 kHz [43], and similarly, signals transmitted from on-board systems (like cars or aeroplanes) may present band-limitations [1], [9]. Low-sampling rates also impose an upper limit to the available spectrum [35], as can be seen in recordings made with personal portable devices. Oftentimes, band-limited recordings of different qualities may be interspersed, for example in documentaries, broadcast news, etc; a situation that further complicates ASR (Fig. 1).

Although speech understandability is assumed to be maximal for signals that expand over the whole spectrum (both for humans [36] and ASR systems [42]), several studies have shown that given the redundancy of speech signals humans achieve high accuracy rates using only a fraction of the complete spectrum (low-pass or high-pass filters [3]; spectral slits of 1/3 octave width in the region 1100–2100 Hz [45]). ASR systems trained specifically for a particular bandwidth are only slightly outperformed by full-bandwidth systems provided that a reasonable part of the spectrum remains available. Therefore, the major cause of degradation in ASR of band-limited data and full-bandwidth acoustic models seems to be the mismatch and not the loss of information in the signal.

Typical solutions to the problem of mismatch are training specific models for the new condition, or adapting previously existing models. Both approaches are reliable, but nevertheless, under particular circumstances feature compensation may constitute a better solution. For example when different distortions affect different segments of test data in an unknown and unpredicted manner, a model-side solution would require important modifications in the search process (analysis of available bandwidth and/or combination of outputs from multiple recognizers, as in ROVER [14]), or sophisticated acoustic models valid for the variety of conditions (e.g., multistyle-trained models or 3-D Viterbi decoding [44], [46]), with a significant increase in computation time. On the contrary, feature compensation performed

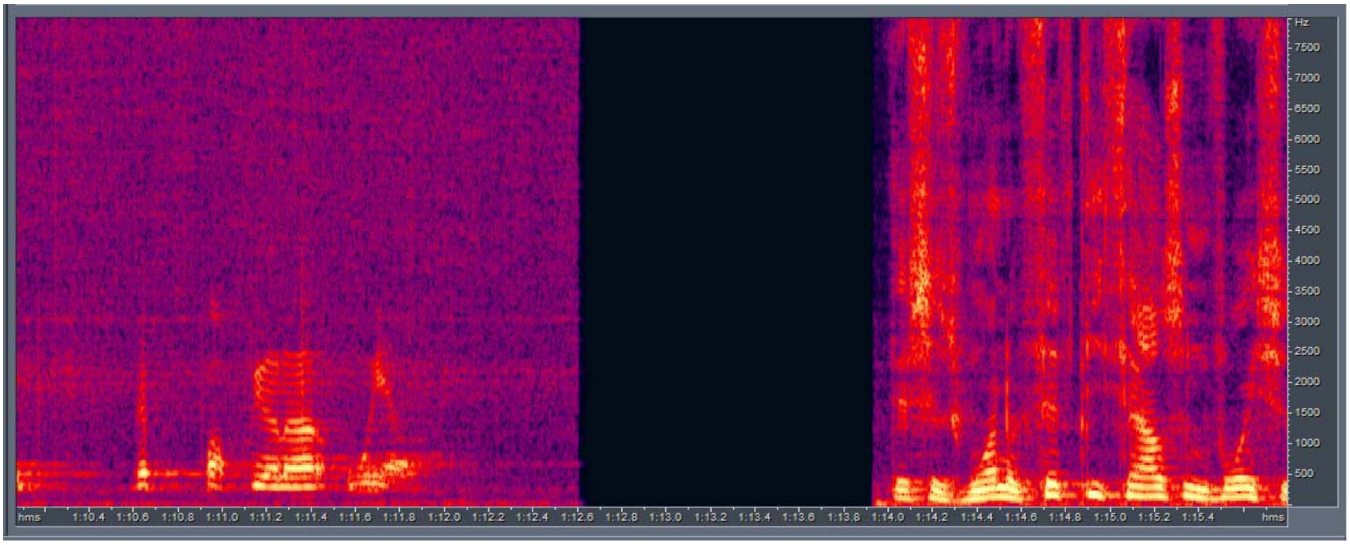


Fig. 1. This spectrogram shows an example of a file from the NGSW collection [34] where the available bandwidth is below 2500 Hz: a speech made on September 22nd, 1912 by US presidential candidate, Theodore Roosevelt (left part of the spectrogram). This is followed by a zero-content portion (represented as black in the spectrogram) and a third part that spans the entire spectrum up to 8 kHz, where the anchor discusses the historical recording.

in an independent module is able to automatically detect and compensate multiple types of distortion in a process that remains transparent to the decoder module. Feature compensation is also convenient in systems where memory and computational costs are an important limitation as it offers a light and reliable solution, allowing to store a single set of acoustic models and performing ASR with only one recognition engine (the memory space required to store corrector functions is between one and two orders of magnitude below that of full-bandwidth acoustic models [31]). Finally, if no time-varying band limitations are expected and computational resources are not an issue, feature compensation can be combined with model-side techniques for improved accuracy.

Bandwidth expansion has been previously used in the time domain on telephone-transmitted signals in order to obtain more natural sounds [4], [6], [22], [24], [47]. Although the same approach may be used when the goal is ASR, the effort of reconstructing the whole time-domain signal seems unnecessary (e.g., phase information is discarded in parameterizations based on the signal's power spectrum). Feature compensation has been previously applied for compensation of noise distortions [10], [12], [33], [50], [51], but only recently for the problem of band-limiting distortions—to our knowledge the earliest work in this direction is [25]. Both, noise and bandwidth limitation introduce uncertainty in the spectrum, as proposed in [50], and in both cases this uncertainty can be time-varying [51]. However, we show in this study that bandwidth limitations present some peculiarities that require a different treatment than noise. In [27] and [28], we proposed linear compensations of the Mel frequency cepstrum coefficients (MFCCs) [7] based on knowledge-based or data-driven criteria, respectively. Similarly, in [41], an extension of the SPLICE framework for noise compensation has been proposed for restoration of band-limited signals. In [23], the masking approach originally designed for compensation of additive noise [40] is modified to fit band-limiting distortions.

In the present work, we propose a number of algorithms for feature compensation whose common ground is the learning of a transformation between the distorted (band-limited spectrum) and undistorted (full-bandwidth) feature spaces. Different constraints studied are availability or not of stereo-data for training (speech samples recorded simultaneously in full and band-limited environments), training data scarcity and blind classification and compensation of multiple distortions. Evaluation is performed on distorted TIMIT data, including real telephone channel distortions (NTIMIT [21] and our own TIMIT-derived corpus, STC-TIMIT [32]), as well as artificial low-pass and bandpass filters. Results are always compared to other robustness methods: Cepstral mean normalization (CMN), model adaptation, and model retraining.

The rest of this paper is organized as follows. In Section II, we present a novel mathematical model of the effect of band-limiting distortions on MFCC features. This gives rise to a unified framework for feature compensation explained in Section III. In Section IV, we make theoretical and empirical observations on the topic of correlation of MFCC vector elements for full-bandwidth and band-limited speech, and show that in the later case MFCCs are significantly more correlated. This justifies an important modification in the compensation schemes proposed. In Section V, we describe the experimental framework, and Section VI shows an extensive collection of results under different settings and problem constraints. In Section VII, we summarize our results and extract conclusions.

II. Mathematical Model of the Effects of Band-Limiting Distortions on MFCCs

The effect of a purely convolutional distortion may be expressed for the power spectrum as

$$|Y_t(f)|^2 = |H_t(f)|^2 \cdot |X_t(f)|^2 \tag{1}$$

where $Y_t(f)$ and $X_t(f)$ represent the spectra for the distorted and original signals, respectively, for a time frame t, and $H_t(f)$ is the frequency response of the distortion. When the front-end employed is derived from a bank of filters, the following approximation is typically assumed for each filter (j is the order of the filter, and f_j the center frequency of the filter; the second part is just a lighter notation)

$$|Y_t(f_j)|^2 \approx |H_t(f_j)|^2 \cdot |X_t(f_j)|^2 \Rightarrow |Y_{j,t}|^2 \approx h_{j,t} \cdot |X_{j,t}|^2. \tag{2}$$

The singularity of band-limiting distortions is that parts of the spectrum are completely removed, whereas in general convolutional distortions different spectral bands are normally multiplied by nonzero values. For complete removal of bands, the approximation in (2) remains valid, but $h_{j,t}$ follows a particular form

$$|Y_{j,t}|^2 = h_{j,t} \cdot |X_{j,t}|^2 + e_{j,t}, \quad \text{where} \begin{cases} h_{j,t} = 0, & \text{if } j \in F \\ h_{j,t} = 1, & \text{if } j \notin F \end{cases} \tag{3}$$

and F represents the channels in the filterbank removed by the bandwidth-limitation (a further approximation since we assume constant and binary values for the band-limitation in the outputs of the bank of filters). Term $e_{j,t}$ is introduced for taking into account the approximation made in (2). This is particularly important when $h_{j,t} = 0$, so as to reflect the fact that $|Y_{j,t}|$ is nonzero. Complete removal of bands implies that compensation techniques based on signal restoration using the channel's inverse transfer function (equalization techniques, CMN, etc.) are not useful anymore, simply because the inverse of the transfer function $h_{j,t} = 0$ does not exist (information from removed channels is completely lost). Therefore, reconstruction of removed parts of the spectrum should be attempted in a different fashion, for example with information from the available parts (assuming important correlation between different spectral regions).

The general definition of MFCCs is [52]

$$x_{i,t} = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} \log\left(|X_{j,t}|^2\right) \cos\left(\frac{\pi i}{N}(j - 0.5)\right) \tag{4}$$

where subindex i is the order of the MFCC coefficient,[1] t represents a time frame, and N is the number of channels in the filterbank. In subsequent equations subindex t is dropped for simplicity of notation. We may rewrite (4) as

$$x_i = \sum_{j=1}^{N} C_{ij} \cdot \log\left(|X_j|^2\right)$$
$$C_{ij} = \sqrt{\frac{2}{N}} \cdot \cos\left(\frac{\pi i}{N}(j - 0.5)\right) \tag{5}$$

[1]MFCC stands for Mel-frequency cepstrum coefficient. Thus, instead of MFCC coefficients, it would be more appropriate to say MFC coefficients. However, for clarity MFCC will be maintained as a full unit of meaning.

where C_{ij} are individual elements of the discrete cosine transformation matrix. Similarly, MFCC features of band-limited speech are

$$y_i = \sum_{j=1}^{N} C_{ij} \cdot \left(\log\left(h_j \cdot |X_j|^2 + e_j\right)\right). \tag{6}$$

The difference between full-bandwidth and band-limited MFCC vectors for a particular frame is obtained from (5) and (6)

$$x_i - y_i = \sum_{j=1}^{N} C_{ij} \cdot \left[\log\left(|X_j|^2\right) - \log\left(h_j \cdot |X_j|^2 + e_j\right)\right]. \tag{7}$$

Now we decompose the sum over all filters in the filterbank into 2 terms corresponding to channels affected by the bandwidth restriction and intact channels, respectively

$$x_i - y_i = \sum_{\substack{j=1 \\ j \notin F}}^{N} C_{ij} \cdot \left[\log\left(|X_j|^2\right) - \log\left(h_j \cdot |X_j|^2 + e_j\right)\right] + \sum_{\substack{j=1 \\ j \in F}}^{N} C_{ij} \cdot \left[\log\left(|X_j|^2\right) - \log\left(h_j \cdot |X_j|^2 + e_j\right)\right]. \tag{8}$$

For band-limited speech, $h_j \to 1$ for unmodified parts of the spectrum and $h_j \to 0$ for removed channels (or for expanded regions in upsampled data). For channels unaffected by the distortion $h_j \cdot |X_j|^2 \to |X_j|^2$, and $e_j \ll |X_j|^2$, so the sum over the unaffected channels in (8) can be discarded. However, for removed bands $h_j \cdot |X_j|^2 \ll e_j$. We may then approximate full-bandwidth MFCCs as

$$x_i \approx y_i + \sum_{\substack{j=1 \\ j \in F}}^{N} C_{ij} \cdot \left[\log\left(|X_j|^2\right) - \log(e_j)\right]. \tag{9}$$

