Estimating Cepstrum of Speech Under the Presence of Noise Using a Joint Prior of Static and Dynamic Features

Li Deng, *Senior Member, IEEE*, Jasha Droppo, and Alex Acero, *Fellow, IEEE*

***Abstract*—In this paper, we present a new algorithm for statistical speech feature enhancement in the cepstral domain. The algorithm exploits joint prior distributions (in the form of Gaussian mixture) in the clean speech model, which incorporate both the static and frame-differential dynamic cepstral parameters. Full posterior probabilities for clean speech given the noisy observation are computed using a linearized version of a nonlinear acoustic distortion model, and, based on this linear approximation, the conditional minimum mean square error (MMSE) estimator for the clean speech feature is derived rigorously using the full posterior. The final form of the derived conditional MMSE estimator is shown to be a weighted sum of three separate terms, and the sum is weighted again by the posterior for each of the mixture component in the speech model. The first of the three terms is shown to arrive naturally from the predictive mechanism embedded in the acoustic distortion model in absence of any prior information. The remaining two terms result from the speech model using only the static prior and only the dynamic prior, respectively. Comprehensive experiments are carried out using the Aurora2 database to evaluate the new algorithm. The results demonstrate significant improvement in noise-robust recognition accuracy by incorporating the joint prior for both static and dynamic parameter distributions in the speech model, compared with using only the static or dynamic prior and with using no prior.**

***Index Terms*—Acoustic distortion model, Bayesian estimation, conditional MMSE, dynamic prior, noise reduction, weighted summation.**

I. Introduction

ONE OF THE major problems that still remains unsolved in the current speech recognition technology is noise robustness (r.f., [27], [32]). Two major classes of techniques for noise robust speech recognition include: 1) the model-domain approach, where the speech models in the recognizer are modified or adapted to match the statistical properties of the unmodified noisy test speech; and 2) the feature-domain approach, where the noisy test speech (possibly the "noisy" training speech as well) is modified or enhanced to move toward clean speech as closely as possible. Our earlier work [2], [8], [9], demonstrated remarkably superior performance of the feature-domain approach over the model-domain one. When the training speech is corrupted intentionally, followed by feature enhancement (i.e., front-end denoising) and retraining of the hidden Markov model (HMM) system, a higher performance is achieved than that under the matched noisy condition which sets the limit of the model-domain approach.

Toward solving the noise robustness problem based on the feature-domain approach, we recently have successfully developed a family of speech feature enhancement algorithms that make use of the availability of stereo training data consisting of simultaneously collected clean and noisy speech under a variety of real-life noisy conditions [8], [9], [13]–[15]. While high performance under severe noise distortion conditions is achievable, it is desirable to remove or reduce the need for the stereo training data, and to overcome the potential problem of unexpected mismatch between the acoustic environments for recognizer deployment and for stereo training. To this end, we have more recently focused on the development of a new, alternative family of statistical and parametric techniques for noise-robust speech recognition. In this paper, we present a new algorithm for statistical speech feature enhancement free from the use of stereo training data. It has been built upon a series of published work on parametric modeling of nonlinear acoustic distortion [1], [2], [11], [19], [24], [25], [30] and represents a significant extension of these earlier work. The main innovations of the current work are: 1) incorporation of the dynamic cepstral features in the Bayesian framework for effective speech feature enhancement; 2) a new enhancement algorithm using the full posterior that elegantly integrates the predictive information from the nonlinear acoustic distortion model, the prior information based on the static clean speech cepstral distribution, and the prior based on the frame-differential dynamic cepstral distribution; and 3) efficient implementation of the new algorithm.

In addition to extending the use of the nonlinear acoustic distortion model, our new enhancement algorithm can also be viewed as a generalization or modification of some major speech enhancement techniques in existence. Conventional spectral subtraction methods [5], [6] work by obtaining a noise estimate in the linear spectral domain and then subtracting that from the noisy observation in the same domain. The subtraction residual gives the spectral estimate for clean speech but there are no mechanisms to reject the subtraction result even if it deviates substantially from the clean speech statistics (due, for instance, to a poor noise estimate). This is because there is no prior model or template for clean speech that can be used for verifying the "reasonableness" of the spectral subtraction result. The new statistical technique presented in this paper provides

Manuscript received November 14, 2001; revised October 22, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hermann Ney.

The authors are with Microsoft Research, Redmond, WA 98052 USA (e-mail: deng@microsoft.com; jdroppo@