In practice, values of e_j are random and significantly smaller than the values of the original signal in (9). Therefore, we can ignore this term, as in (10), for a better understanding of the physical meaning of the equation

$$x_i \approx y_i + \sum_{\substack{j=1 \\ j \in F}}^{N} C_{ij} \cdot \left[\log\left(|X_j|^2\right)\right]. \tag{10}$$

This equation shows full-bandwidth features as being a combination of limited-bandwidth features, plus the would-be contribution to MFCC vectors of missing parts of the spectrum (something that is also intuitive). However, it does not seem to help in estimating full-bandwidth features because the outputs of the filters in the missing parts of the spectrum are unavailable for reconstruction. Nevertheless, we can still make use of it thanks to a central idea of bandwidth extension: different parts of the spectrum are highly correlated because the configuration adopted by the vocal tract determines the whole shape of the spectral envelope in a particular instant. Under this assumption and for a cluster of observations k that share similar acoustic

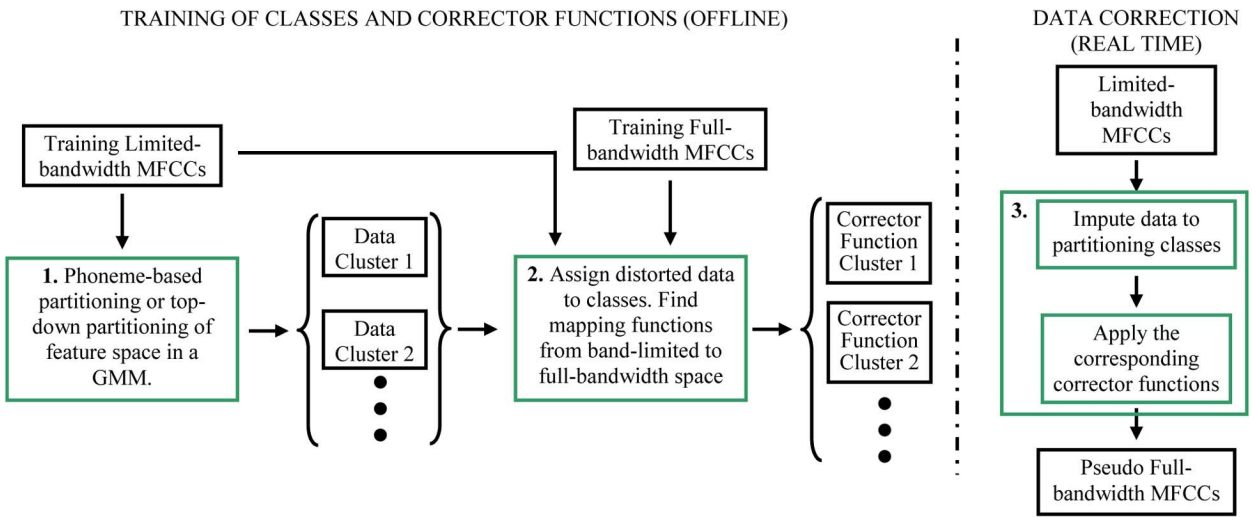


Fig. 2. Schematic representation of the proposed architecture for training of classes and corrector functions and for compensation of band-limited feature vectors to generate pseudo full-bandwidth feature vectors.

characteristics (similar configuration of the vocal tract), the filterbank outputs in the removed regions of the spectrum when input is full-bandwidth speech are related to those in the available parts (a_m) through some unknown functions E_j^k

$$|X_j|^2 \approx E_j^k\left(|X_{a_1}|^2,\ldots,|X_{a_M}|^2\right);\quad j\in F \text{ and } a_m\notin F. \tag{11}$$

Our goal being direct reconstruction using MFCCs, we may substitute in (11) the outputs of individual filters in the filterbank $\left(|X_{a_1}|^2,\ldots,|X_{a_M}|^2\right)$, by the MFCC vectors of restricted bandwidth speech, which relate to the former via the Moore–Penrose inverse transformation [53]. Thus

$$|X_j|^2 \approx G_j^k(\mathbf{y});\quad j\in F \tag{12}$$

where G_j^k is a set of mapping functions between band-limited MFCC vectors and filters' outputs from the original signal. Finally, inserting this into (10) and combining G_j^k with the logarithm and the C_{ij} terms into the new functions H_j^k we obtain a new set of equations

$$\begin{aligned} x_i &\approx y_i + \sum_{\substack{j=1\\ j\in F}}^{N} C_{ij}\cdot\left[\log\left(G_j^k(\mathbf{y})\right)\right] \\ &= y_i + \sum_{\substack{j=1\\ j\in F}}^{N} H_j^k(\mathbf{y}) = J_j^k(\mathbf{y}). \end{aligned} \tag{13}$$

Equation (13) establishes that full-bandwidth features belonging to a particular class of sounds k may be approximately reconstructed using unknown functions of the band-limited features $J_j^k(\mathbf{y})$. Section III deals with methods to partition speech into classes of sounds k and different methods for estimation of the $J_j^k(\mathbf{y})$ functions.

III. Feature Compensation Framework

In this section, we propose a unified feature compensation framework. In Section III-A, we present the foundations for two proposed approaches for feature compensation (a data-driven approach, *Gaussian class-based* and a knowledge-driven approach, *Phoneme-based*). As shown in Fig. 2, both approaches involve three steps that we describe in Sections III-B–D. First, training data is divided into clusters (Section III-B, box 1 in Fig. 2); second, for each cluster a set of corrector functions is trained (Section III-C, box 2 in Fig. 2); and third, full-bandwidth features are estimated from band-limited data (Section III-D, box 3 in Fig. 2). We note that the first two steps are conducted offline, whereas the third one is performed at recognition time.

A. Foundations for Full-Bandwidth Estimation

1) Data-Driven Approach (Gaussian Class-Based): The formulation presented here is in many aspects similar to that used for feature compensation of noisy speech [20]. We will highlight later the main differences we find when compensating band-limited speech instead of noisy speech.

We assume that both the full-bandwidth and limited-bandwidth spaces may be modeled using Gaussian mixture models (GMMs). Thus, for the limited-bandwidth space the probability of observing a feature vector is

$$p(\mathbf{y}) = \sum_{k=1}^{K} p(\mathbf{y}|k)\cdot P(k) = \sum_{k=1}^{K} N(\mathbf{y};\boldsymbol{\mu}_k,\Sigma_k)\cdot P(k) \tag{14}$$

where $P(k)$ is the *a priori* probability of mixture k in the GMM, $\boldsymbol{\mu}_k$ the mean vector, and $\boldsymbol{\Sigma}_k$ the covariance matrix. Assuming $\mathbf{x}$ and $\mathbf{y}$ jointly Gaussian within a cluster of data pairs k the conditional expectation of clean feature vectors given the distorted vectors and the cluster is

$$\begin{aligned} E\{\mathbf{x}|\mathbf{y},k\} &= \boldsymbol{\mu}_{X,k} + \Sigma_{XY,k}\left(\Sigma_{Y,k}\right)^{-1}\left(\mathbf{y}-\boldsymbol{\mu}_{Y,k}\right) \\ &= \mathbf{B}_k\mathbf{y} + \mathbf{b}_k \end{aligned} \tag{15}$$

where $\boldsymbol{\mu}_{X,k}$ and $\boldsymbol{\mu}_{Y,k}$ are the vectors of means for mixture k for full-bandwidth and limited-bandwidth speech, respectively, $\boldsymbol{\Sigma}_{Y,k}$ is the covariance matrix of band-limited data, and $\boldsymbol{\Sigma}_{XY,k}$

is the joint covariance matrix. We call $\mathbf{B}_k$ and $\mathbf{b}_k$ the compensation matrix and offset vector, respectively. The joint pdf is

$$\begin{aligned} p(\mathbf{x},\mathbf{y}) &= \sum_{k=1}^{K} p(\mathbf{x},\mathbf{y}\,|k)\cdot P(k) \\ &= \sum_{k=1}^{K} p(\mathbf{x}\,|\mathbf{y},k)\,p(\mathbf{y}\,|k)\cdot P(k) \\ &= \sum_{k=1}^{K} N(\mathbf{x};\mathbf{B}_k\mathbf{y}+\mathbf{b}_k,\boldsymbol{\Gamma}_k)\cdot N(\mathbf{y};\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\cdot P(k) \end{aligned} \tag{16}$$

where $\boldsymbol{\Gamma}_k$ is the covariance matrix of the conditional probability of $\mathbf{x}$ and $\mathbf{y}$ for class k.

Estimation of undistorted features from distorted features is possible maximizing the joint probability, using for example the minimum mean squared error (MMSE) criterion. In [19], it is shown that the solution is given by mapping functions taking the form of the conditional expectation, and so, for a particular observation $\mathbf{y}$

$$\begin{aligned} \mathbf{x}^{\mathrm{MMSE}} &= E\{\mathbf{x}\,|\mathbf{y}\} = \sum_{k=1}^{K} P(k\,|\mathbf{y})\cdot E\{\mathbf{x}\,|\mathbf{y},k\} \\ &= \sum_{k=1}^{K} P(k\,|\mathbf{y})\cdot(\mathbf{B}_k\mathbf{y}+\mathbf{b}_k). \end{aligned} \tag{17}$$

In the general form, $\mathbf{B}_k$ is a complete matrix and compensation of each individual element in the feature vector is made as a linear compensation of all the other elements in the vector (we call this *multivariate feature compensation*, in which, for instance, cepstrum coefficient C2 can be compensated using C2 and any other components of the feature vector). However, for MFCCs the assumed uncorrelation of different elements seems to indicate that similar compensation ability may be obtained simply using a diagonal matrix (in this case we talk about *univariate linear compensation*, in which, for instance, cepstrum coefficient C2 can be compensated using only C2). In later sections, we show that, while this simplification seems to work well for compensating noise, it is suboptimal for the case of band-limited speech.

Feature compensation methods with GMMs previously used for noisy speech differ in the methodology used for estimating GMMs, and the transformation matrix $\mathbf{B}_k$ and offset vector $\mathbf{b}_k$ for each Gaussian mixture (normally assuming diagonal transformation matrixes, and sometimes even further simplifications, such as assuming $\mathbf{B}_k = \mathbf{I}$). In RATZ, GMMs are the emitting states in the set of acoustic models of an ASR engine [33]. In SPLICE, specific GMMs defined in the noisy feature space are used, and the number of such partitioning classes may be tuned according to the amount of adaptation data and the complexity of the distortion [12]. In [5], it is proposed to model the feature space as GMMs for both, the clean and distorted spaces, and associating Gaussians from one and the other space prior to compensation. Also, in [2], clean and distorted features are concatenated prior to feature space partitioning in order to obtain groups of data clustered both in the original and distorted spaces. A similar approach to SPLICE is followed in [40], but

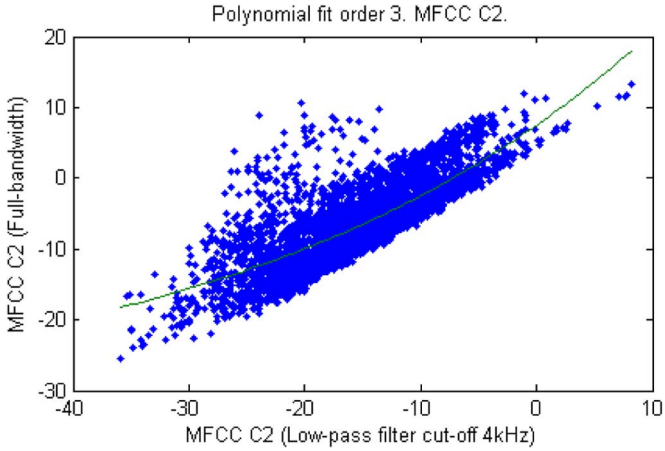


Fig. 3. Mapping of LP4 kHz data (x-coordinate) to full-bandwidth data (y-coordinate) for MFCC C2 in a particular Gaussian class. The plot also shows a third-order polynomial fit.

compensation is made on the outputs of the filterbank, thus removing the extra complication of the cosine transformation and allowing for straightforward exploitation of the redundancy of different spectral regions. In our work, the mathematical model of band-limiting distortions made in Section II motivates the use of such Gaussian class-based techniques [2], [5], [12], [33] for the problem of bandwidth limitation.

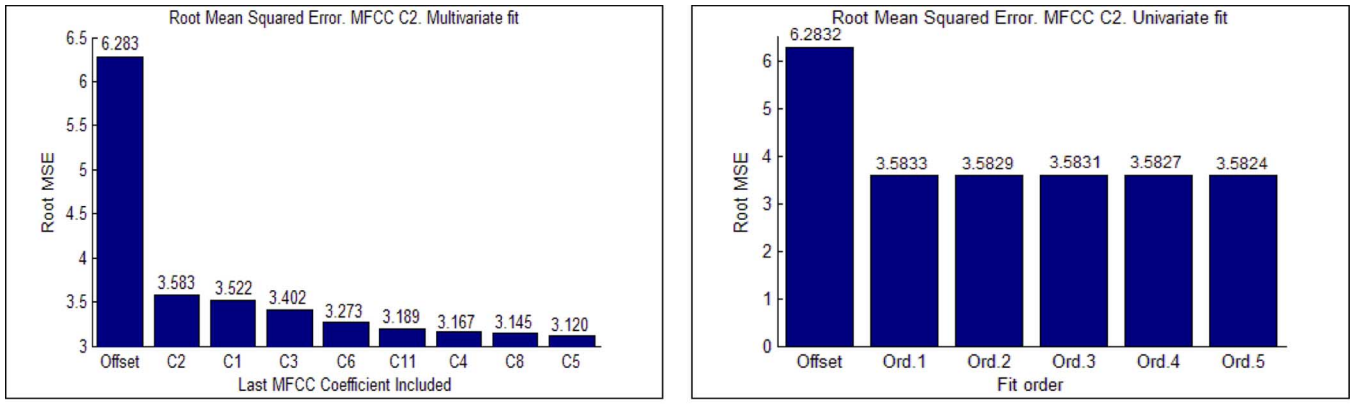
2) Knowledge-Based Approach (Phoneme-Based): In the previous section, we proposed data clustering in the band-limited feature space according to a maximum-likelihood criterion. Alternatively, we may do this partitioning based on the phonetic content of frames (this requires phonetic labeling of training data). Training of corrector functions is independent of the method used for data clustering and so it can be done exactly in the same way as in the data-driven method. However, feature compensation would require phonetic labeling of test data, something that is never available (otherwise, ASR would not be needed). Different solutions, as well as an upper performance bound that we call *oracle*, are presented in Section III-D1.

B. Partitioning the Acoustic Space

In the training stage, the feature space is first partitioned into clusters (partitioning classes). It is assumed that data within a cluster in the distorted feature space suffered a similar transformation as a result of the distortion and thus, to a certain extent, full-bandwidth data will show a similar clustering structure. Therefore, a set of corrector functions is trained for each cluster, mapping observations from the distorted space to the full-bandwidth space.

1) Phoneme-Based Partitioning: In phoneme-based partitioning, it is assumed that realizations of any given phoneme present a more or less stable energy distribution and therefore may be compensated with the same mapping functions. The feature space is divided straightforwardly using phonetically aligned labels of training data.

2) Gaussian Class-Based Partitioning: In our work, the clusters of data k shown in (15) are found using an iterative method. Partitioning is initialized with a single cluster defined as a Gaussian distribution with mean $\boldsymbol{\mu}_{0,0}$ and covariance $\boldsymbol{\Sigma}_{0,0}$,



a) b)

Fig. 4. Evolution of RMSE for stepwise (a) multivariate and (b) univariate estimation of full-bandwidth MFCC C2 from limited-bandwidth MFCCs (TIMIT LP4 kHz). Significant improvements in signal restoration are possible with the inclusion of multiple elements in the multivariate fit, whereas univariate fits of order 2 and larger provide no improvements compared to a linear fit.

computed for all training data. This initial cluster is divided into two by perturbing the mean vector by $\pm\eta$ times the vector of standard deviations, where η is a perturbation factor (0.2 in all our experiments). Training data are reassigned to either cluster, and means and covariances are recalculated. Reassignment is repeated a number of times (3 in our experiments) before increasing again the number of Gaussians, by splitting a new Gaussian mixture according to a partitioning criterion. The process continues until the desired number of Gaussians is achieved.

C. Training of Corrector Functions

Two methodologies are proposed depending on whether stereo data are available for training, or not.

1) Training of Corrector Functions With Stereo Data: When stereo data are available, it is possible to learn a mapping from limited-bandwidth to full-bandwidth data using linear least squares curve fitting techniques. For example, for univariate polynomial compensation, the coefficients for each corrector class were obtained independently for each element in the feature vector by fitting pairs of observations (full-bandwidth and limited-bandwidth) to a polynomial curve [39] (an example for coefficient MFCC C2 is shown in Fig. 3; comparable plots exist for other MFCCs).

In multivariate compensation, we used a step-wise strategy that successively introduces new elements in a multivariate fit using analysis of variance. The number of final coefficients is set dynamically so as to reach a point close to the minimum of the root mean squared error (RMSE) for each coefficient to be compensated. Fig. 4(a) shows an example where the RMSE of univariate linear compensation, 3.583 may be improved to down to 3.189 by the inclusion of four additional elements using multivariate compensation. However, inclusion of additional coefficients offers smaller improvement, suggesting that in this case, the multivariate fit may be truncated after the first six terms. Moreover, extra coefficients might cause over-fitting and have a negative impact in reconstruction of test data unseen during training. In contrast with the important reduction of RMSE using multivariate feature compensation, in Fig. 4(b) we show that for univariate polynomial fits (another possible extension to the basic compensation approach) no advantage is obtained by using large orders of polynomial fits (the same conclusion was obtained directly in ASR accuracy in previous work [28]).

2) Training of Corrector Functions With Nonstereo Data: In situations where collecting stereo data is complicated, it is still possible to train corrector functions. As it was previously shown, the partitioning techniques used do not require stereo data. The following shows how to estimate the feature compensation coefficients.

In the most general case, the vector of means $\boldsymbol{\mu}_{X,k}$ and matrix of covariance $\boldsymbol{\Sigma}_{X,k}$ of a cluster k in the full-bandwidth feature space are related to their limited-bandwidth counterparts as

$$\boldsymbol{\mu}_{X,k} = \boldsymbol{\mu}_{Y,k} + \mathbf{r}_k \tag{18}$$

$$\boldsymbol{\Sigma}_{X,k} = \boldsymbol{\Sigma}_{Y,k} + \mathbf{R}_k \tag{19}$$

where for the moment no constraints are imposed on the correction factors $\mathbf{r}_k$ and $\mathbf{R}_k$. The probability of observation of a feature vector in the full-bandwidth space is

$$p(\mathbf{x}) = \sum_{k=1}^{K} N\left(\mathbf{x}; \boldsymbol{\mu}_{X,k}, \boldsymbol{\Sigma}_{X,k}\right) \cdot P(k). \tag{20}$$

In Appendix A, we show how this probability may be maximized using an expectation-maximization (EM) strategy [8]. For corrections in the form of simple offsets, the update equations for the vector of means take the form

$$\bar{\mathbf{r}}_k = \frac{\sum_{t=1}^{T} p(k|\mathbf{x}_t, \phi) \cdot \mathbf{x}_t}{\sum_{t=1}^{T} p(k|\mathbf{x}_t, \phi)} - \boldsymbol{\mu}_{Y,k} \tag{21}$$

where ϕ is the set of parameters for Gaussian mixtures in the previous iteration. It is also shown that $\bar{\mathbf{R}}_k = \mathbf{0}$ (note that this is only for the assumption of corrections in the form of simple offsets).

In practice, we start our algorithm by partitioning the band-limited acoustic space. Initially, we set $\bar{\mathbf{r}}_k = \mathbf{0}$; then for each

iteration we use the current value of $\boldsymbol{\mu}_{X,k}$ to estimate $\bar{\mathbf{r}}_k$ according to (21), and use this value to update $\boldsymbol{\mu}_{X,k}$ as in (18). In Section VI-D, we also show results for a similar approach with an initial partitioning of data in the full-bandwidth instead of in the limited-bandwidth space.

D. Compensation of Band-Limited Features

In order to perform decoding of band-limited data, frames need to be classified as belonging to a particular data cluster (or combination of them) and compensation may then be applied accordingly.

1) Phoneme-Based Compensation: Several solutions are proposed for associating speech frames of unknown phonetic content to the defined phonetic classes [27].

Oracle Phoneme-Specific Compensation: Oracle Phoneme-specific correction employs manually time-aligned phonetic transcriptions of test data in order to apply the appropriate corrector function to each frame. This Oracle method is not a practical solution but an upper bound on performance for feature compensation.

General Compensation: Here, the complete distorted feature space is modeled using a single class. This is a simplification of phoneme-based compensation with which phonetic labels are not needed anymore. General feature compensation is a global transformation of the feature space, similar to CMN (the compensation is now computed for all available observations instead of in a per file mode, as is usually done in CMN).

Two-Stage Compensation: General compensation as in the previous section is followed by a pass of an ASR phonetic recognizer that provides a set of tentative transcripts for each test utterance (using N-best or a lattice output). In the second stage, the generated phonetic transcriptions are employed for Phoneme-specific correction. The ASR final decision is chosen as the one with maximum likelihood, among the possible initial transcripts for each utterance.

Compensation Embedded in the Decoder Module: Although out of the general framework of compensation in an independent module, it is possible to apply compensation embedded in the Viterbi decoding process, prior to likelihood computation for each pair observation-acoustic model. For each observation, phoneme-specific compensation is applied before computing the likelihood probability against states that belong to each phonetic model. Our hypothesis is that corrector functions would work best for the correct combination of band-limited feature vector and full-bandwidth phonetic models and would lead to lower likelihoods when the band-limited feature vector does not correspond to the full bandwidth phonetic model. We note the similarity between this approach and constrained linear adaptation [11].

2) Gaussian-Based Compensation: In this case, feature compensation is made according to (17) by weighting the contribution of the compensation for each class by the posterior of that class [28].

3) Multi-Environment Compensation: In real applications, a system may receive speech from a variety of sources with different band-limitations and different corrections should be applied accordingly (e.g., speech downloaded from the Internet may have been recorded at different sampling rates). Here, we hypothesize that different bandwidth limitations leave a distinct footprint in the MFCC feature vector that can be used to identify the type of distortion and so, classification can be made within the compensation framework.

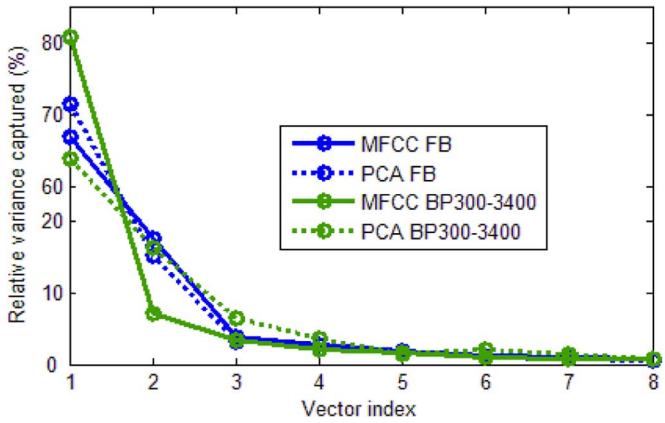


Fig. 5. Relative variance (in %) captured by PCA-derived eigenvectors and MFCC transformation vectors for full-bandwidth and bandpass-filtered data (300–3400 Hz bandpass filter). Data points correspond to Table I.

Gaussian mixture modeling as in (14) is now extended to the case of multiple distortions

$$p(\mathbf{y}) = \sum_{d=1}^{D} \sum_{k^d=1}^{K^d} p\left(\mathbf{y}\,|k^d, d\right) \cdot P\left(k^d\,|d\right) \cdot P(d) = \sum_{d=1}^{D} \sum_{k^d=1}^{K^d} N\left(\mathbf{y}; \boldsymbol{\mu}_{k^d}, \boldsymbol{\Sigma}_{k^d}\right) \cdot P\left(k^d\,|d\right) \cdot P(d). \quad (22)$$

In (22), d refers to a specific band-limitation in a set of D possible distortions and $P(d)$ and $P\left(k^d\,|d\right)$ are *a priori* probabilities of distortion d and mixture k^d given distortion d, respectively. The final GMM is obtained by combination of GMMs from individual environments.

If we denote the set of all the Gaussians in the GMMs corresponding to all the environments as $\Psi = \{\psi\}$, where ψ denotes one of these Gaussians, and define

$$P(\psi) = P\left(k^d, d\right) = P\left(k^d\,|d\right) \cdot P(d) \quad (23)$$

we can rewrite (22) with an expression similar to (14), where the sum over distortions is substituted by an extended GMM from multiple distortions

$$p(\mathbf{y}) = \sum_{\psi} N\left(\mathbf{y}; \mu_\psi, \Sigma_\psi\right) \cdot P(\psi). \quad (24)$$

In synthesis, classes are created independently for each possible distortion and they are combined into a super set of classes for automatic multi-environment compensation. Alternatively it is possible to start pooling together data from different distortions and so, classes are trained in a multidistortion mode (this is similar to multistyle training of acoustic models). However, in our experiments, performance with this approach was worse than training classes for each environment separately, probably because the later assures a partition of the feature space based on the type of distortion.

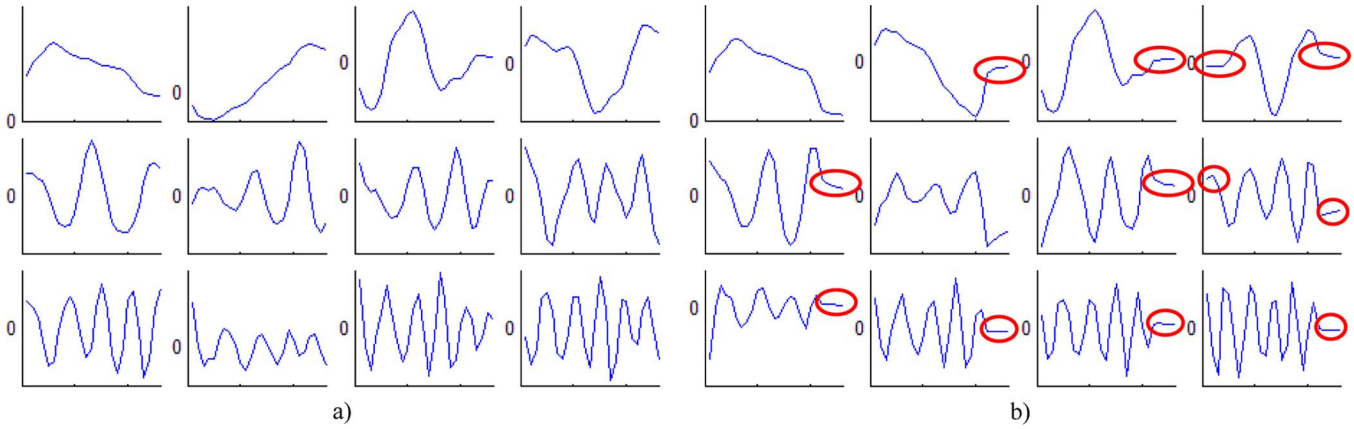


Fig. 6. Eigenvectors for a log-Mel frequency energy representation of (a) full-bandwidth and (b) limited-bandwidth speech (BP300–3400 Hz), derived from the covariance matrix. The first set presents a resemblance with sinusoidal functions, whereas the second set shows deviations, especially near the borders (red circles). In each inset, the x-axis represents Mel-scaled frequencies. In the y-axis, zero-values are shown.

IV. Cepstral Decorrelation and Band-Limiting Distortions

MFCCs are highly uncorrelated, making them a convenient set of features. For example, this allows using diagonal covariance matrices with only small accuracy loss. For the problem of feature compensation, uncorrelation suggests that in practice no information pertaining to a given MFCC coefficient may be obtained from any other coefficient and so, in (17) full compensation matrices $\mathbf{B}_k$ may be substituted by diagonal ones without significant accuracy loss. In fact, this is normally used for noisy speech [5], [12]. Although the same simplification has been adopted in the past for band-limiting distortions [28], [41], we show now that in this case this assumption is not as clearly justified.

The assumption of decorrelation of MFCCs is tied to the relation of this transformation basis to principal component analysis (PCA) [13]. PCA was used in the past as a means for optimal representation of a particular distribution with maximal compactness. This is done extracting the directions of maximum variability of a particular set of data samples; in practice, the eigenvectors of the sample covariance matrix of the distribution (the eigenvalues representing the variability captured by each direction). An additional property of these vectors is that they are uncorrelated and orthogonal. However, PCA computation is expensive and depends on the available data. MFCCs were originally formulated as a fixed transformation that is, nevertheless, similar to PCA [38].

A transformation of spectral features using sinusoidal functions was first proposed by Yilmaz [48], [49]. Pols [38] later showed that for a particular set of speech files, PCA-derived vectors resembled the shapes of sinusoidal functions, and the variabilities captured by each PCA-derived eigenvector and each MFCC were similar, too. In the first two columns of Table I (darker lines in Fig. 5), we show the variability captured by the first MFCCs and PCA-derived eigenvectors in a full-bandwidth (FB; 8 kHz) distribution using TIMIT data. These results are similar to those by Pols [38]. Also, in Fig. 6(a) we show the shapes of the first eigenvectors which roughly resemble sinusoidal functions.

TABLE I
Relative Variance (in %) Captured by PCA-Derived Eigenvectors and MFCC Transformation Vectors for Full-Bandwidth and Bandpass-Filtered Data (300–3400 Hz Bandpass Filter). Vectors in Both Bases Seem to Capture Similar Amount of Variation in the Case of Full-Bandwidth Speech, but not so Much in Band-Limited Speech

Vector basis → / **Vector index ↓**	MFCC FB	PCA FB	MFCC BP300-3400	PCA BP300-3400
1	66.90	71.52	80.85	63.86
2	17.68	15.27	7.14	16.24
3	3.93	3.16	3.38	6.46
4	2.73	2.80	2.07	3.58
5	1.87	1.55	1.69	1.35
6	1.30	0.96	1.00	2.14
7	0.95	0.83	0.78	1.36
8	0.82	0.58	0.68	0.76

The situation changes for band-limiting distortions, because the energy in missing parts of the spectrum is several dBs below that in the unfiltered bands. MFCC extraction over band-limiting channels may be viewed as a whole, as a transformation of full-bandwidth speech where the sinusoidal functions are flattened in the filtered regions. In Fig. 7, we show the theoretical cosine transformation vectors of orders 1 and 3 on top and an equivalent transformation with bandpass filtering in the bottom. The basis that defines the MFCC transformation over band-limited speech is not orthogonal anymore and therefore, we expect MFCCs to resemble less clearly PCA-derived features. Fig. 7 is a simplification because the result of the logarithm in the filtered regions has been zeroed; in reality, the logarithm is never null, but very significant energy attenuation exists in band-limited regions, so the orthogonality of the transformation is still broken.

In the last two columns of Table I (lighter lines in Fig. 5), we show that the relationship between the variability captured by eigenvectors and MFCCs derived from band-limited TIMIT is less obvious than that for full-bandwidth speech, and similarly, in Fig. 6(b) the shapes of the eigenvectors are less similar to sinusoids, particularly near the borders. These observations suggest that MFCC transformations over band-limited speech differ from the PCA-derived transformation vectors much more

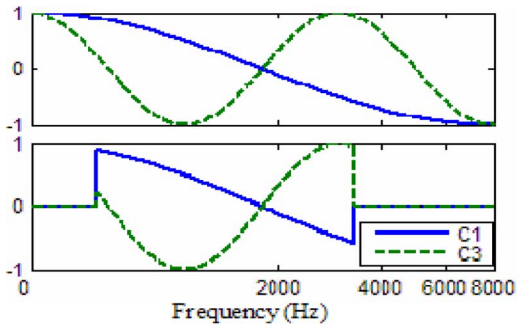


Fig. 7. Cepstral transformations of orders 1 and 3 for full-bandwidth speech (top) and equivalent cepstral transformations for limited-bandwidth speech (bottom: 300–3400 Hz bandpass filter). The band-limited transformation basis is no longer orthogonal.

than with full-bandwidth speech. As a side-effect they will also probably not have the other properties of PCA, namely feature decorrelation.

In order to empirically verify our hypothesis that MFCCs of band-limited speech are more correlated than those of full-bandwidth speech, we define a measure of nondiagonality of a covariance matrix based on the correlation of pairs of MFCC elements

$$\text{nonDiag} = \sum_{i}^{\text{staticMFCCs}} \sum_{j,\, j\neq i}^{\text{MFCCs}} \delta_{ij};$$
$$\text{where } \delta_{ij} = \begin{cases} 1, & \text{if } |\rho(i,j)| \equiv \left| \frac{\text{cov}(i,j)}{\sqrt{\text{cov}(i,i)\cdot\text{cov}(j,j)}} \right| \geq \tau \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

Such metric establishes that a significant correlation exists between two elements of an MFCC vector if their correlation coefficient is larger than a threshold; a binary metric because coefficients with a significant correlation are given a score of 1 and the rest, a score of 0. This was chosen in order to show how many features are more correlated than a threshold. If instead, a continuous measure had been used (like multiplication of correlation coefficients) very low correlation between two terms could hide the fact that a large number of other coefficients are correlated.

Experiments setting the threshold to $\tau = 0.2$ and using all TIMIT training data to compute the covariance matrix produced a nondiagonality of 51 for full-bandwidth MFCC coefficients, 108 for LP4 kHz coefficients and 110 for BP300-3400 Hz coefficients. Actual numbers vary for different values of τ, but the interesting point is that the number of correlated elements in the MFCC vector is more than doubled for band-limiting distortions compared to full-bandwidth speech.

The previous argumentation seems to indicate that when the distortion affecting speech is of a band-limiting type, better reconstruction performance may be obtained by multivariate feature compensation (something also suggested by the decrease in RMSE shown in Fig. 4(a) for multivariate compensation).

V. Experimental Framework

A. Description of ASR Engines

HTK tools are used for training of language models and acoustic models, model adaptation and decoding [52].

The front-end employed uses pre-emphasis filtering ($\alpha = 0.97$) and Hamming windows of 25-ms length and 10-ms shift. Thirteen MFCC coefficients including C0, and their respective first- and second-order derivatives (39 total features) are computed from a filter-bank of 26 Mel-scaled filters distributed in the region 0–8 kHz. Experiments reported are for acoustic model sets with 51 context-independent hidden Markov models (HMMs) and a phone bigram language model. HMM models are trained using TIMIT [15]. For training, we use all 4680 files in the training partition and evaluation is made on all the 1620 files in the test partition.

Scoring metrics employed are phonetic percent correct and percent accuracy

$$\%\text{Correct} = 100\% \cdot \frac{C}{N} = 100\% \cdot \frac{N-S-D}{N} \tag{26}$$
$$\%\text{Accuracy} = 100\% \cdot \frac{C-I}{N} = 100\% \cdot \frac{N-S-D-I}{N} \tag{27}$$

where C represents correct hypotheses, I are insertions, S represents substitutions, D stands for deletions, and N is the total of units in the reference transcripts (i.e., $N = C + S + D$).

Experiments compare performance of the proposed feature compensation approaches with those as follows.

- — *No Compensation*: Acoustic models trained with full-bandwidth data and tested with band-limited data.
- — *Model Adaptation*: Full-bandwidth acoustic models adapted with data from the band-limited condition (unless otherwise specified, global MLLR followed by 28-class MLLR, followed by MAP).
- — *Matched Models*: Acoustic models trained and tested with band-limited data.
- — *CMN*: Models trained with CMN full-bandwidth data and tested with CMN limited-bandwidth data.

B. Speech Corpora

Experiments have been conducted with the DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus [15] (commonly known as TIMIT) for the following reasons: first, TIMIT has been extensively used in the past as a benchmark for algorithm testing. Second, the corpus was recorded under clean conditions and the vocabulary size is small; therefore, high accuracy may be achieved with simple systems and tuning is not complicated. This also simplifies analysis of the impact of different distortions and robustness methods as they affect performance in a more direct manner than with a more sophisticated system. Finally, a number of associated corpora exist as rerecordings of the original data under different distortions, thus making it possible to evaluate the impact of real distortions and allowing the use of stereo-based compensation techniques.

TABLE II
PERFORMANCE OF DIFFERENT LINEAR FEATURE COMPENSATION APPROACHES, COMPARED TO NO COMPENSATION, MODEL ADAPTATION, MATCHED MODELS, AND CMN FOR MULTIPLE REAL AND ARTIFICIAL BAND-LIMITING DISTORTIONS

MODE	DISTORTION	%CORR	%ACC	DISTORTION	%CORR	%ACC
No Compensation	**Full-Bandwidth**	75.40	71.18			
CMN		75.71	71.61			
No Compensation	**LP6kHz**	64.32	58.30	**BP300-3400 Hz**	41.13	32.67
Model Adaptation		75.46	70.85		70.63	64.90
Matched Models		75.45	71.03		71.86	65.73
CMN		74.30	69.95		60.91	54.71
Oracle Phoneme		75.62	71.76		74.31	71.01
2-stage Phon		74.63	70.69		54.44	50.18
Gaussian-based		74.92	70.63		66.48	60.18
No Compensation	**LP4kHz**	55.93	44.67	**STC-TIMIT**	30.98	21.23
Model Adaptation		73.57	68.64		62.63	58.26
Matched Models		74.73	69.33		69.10	61.80
CMN		68.00	62.28		51.59	46.98
Oracle Phoneme		75.35	71.44		71.86	67.56
2-stage Phon		69.05	64.31		47.31	43.14
Gaussian-based		72.61	67.45		56.61	49.72
No Compensation	**LP2kHz**	30.45	26.10	**NTIMIT**	36.15	26.27
Model Adaptation		63.48	57.96		55.96	50.71
Matched Models		68.67	61.57		62.45	53.76
CMN		51.70	45.63		39.62	34.05
Oracle Phoneme		74.52	70.32		67.98	64.61
2-stage Phon		48.91	44.42		28.46	24.96
Gaussian-based		55.73	49.53		40.84	34.14

Experiments on real telephone data employ NTIMIT [21] and STC-TIMIT [32] corpora. In NTIMIT, the original utterances from TIMIT were rerecorded after sending each of them in a separate telephone call (involving all sorts of network combinations). Thus, in effect each file is affected by a different distortion. However, a constraint for the success of feature compensation approaches proposed in this work is the existence of well-defined and steady distortions. This constraint may be alleviated using a multi-environment compensation approach, which nevertheless requires individual training of corrections for well-defined distortions. Therefore, NTIMIT is not well suited for the setup proposed in this work (very scarce data exists from each individual distortion). In order to be able to test feature compensation techniques on real data we created STC-TIMIT, where the whole original corpus was passed through a single telephone call. In the new corpus, an important effort was made for obtaining an accurate alignment with the original files, and as shown in [31] this has a positive impact in stereo-based robustness methods (in NTIMIT we detected an average misalignment of 15 ms, and this may cause a relative accuracy decrease of up to 3% in our experiments). STC-TIMIT is distributed through LDC [26].

VI. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present experimental results for a variety of conditions and problem constraints.

A. Phoneme-Based and Gaussian-Based Univariate Feature Compensation

Table II shows performance of univariate phoneme-based and Gaussian-based compensation strategies for a variety of distortions. Comparison of accuracy in the first row (full-bandwidth training and testing data) with that of matched models for each band-limitation shows that only moderate degradation exists for most band-limiting channels considered. On the contrary, the difference in performance between the cases of No Compensation and matched models shows the very important impact of the mismatch and the need for robustness methods.

The upper bound on feature compensation performance set by the Oracle phoneme-based compensation shows that assuming we could reliably select the appropriate corrector function we could obtain with simple linear univariate feature compensations accuracy close to that obtained with full bandwidth speech. Performance of the two-stage solution implemented is far from that with the Oracle technique, especially in severe distortions because of the poor performance of general compensation (performance is similar to CMN, which is also a form of general correction). On the contrary, Gaussian-based partitioning shows significant improvements compared to CMN and is also close to model adaptation for moderate distortions. However, for more severe distortions, relative performance decreases compared to model-side robustness approaches. The reasons are twofold: 1) a small number of modeling parameters for the compensation (linear compensations in feature compensation compared to full compensation matrixes for MLLR) and 2) univariate compensation is not able to capture all data information useful for reconstruction.

B. Univariate and Multivariate Feature Compensation

It was previously shown that correlation of MFCCs is larger in band-limited data than in full-bandwidth data, and we hypothesized that as a consequence multivariate (instead of univariate) compensation should be used for compensation of band-limiting distortions. Results in Table III corroborate our hypothesis and show more evident improvements for severe distortions. For the more complicated distortions BP300–3400 Hz and STC-TIMIT,

TABLE III
ASR PERFORMANCE USING UNIVARIATE AND MULTIVARIATE GAUSSIAN-BASED COMPENSATION FOR DIFFERENT DISTORTIONS. RESULTS ARE COMPARED TO NO COMPENSATION AND MODEL-SIDE ROBUSTNESS APPROACHES. IN UNIVARIATE AND MULTIVARIATE COMPENSATION, THE NUMBER THAT FOLLOWS INDICATES THE AMOUNT OF CLASSES EMPLOYED FOR BAND-LIMITED SPACE PARTITIONING

MODE	DISTORTION	%CORR	%ACC	DISTORTION	%CORR	%ACC
No Compensation	**Full-Bandwidth**	75.40	71.18			
No Compensation		64.32	58.30		41.13	32.67
Model Adaptation		75.46	70.85		70.63	64.90
Matched Models		75.45	71.03		71.86	65.73
Univariate-32	**LP6kHz**	74.88	70.65	**BP300-3400 Hz**	65.63	58.46
Multivariate-32		75.22	70.95		69.29	63.44
Univariate-256		75.11	70.82		68.08	60.78
Multivariate-256		75.25	70.97		70.62	64.79
No Compensation		55.93	44.67		30.98	21.23
Model Adaptation		73.57	68.64		62.63	58.26
Matched Models		74.73	69.33		69.10	61.80
Univariate-32	**LP4kHz**	72.41	66.97	**STC-TIMIT**	56.03	49.14
Multivariate-32		73.16	68.46		62.53	56.78
Univariate-256		72.79	67.65		60.32	53.38
Multivariate-256		73.69	68.94		64.67	58.79

we show accuracy for different numbers of partitioning classes (32 and 256). In our experiments, we truncated multivariate compensation at an average number of ten elements, whereas for univariate linear compensation two coefficients are used for each cluster and MFCC coefficient (i.e., five times more modeling parameters in multivariate compensation in average). Interestingly, accuracy of multivariate feature compensation with 32 Gaussian clusters is clearly superior to univariate compensation with 256 classes, even when more modeling parameters are used for univariate compensation in this comparison (512 versus $\sim$ 320 for each MFCC coefficient). This shows that the improvements with multivariate compensation are not due only to having more modeling parameters, but better modeling of the compensation. Also, from a computational point of view, the cost of computing posterior likelihoods is much higher than that of increasing the number of elements in the linear compensation. From this perspective, it makes sense to compare performance of both approaches for the same number of Gaussian classes, and in that case the improvement is obviously much greater.

In addition to outperforming univariate compensation, multivariate feature compensation is as accurate as model adaptation, each approach having its advantages; model adaptation does not require stereo-data for training the compensation, but performs better in supervised mode, whereas feature compensation is unsupervised but employs stereo-data (a nonstereo implementation is possible, too).

C. Compensation Embedded in the Decoder Module

In Table IV, we include ASR performance measures as well as the average log-likelihood per utterance (computed by adding the log-likelihood of the winner state sequence for each utterance in the test set) for compensation embedded in the decoder, Oracle, and General Correction. Compensation embedded in the decoder requires compensation of dynamic features (the rest of experiments in this work use linear regression of compensated static features) as a side-effect of the modification of the search algorithm and for the sake of comparison results for Oracle compensation and General compensation in Table IV also use it (this explains differences with Table II).

Unsurprisingly, compensation embedded in the decoder obtains for each distortion the maximum average likelihood, because for each frame this approach applies the compensation that maximizes the likelihood probability. Unfortunately, maximization of the likelihood does not guarantee optimal ASR performance; for example, ASR performance with Oracle compensation is clearly superior in spite of smaller likelihood. This behavior may be explained by the following limit case: take two different phonemes for which training data for a particular feature present great dispersion. In such a case, the polynomial fit for both phonemes may resemble a straight and horizontal line. Then, the result of compensating any observation would not depend on the observation itself, but would be the means of the full-bandwidth features in the training data used in the polynomial fit; a clearly undesirable situation. In order to reduce this problem, we propose the use of multivariate feature compensation. As the fit now depends on a larger number of coefficients in the feature vector, compensated features will be less subject to the problem described. Results in Table IV support this hypothesis.

D. Nonstereo Training of Corrector Functions

As shown in Section III-C2, it is possible to train an offset feature compensation using nonstereo data and an iterative EM strategy. In Fig. 8, we show performance for a low-pass filter distortion with cutoff frequency 4 kHz, for two versions of the algorithm; when Gaussian classes are defined in the limited-bandwidth (LB) or full-bandwidth (FB) spaces. Results show the advantage of training classes directly in the distorted space for more accurate classification of speech frames. Accuracy is only 2% absolute worse than with stereo-based training. However, the solution in (21) is only for univariate feature compensation, which limits the success of feature compensation.

E. Blind Compensation of Corrector Functions

In this section, we study compensation of data from multiple distorting environments seen and unseen during training in a unified multi-environment feature compensation framework. This approach, described in Section III-D3, allows the system to automatically detect the type of distortion (or estimate it in the

TABLE IV
ASR PERFORMANCE USING PHONEME-BASED COMPENSATION EMBEDDED IN THE DECODER, OR OUTSIDE OF THE DECODER (ORACLE AND GENERAL CORRECTION)

MODE		DISTORT.	%CORR	%ACC	AVG. LOG-LIKELY (X100)
Univar	Oracle	LP4kHz	74.53	70.93	-220
	Gen. corr.		67.02	62.60	-226
	Embedded		67.67	61.92	-219
Multivar	Oracle		75.34	71.01	-245
	Gen. corr.		70.70	65.92	-249
	Embedded		72.48	67.49	-220
Univar	Oracle	LP2kHz	69.98	66.30	-208
	Gen. corr.		46.97	41.87	-216
	Embedded		49.14	44.00	-203
Multivar	Oracle		73.51	69.53	-251
	Gen. corr.		51.19	46.21	-256
	Embedded		56.58	51.52	-204

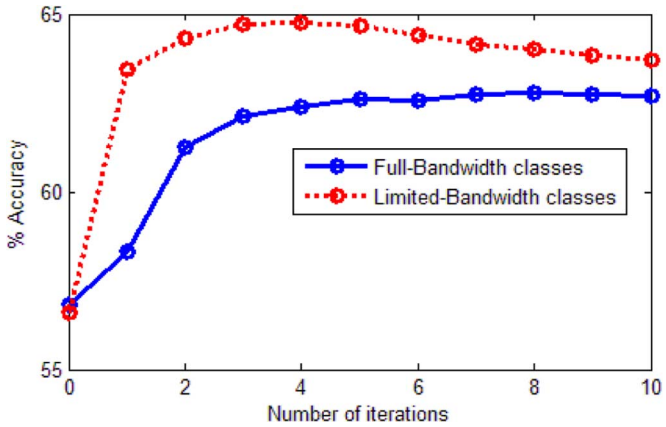


Fig. 8. Accuracy versus number of iterations in two different nonstereo compensation modes for TIMIT LP4 kHz; Gaussian classes defined in the full-bandwidth space and in the limited-bandwidth space.

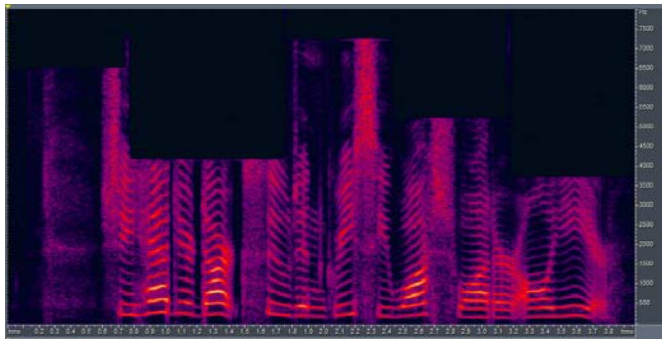


Fig. 9. Spectrogram of file DR1_FAKS0_SA1 from TIMIT after random length filtering with randomly chosen low-pass filters.

TABLE V
INPUT DISTORTION CLASSIFICATION ACCURACY (IN PERCENTAGE OF FRAMES) USING AUTOMATIC ENVIRONMENT CLASSIFICATION IN FEATURE COMPENSATION UNDER THE SETUP DESCRIBED IN SECTION VI-E. FOR OBSERVED BANDWIDTHS "HIT" CORRESPONDS TO CORRECT CHOICE AND "±1" IS THE SUM OF "HIT" PLUS THE CASES WHERE FIRST CHOICE BELONGS TO THE IMMEDIATELY PREVIOUS OR POSTERIOR OBSERVED DISTORTIONS. FOR UNOBSERVED BANDWIDTHS "±1" SHOWS THE PERCENTAGE OF CASES WHEN THE FIRST CHOICE BELONGS TO THE IMMEDIATELY PREVIOUS AND POSTERIOR DISTORTIONS AVAILABLE DURING TRAINING

OBSERVED FILTERS	HIT (%)	+/- 1 (%)	UNOBSERVED FILTERS	+/- 1 (%)
O1: 8000 Hz	68.78	97.70	U1: 7588 Hz	97.20
O2: 7196 Hz	58.35	96.92	U2: 6823 Hz	86.01
O3: 6467 Hz	83.20	97.78	U3: 6128 Hz	93.25
O4: 5805 Hz	77.52	96.42	U4: 5497 Hz	93.27
O5: 5204 Hz	88.07	98.45	U5: 4925 Hz	96.96
O6: 4659 Hz	88.26	98.69	U6: 4405 Hz	96.66
O7: 4164 Hz	90.52	99.74	U7: 3933 Hz	99.78
O8: 3714 Hz	98.66	99.91		

best possible way) and compensate it, keeping active a single set of acoustic models at all times [29], [30].

We designed an experiment where TIMIT files suffer abrupt changes in available bandwidth (with random values for both the cutoff frequency of the low-pass filter in each chunk and its length), as in Fig. 9. Fifteen different low-pass filters were considered, with cutoff frequencies distributed in a Mel-scale (Table V). Training of corrector functions was made only for eight of these filters, leaving 1 unseen distortion between each of the distortions used for training (seven unseen distortions in total). The success of multi-environment compensation may be initially estimated evaluating the success of automatic classification of the distortion (the distortion containing the Gaussian class with best posterior probability for each frame). In Table V, we show that distortions seen during training (*O#*) are accurately identified and therefore we expect successful compensation (also, misclassified frames are identified as belonging to the immediately superior or inferior distortion which would not cause significant errors in the compensation). Analysis of classification errors showed that most errors are located in silent parts of files or correspond to isolated frames; an effect that may be mitigated by using smoothing windows over consecutive frames [30]. The Table also shows classification of band-limitations unseen during training (*U#*), which are in most cases accurately classified as one of the immediately superior or inferior distortions available during training.

In Table VI, we show the superiority, in this setup, of feature compensation (first row) compared to adaptation of acoustic models using data from any individual distortion (intermediate rows) as well as models adapted using data from all seen distortions (final row). Performance could be improved in model-side robustness by using a distortion classifier as a selector for acoustic models, or using several speech recognizers in parallel, but these solutions require significant modifications to the recognition framework, and would also increase computational costs and memory load.

As the previous experiment is somehow biased towards feature compensation approaches, we propose a further experiment to compare performance of our approach with that of model-side approaches. Here, we assume test data are band-limited with a fixed channel from the unseen channels *U#* defined in Table V. In Fig. 10, we compare performance of multi-environment feature compensation trained only with the observed distortions O# in Table V versus performance of acoustic models adapted in each case with data from either the immediately superior or inferior observed band-limitations. For the less severe distortions, in which the mismatch between the observed and unobserved channels is smaller feature compensation clearly outperforms

TABLE VI
ASR PERFORMANCE ON TIMIT CORPUS DISTORTED WITH ALL LOW-PASS DISTORTIONS (SEEN AND UNSEEN) IN TABLE V, IN DIFFERENT FRAGMENTS OF RANDOM LENGTH (AS IN FIG. 9). FEATURE COMPENSATION USES AUTOMATIC BANDWIDTH CLASSIFICATION. MODEL ADAPT RESULTS ARE GIVEN FOR ACOUSTIC MODELS ADAPTED WITH DATA FROM A PARTICULAR DISTORTION AMONG THE OBSERVED ONES AND IN THE FINAL ROW FOR MODELS ADAPTED WITH DATA FROM ALL THE OBSERVED DISTORTIONS

MODE	%CORR	%ACC
Feature Compensation	73.54	68.47
Model Adapt: O1	53.60	48.61
Model Adapt: O2	57.30	51.98
Model Adapt: O3	59.34	53.23
Model Adapt: O4	60.62	53.95
Model Adapt: O5	62.20	54.68
Model Adapt: O6	60.93	53.01
Model Adapt: O7	63.52	55.35
Model Adapt: O8	62.55	54.24
Model Adapt: All	68.04	61.32

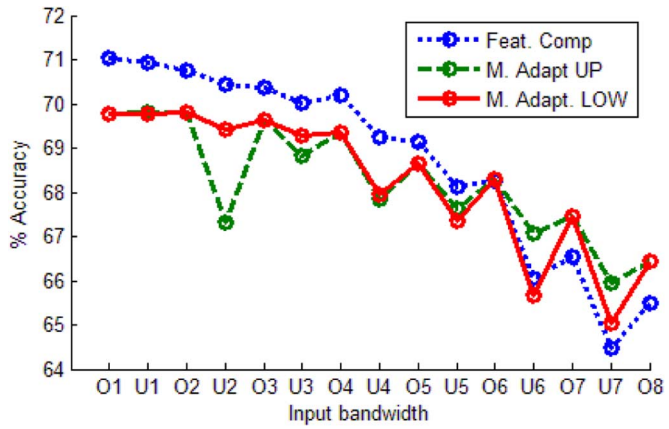


Fig. 10. ASR accuracy for feature compensation and model adaptation in the experiment described in Section VI-E. The x-axis indicates the filter corrupting data. Distortions labeled O# are observed during training, and those labeled U# are unobserved.

model adaptation, probably benefiting from the possibility of combining compensations from either the immediately upper or lower observed band-limitations in a per-frame basis. On the contrary, for more severe distortions where the mismatch is larger model adaptation is more accurate. However, this experiment is biased towards model adaptation because it assumes correctly identified fixed distortions. Model adaptation performance would be degraded in a variable distortion situation and additionally an important modification to the structure of the recognizer would be needed.

These experiments show that provided sufficient resolution during training, any low-pass filter within the considered range may be efficiently dealt with using feature compensation, even variable distortions, smooth or abrupt and of duration as small as a few frames.

F. Compensation With Limited Amounts of Training Data

In previous sections, we employed large amounts of training data for training of both, feature compensation and model adaptation schemes. In this section, we compare performance in situations of data scarcity. In Fig. 11, we show accuracy of multivariate feature compensation using 2, 32, and 256 partitioning

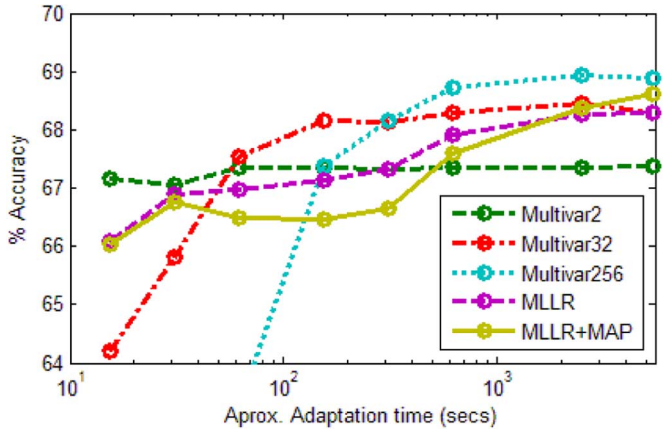


Fig. 11. ASR accuracy versus available adaptation data for feature compensation and model adaptation. Distortion is LP4 kHz.

classes, respectively, and model adaptation (we show results for MLLR-only adaptation, and MLLR followed by MAP).

It is clear that multivariate feature compensation suffers the problem of under-training for large numbers of partitioning classes and very limited training data. This problem could be solved by applying a minimum occupancy threshold, as in fact is done in MLLR in this experiment. Using such threshold in the creation of partitioning classes we would expect to achieve for each experimental point a result similar to that of the best performing of the three multivariate compensation results shown and so, multivariate feature compensation would outperform model adaptation in all cases (differences tend to decrease for larger amounts of data due to the successful combination of MLLR and MAP and for large enough amounts of data similar performance is achieved).

These results show that feature compensation is an excellent solution for situations of data scarcity.

G. Combination of Different Approaches

The proposed feature compensation approaches may be combined with other robustness methods for improved accuracy, as shown in Table VII.

Feature compensation on CMN features does not provide significant improvements. This was expected because CMN is approximately a simplification of feature compensation; combination of both approaches is redundant and performance is similar to that with feature compensation only. On the contrary, combination of model-side robustness methods (model adaptation or model retraining) and feature compensation achieves significant improvements compared to the best individual method in each case. Combination of both approaches performs better than feature compensation alone because the system is able to recover from the mismatch caused by artificial distortions introduced by the feature compensation method. Also, compared to the case of model-side robustness only, usage of feature compensation plays an important role in reducing data variability and the resulting acoustic space is simpler to model. Of particular interest is the combination of feature compensation and matched models (i.e., compensating features to produce pseudo full-bandwidth features and train models on those features), which clearly outperforms any of the individual approaches alone.

TABLE VII
COMPARISON OF DIFFERENT INDIVIDUAL APPROACHES (MODEL ADAPTATION, MATCHED MODELS, CMN AND MULTIVARIATE FEATURE COMPENSATION) WITH COMBINED METHODS

MODE	DISTORTION	%CORR	%ACC
No Compensation	**Full-**	75.40	71.18
CMN	**Bandwidth**	75.71	71.61
No Compensation		41.13	32.67
Model Adapt		70.63	64.90
Matched Models		71.86	65.73
CMN	**BP300-**	60.91	54.71
Feature Compensation	**3400 Hz**	70.62	64.79
CMN + Feature Comp.		70.12	64.31
M. Adapt + Feature Comp.		70.66	65.14
Matched M. + Feature Comp.		73.05	66.87
No Compensation		30.98	21.23
Model Adapt		62.63	58.26
Matched Models		69.10	61.80
CMN	**STC-**	51.59	46.98
Feature Compensation	**TIMIT**	64.67	58.79
CMN + Feature Comp.		64.80	58.66
M. Adapt + Feature Comp.		66.25	60.12
Matched M. + Feature Comp.		71.32	63.96

H. Discussion and Further Experimental Evidence

Due to space constraints, we presented in previous sections a selection of numerous experimental results available (other results may be found in [31]). A summary of other interesting experimental results is given here.

As an extension to univariate linear compensation, we also evaluated compensation as a univariate polynomial expansion. In Fig. 4(b), we already showed that MSE does not significantly improve for polynomial expansions larger than order 1. Experiments conducted on ASR of band-limited data compensated with such polynomial expansions showed little difference in performance, and worse results for larger orders, due to large compensation errors for outlier frames.

An interesting point in feature compensation concerns the convenience of compensation of dynamic features or reconstruction of these using the usual linear regression of reconstructed static features. In our experiments, we observed that MSE was smaller between full-bandwidth feature vectors and band-limited speech vectors compensated with dynamic feature compensation. This is not surprising because the criterion used for training corrections is precisely minimization of MSE. However, this creates *unrealistic* feature vectors because dynamic features are not directly related to the static ones (dynamic features are usually computed from the static ones) and the result is a mismatch between feature vectors and models that in our experiments degraded ASR performance.

In [2], a new iterative method for partitioning the acoustic space was employed with the goal of finding clusters of data that follow a similar distortion and may therefore be compensated using the same transformations. This seems in principle more adequate than a partitioning strategy that simply maximizes the probability of observation in the limited-bandwidth space. Applying the same strategy for the case of band-limited speech we observed significant improvements in percent correct but not in percent accuracy, indicating that the number of phoneme insertions is notably incremented. More experiments are required in this field, as finding an optimal partitioning criterion for the problem of feature compensation might provide significant improvements.

VII. SUMMARY AND CONCLUSION

Feature compensation has been proposed for the problem of ASR of band-limited test data and full-bandwidth acoustic models. In Section II, we present a novel mathematical model of the distortions introduced in a typical front-end as a result of bandwidth limitations (with particular focus on MFCCs). From it, we derive a number of approaches for estimation of full-bandwidth features from limited-bandwidth features that are organized in a unified feature compensation framework. Furthermore, in Section IV we presented an original theoretical and empirical discussion on how band-limitations modify the correlation of cepstral features and hypothesized the very superior performance of nondiagonal compensation matrixes. This hypothesis was confirmed in the experimental section. In Section VI, a large number of experimental setups and problem constraints are studied in which constitutes to our knowledge the most extensive study on the problem of feature compensation for band-limitations. In all cases, performance of feature compensation is benchmarked against the more usual approach of model-side robustness.

A practical requirement common to all feature compensation solutions is the need to define partitioning classes in an offline step (with the exception of General Compensation), whereas training of corrector functions may be done offline or in decoding time. Similarly, model retraining is performed offline, whereas model adaptation can be offline or on-the-fly. Additionally, training is supervised in phoneme-based feature compensation methods and in model-side solutions, whereas Gaussian-class based feature compensation is free of this constraint. Finally, concerning required modifications to the basic architecture of an ASR system, two-stage feature compensation is offline and compensation embedded in the decoder requires performing feature compensation using corrector functions trained for each phoneme before likelihood computation. On the other hand, General feature compensation and Gaussian-class based compensation may be performed independently from the rest of the recognizer modules, inserted between the front-end and the decoder, even for the cases of multiple or time-varying distortions. This is a great advantage over model-side approaches that in order to deal with multiple and time-varying distortions require the simultaneous use of several acoustic model sets and a frequency detector module, or a 3-D type decoder architecture.

Feature compensation approaches proved to be useful both in terms of accuracy (similar to that with model-adaptation) and for their suitability in particular situations, like systems subject to multiple distortions, rapidly varying distortions, etc. In addition, we showed potential better performance than MLLR and MLLR+MAP in situations of training data scarcity, and the possibility of combining feature-side and model-side approaches for better performance than model-side approaches alone (including retraining models for band-limited data, which is currently the most common solution to the problem of bandwidth limitation).

Among the possible lines of future improvement of our approach, an interesting topic is the optimization of the partitioning method for the goal of feature compensation, in order

to find clusters of data better suited for a linear compensation framework. Another interesting line of work is that of training methods that do not require stereo-data. In the future, we expect to derive a nonstereo strategy for multivariate training compensation. An alternative is to obtain an estimate of the channel distortion and apply this to clean data in order to generate pseudostereo data as was done in [18].

APPENDIX
DERIVATION OF THE UPDATE EQUATIONS FOR TRAINING CORRECTOR FUNCTIONS WITH NONSTEREO DATA

It can be shown that maximization of (20) for all available data is equivalent to iteratively maximizing an auxiliary function Q defined as

$$Q\left(\phi,\bar{\phi}\right)=\sum_{t=1}^{T}\sum_{k=1}^{K}\frac{p\left(\mathbf{x}_t,k\left|\phi\right.\right)}{p\left(\mathbf{x}_t\left|\phi\right.\right)}\log\left(p\left(\mathbf{x}_t,k\left|\bar{\phi}\right.\right)\right) \tag{28}$$

where ϕ and $\bar{\phi}$ are the sets of parameters for Gaussian mixtures in the previous and current iteration, respectively. Substituting (18) and (19) in (28), expanding the term of posterior probability and simplifying, we obtain

$$\begin{aligned}Q\left(\phi,\bar{\phi}\right)&=\text{const.}+\sum_{t=1}^{T}\sum_{k=1}^{K}p\left(k\left|\mathbf{x}_t,\phi\right.\right)\\&\quad\cdot\left\{-\frac{1}{2}\log\left|\boldsymbol{\Sigma}_{Y,k}+\bar{\mathbf{R}}_k\right|-\frac{1}{2}\left(\mathbf{x}_t-\boldsymbol{\mu}_{Y,k}-\bar{\mathbf{r}}_k\right)^T\right.\\&\qquad\left.\times\left(\boldsymbol{\Sigma}_{Y,k}+\bar{\mathbf{R}}_k\right)^{-1}\left(\mathbf{x}_t-\boldsymbol{\mu}_{Y,k}-\bar{\mathbf{r}}_k\right)\right\}.\end{aligned} \tag{29}$$

Now, differentiating with respect to $\bar{\mathbf{r}}_k$ and $\bar{\mathbf{R}}_k$ and equating to zero, we obtain solutions that maximize Q. Differentiating is simple if we assume diagonal covariance and compensation matrixes, as shown in (30) and (31) at the bottom of the page, where $\text{diag}\{\mathbf{A}\}$ is a vector containing diagonal elements in matrix $\mathbf{A}$, and $\mathbf{0}$, a vector of zeros.

Solving for the corrector coefficients, we get (32) and (33), as shown at the bottom of the page, where $\odot$ is the product of two matrixes term by term.

These solutions are equivalent to those in RATZ, originally designed for noise compensation, with the difference that classes in RATZ were defined in the clean space [33].

Assuming now that the effect of the distortion in each class is simply an offset for each observation, the vectors of means and covariances are modified as

$$\boldsymbol{\mu}_{X,k}=E\left\{\mathbf{x}_k\right\}=E\left\{\mathbf{b}_k+\mathbf{y}_k\right\}=\mathbf{b}_k+\boldsymbol{\mu}_{Y,k} \tag{34}$$

$$\begin{aligned}\boldsymbol{\Sigma}_{X,k}&=E\left\{\left(\mathbf{x}_k-\boldsymbol{\mu}_{X,k}\right)^2\right\}\\&=E\left\{\left(\mathbf{b}_k+\mathbf{y}_k-\mathbf{b}_k-\boldsymbol{\mu}_{Y,k}\right)^2\right\}\\&=E\left\{\left(\mathbf{y}_k-\boldsymbol{\mu}_{Y,k}\right)^2\right\}=\boldsymbol{\Sigma}_{Y,k}.\end{aligned} \tag{35}$$

Identifying (34) and (35) with (18) and (19), we obtain

$$\mathbf{r}_k=\mathbf{b}_k \tag{36}$$

$$\mathbf{R}_k=\mathbf{0}. \tag{37}$$

Therefore, the extra constraint of simple offsets suppresses (33) in the EM algorithm, so covariance matrixes are not updated anymore and $\mathbf{r}_k$, as defined in (32) will be the offset vector to apply for feature compensation.

$$\frac{\partial Q}{\partial\bar{\mathbf{r}}_k}=\sum_{t=1}^{T}p\left(k\left|\mathbf{x}_t,\phi\right.\right)\cdot\left(\boldsymbol{\Sigma}_{Y,k}+\bar{\mathbf{R}}_k\right)^{-1}\cdot\left(\mathbf{x}_t-\boldsymbol{\mu}_{Y,k}-\bar{\mathbf{r}}_k\right)=\mathbf{0} \tag{30}$$

$$\begin{aligned}\frac{\partial Q}{\partial\bar{\mathbf{R}}_k}&=-\frac{1}{2}\cdot\sum_{t=1}^{T}p\left(k\left|\mathbf{x}_t,\phi\right.\right)\\&\quad\cdot\text{diag}\left\{\left(\boldsymbol{\Sigma}_{Y,k}+\bar{\mathbf{R}}_k\right)^{-1}-\left(\mathbf{x}_t-\boldsymbol{\mu}_{Y,k}-\bar{\mathbf{r}}_k\right)^T\cdot\left(\boldsymbol{\Sigma}_{Y,k}+\bar{\mathbf{R}}_k\right)^{-2}\cdot\left(\mathbf{x}_t-\boldsymbol{\mu}_{Y,k}-\bar{\mathbf{r}}_k\right)\right\}=\mathbf{0}\end{aligned} \tag{31}$$

$$\bar{\mathbf{r}}_k=\frac{\sum_{t=1}^{T}p\left(k\left|\mathbf{x}_t,\phi\right.\right)\cdot\mathbf{x}_t}{\sum_{t=1}^{T}p\left(k\left|\mathbf{x}_t,\phi\right.\right)}-\boldsymbol{\mu}_{Y,k} \tag{32}$$

$$\bar{\mathbf{R}}_k=\frac{\sum_{t=1}^{T}p\left(k\left|\mathbf{x}_t,\phi\right.\right)\cdot\mathbf{I}\odot\left\{\left(\mathbf{x}_t-\boldsymbol{\mu}_{Y,k}-\bar{\mathbf{r}}_k\right)\cdot\left(\mathbf{x}_t-\boldsymbol{\mu}_{Y,k}-\bar{\mathbf{r}}_k\right)^T\right\}}{\sum_{t=1}^{T}p\left(k\left|\mathbf{x}_t,\phi\right.\right)}-\boldsymbol{\Sigma}_{Y,k} \tag{33}$$

REFERENCES

[1] , H. Abut, J. H. L. Hansen, and K. Takeda, Eds., *DSP for In-Vehicle and Mobile Systems*. New York: Kluwer/Springer-Verlag, 2004.

[2] M. Afify, X. Cui, and Y. Gao, "Stereo-based stochastic mapping for robust speech recognition," in *Proc. ICASSP*, HI, Apr. 2007, vol. 4, pp. 377–380.

[3] J. B. Allen, "How do humans process and recognize speech?," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 567–577, Oct. 1994.

[4] C. Avendano, H. Hermansky, and E. A. Wan, "Beyond Nyquist: Towards the recovery of broad-bandwidth speech from narrow-bandwidth speech," in *Proc. EuroSpeech*, Madrid, Spain, Sep. 1995, pp. 165–168.

[5] L. Buera, E. Lleida, A. Miguel, A. Ortega, and O. Saz, "Cepstral vector normalization based on stereo data for robust speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1098–1113, Mar. 2007.

[6] Y. M. Cheng, D. O'Shaughnessy, and P. Mermelstein, "Statistical recovery of wideband speech from narrowband speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 544–548, Oct. 1994.

[7] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, Signal Process.*, vol. ASSP-28, no. 4, pp. 357–366, Aug. 1980.

[8] P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. 39, pp. 1–38, 1977.

[9] L. Denenberg, H. Gish, M. Meteer, T. Miller, J. R. Rohlicek, W. Sadkin, and M. Siu, "Gisting conversational speech in real time," in *Proc. ICASSP*, Minneapolis, MN, Apr. 1993, vol. 2, pp. 131–134.

[10] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 412–421, May 2005.

[11] V. V. Digalakis, D. Rtischev, and L. G. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 357–366, Sep. 1995.

[12] J. Droppo, A. Acero, and L. Deng, "Evaluation of the SPLICE algorithm on the Aurora2 database," in *Proc. EuroSpeech*, Aalborg, Denmark, Sep. 2001, pp. 217–220.

[13] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley-Interscience, 2000, pp. 115–117.

[14] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)," in *Proc. ASRU*, Santa Barbara, CA, Dec. 1997, pp. 347–354.

[15] W. M. Fisher, R. Doddington, and K. M. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," in *Proc. DARPA Workshop Speech Recognition*, Feb. 1986, pp. 93–99.

[16] J. H. L. Hansen, B. Zhou, M. Akbacak, R. Sarikaya, and B. Pellom, "Audio stream phrase recognition for a national gallery of the spoken word: One small step," in *Proc. ICSLP*, Beijing, China, Oct. 2000, vol. 3, pp. 1089–1092.

[17] J. H. L. Hansen, R. Huang, P. Mangalath, B. Zhou, M. Seadle, and J. Deller, "SPEECHFIND: Spoken document retrieval for a national gallery of the spoken word," in *Proc. NORSIG*, Espoo, Finland, Jun. 2004, pp. 1–4.

[18] T. H. Hsieh and J. W. Hung, "Speech feature compensation based on pseudo stereo codebooks for robust speech recognition in additive noise environments," in *Proc. InterSpeech*, Antwerp, Belgium, Aug. 2007, pp. 242–245.

[19] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Upper Saddle River: Prentice-Hall, 2001, pp. 102–104.

[20] X. Huang, A. Acero, and H. W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Upper Saddle River: Prentice-Hall, 2001, pp. 526–526.

[21] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proc. ICASSP*, Albuquerque, NM, Apr. 1990, vol. 1, pp. 109–112.

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

[23] W. Kim and J. H. L. Hansen, "Missing-feature reconstruction for band-limited speech recognition," in *Proc. InterSpeech*, Pittsburgh, PA, Sep. 2006, pp. 2306–2309.

[24] J. Kontio, L. Laaksonen, and P. Alku, "Neural network-based artificial bandwidth expansion of speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 873–881, Mar. 2007.

[25] Y. F. Liao, J. S. Lin, and W. H. Tsai, "Bandwidth mismatch compensation for robust speech recognition," in *Proc. EuroSpeech*, Geneva, Switzerland, Sep. 2003, pp. 3093–3096.

[26] Linguistics Data Consortium (LDC). [Online]. Available: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2008S03.

[27] N. Morales, J. H. L. Hansen, and D. T. Toledano, "MFCC compensation for improved recognition of filtered and band-limited speech," in *Proc. ICASSP*, Philadelphia, PA, Mar. 2005, vol. 1, pp. 521–524.

[28] N. Morales, D. T. Toledano, J. H. L. Hansen, J. Colás, and J. Garrido, "Statistical class-based MFCC enhancement of filtered and band-limited speech for robust ASR," in *Proc. InterSpeech*, Lisbon, Portugal, Sep. 2005, pp. 2629–2632.

[29] N. Morales, D. T. Toledano, J. H. L. Hansen, J. Garrido, and J. Colás, "Unsupervised class-based feature compensation for time-variable bandwidth-limited speech," in *Proc. ICASSP*, Toulouse, France, May 2006, vol. 1, pp. 533–536.

[30] N. Morales, D. T. Toledano, J. H. L. Hansen, and J. Colás, "Blind feature compensation for time-variant band-limited speech recognition," *IEEE Signal Process. Lett.*, vol. 14, no. 1, pp. 70–73, Jan. 2007.

[31] N. Morales, "Robust speech recognition under band-limited channels and other channel distortions," Ph.D. dissertation, Comput. Eng. Dept., Univ. Autonoma de Madrid, Madrid, Spain, 2007.

[32] N. Morales, J. Tejedor, J. Garrido, J. Colás, and D. T. Toledano, "STC-TIMIT: Generation of a single-channel telephone corpus," in *Proc. LREC*, Marrakech, Morocco, May 2008, pp. 391–395.

[33] P. J. Moreno, "Speech Recognition in Noisy Environments," Ph.D. dissertation, Elect. Comput. Eng. Dept., Carnegie Mellon Univ., Pittsburgh, PA, 1996.

[34] National Gallery of the Spoken Word. [Online]. Available: http://www.ngsw.org.

[35] H. Nyquist, "Certain topics in telegraph transmission theory," *Proc. IEEE*, vol. 90, no. 2, pp. 280–305, Feb. 2002.

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

[37] D. Pallett, "A look at NIST's benchmark ASR tests: Past, present, and future," in *Proc. ASRU*, U.S. Virgin Islands, Dec. 2003, pp. 483–488.

[38] L. C. W. Pols, "Spectral Analysis and Identification of Dutch Vowels in Monosyllabic Words," Ph.D. dissertation, Free Univ., Amsterdam, The Netherlands, 1977.

[39] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C: The Art of Scientific Computing*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 1992, pp. 656–680.

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

[41] M. Seltzer, A. Acero, and J. Droppo, "Robust bandwidth extension of noise-corrupted narrowband speech," in *Proc. InterSpeech*, Lisbon, Portugal, Sep. 2005, pp. 1509–1512.

[42] J. J. Sroka and L. D. Braida, "Human and machine consonant recognition," *Speech Commun.*, vol. 45, no. 4, pp. 401–423, Apr. 2005.

[43] Transmission performance characteristics of pulse code modulation channels. ITU-T Rec. G. 712 (11/01)..

[44] A. P. Varga and R. K. Moore, "Hidden Markov model decomposition of speech and noise," in *Proc. ICASSP*, Albuquerque, NM, Apr. 1990, vol. 2, pp. 845–848.

[45] R. M. Warren, K. R. Riener, J. A. Bashford, and B. S. Brubaker, "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," *Percept. Psychophys.*, vol. 57, no. 2, pp. 175–182, 1995.

[46] B. D. Womack and J. H. L. Hansen, "N-channel hidden Markov models for combined stress speech classification and recognition," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 6, pp. 668–677, Nov. 1999.

[47] H. Yasukawa, "Restoration of wide band signal from telephone speech using linear prediction error processing," in *Proc. ICSLP*, Philadelphia, PA, Oct. 1996, vol. 2, pp. 901–904.

[48] H. Yilmaz, "A theory of speech perception," *Bull. Math. Biophys.*, vol. 29, no. 4, pp. 793–825, 1967.

[49] H. Yilmaz, "Statistical theory of speech perception," in *Proc. Conf. Speech Commun. Process.*, 1972, pp. 226–229.

[50] N. B. Yoma and M. Villar, "Speaker verification in noise using a stochastic version of the weighted Viterbi algorithm," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 3, pp. 158–166, Mar. 2002.

[51] N. B. Yoma, I. Brito, and J. Silva, "Language model accuracy and uncertainty in noise canceling in the stochastic weighted Viterbi algorithm," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 2193–2196.

[52] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (HTK Version 3.3)*. Cambridge, U.K.: Cambridge Univ., Eng. Dept., 2005, pp. 61–61.

[53] B. Zheng and R. B. Bapat, "Generalized AT,S and A rank equation," *Appl. Math. Comput.*, vol. 155, no. 2, pp. 407–415, Aug. 2004.

Nicolás Morales (S'04–M'08) received the B.S. degree in physics from the Universidad Autónoma de Madrid (UAM), Madrid, Spain, in 2002 and the Ph.D. degree in telecommunication engineering from UAM in 2007.

From 2003 to 2007, he was a Lab Assistant Professor at the Polytechnic School of UAM. From 2004 to 2005, he spent 12 months as a Visiting Research Student at the University of Colorado at Boulder and the University of Texas at Dallas. From 2006 to 2007, he spent nine months as a Research Intern at IBM Research, Yorktown Heights, New York. Since 2008, he has been a Research Scientist at the group of Acoustic Modelling Research, Nuance Communications, Aachen, Germany. He is the first author of the speech corpus STC-TIMIT, distributed through the LDC catalogue. His research interests include robust speech recognition, acoustic modelling, signal processing, and creation and maintenance of speech resources.

Dr. Morales has served as reviewer for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, as well as for the LREC'08 International Conference.

Doroteo Torre Toledano (M'04) received the telecommunication engineering degree from the Universidad Politecnica de Madrid, Madrid, Spain, in 1997, obtaining the best academic records of his class, and the Ph.D. degree in telecommunication engineering from the same university in 2001, receiving a Ph.D. Dissertation Award from the Spanish Association of Telecommunication Engineers.

He was with the Speech Technology Division of Telefonica R&D from 1994 to 2001. From 2001 to 2002, he was with the Spoken Language Systems Group, Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, as a Postdoctoral Research Associate. After another short period at the Speech Technology Division of Telefonica R&D, he moved in 2004 to the Universidad Autonoma de Madrid, where he is currently an Associate Professor. His current research interests include acoustic modeling, speaker and language recognition, and multimodal biometrics. He has over 60 scientific publications in these fields including journals and conference proceedings.

Dr. Toledano served as a member of the scientific committee of several international conferences as well as a reviewer for several journals such as the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, IEEE SIGNAL PROCESSING LETTERS, *Speech Communication*, and *Computer, Speech and Language*.

John H. L. Hansen (S'81–M'82–SM'93–F'07) received the B.S.E.E. degree from the College of Engineering, Rutgers University, New Brunswick, NJ, in 1982, and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1988 and 1983, respectively.

He joined the Erik Jonsson School of Engineering and Computer Science, University of Texas at Dallas (UTD), Richardson, in Fall 2005, where he is a Professor and Department Chairman of Electrical Engineering, and holds the Distinguished University Chair in Telecommunications Engineering. He also holds a joint appointment as Professor in the School of Brain and Behavioral Sciences (Speech and Hearing). At UTD, he established the Center for Robust Speech Systems (CRSS), which is part of the Human Language Technology Research Institute. Previously, he served as Department Chairman and Professor in the Department of Speech, Language, and Hearing Sciences (SLHS), and Professor in the Department of Electrical and Computer Engineering, at the University of Colorado at Boulder (1998–2005), where he cofounded the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory (RSPL) and continues to direct research activities at the CRSS at UTD.

Prof. Hansen was named IEEE Fellow for contributions in "Robust Speech Recognition in Stress and Noise" in 2007 and is currently serving as Member of the IEEE Signal Processing Society Speech Technical Committee and Educational Technical Committee. Previously, he has served as Technical Advisor to the U.S. Delegate for NATO (IST/TG-01), IEEE Signal Processing Society Distinguished Lecturer (2005/2006), Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1999), Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (1998–2000), and Editorial Board Member for the *IEEE Signal Processing Magazine* (2001–2003). He has also served as a Guest Editor of the October 1994 Special Issue on Robust Speech Recognition for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He has served on the Speech Communications Technical Committee for the Acoustical Society of America (2000–2003), and is serving as a member of the ISCA (International Speech Communications Association) Advisory Council. His research interests span the areas of digital speech processing, analysis, and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human–computer interaction. He has supervised 40 (18 Ph.D., 22 M.S.) thesis candidates, was recipient of the 2005 University of Colorado Teacher Recognition Award as voted by the student body, and author/coauthor of 288 journal and conference papers in the field of speech processing and communications, coauthor of the textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), coeditor of *DSP for In-Vehicle and Mobile Systems* (Springer, 2004), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), and lead author of the report "The impact of speech under 'stress' on military speech technology," (NATO RTO-TR-10, 2000). He also organized and served as General Chair for ICSLP/Interspeech-2002: International Conference on Spoken Language Processing, September 16–20, 2002, and will serve as Technical Program Chair for IEEE ICASSP-2010, Dallas, TX.

Javier Garrido (M'97) was born in Madrid, Spain, in 1954. He received the B.Sc. degree, the M.Sc. degree, and the Ph.D. degree in physics from the Universidad Autónoma de Madrid (UAM), Madrid, Spain, 1974, 1976, and 1984, respectively.

He belongs to the faculty staff at the Computer Science Department, Polytechnic School (EPS),UAM. In 1976, he joined the Group of Semiconductors, UAM, where he participated in research projects related to semiconductor optical devices and their application in waveguides. He has been a Professor at UAM since 1986. Since 1992, he has participated in the implementation of the Computer Science (1992) and Telecommunication (2002) engineering studies at UAM. From 1997 to 2001, he served as the Dean of the Polytechnic School. From his incorporation to EPS, he has extended his research interests to topics related to HW/SW applications on embedded systems (microcontrollers, microprocessors, FPGAs, and SOC devices) such as platforms for wireless sensor networks (WSN) or robotic sensor agents (RSA). In 2003, he cofounded the HCTLab Group, (Human Computer Technology Laboratory) of which he is currently Director. In the HCTLab, he has extended his research interests to the fields of speech recognition and synthesis, with special application to biometric techniques. Also, he has participated in R&D projects related to new technologies with applications for disabled people. Since 1978, he has published more than 40 articles in peer-reviewed journals and 60 papers in archived conference proceedings.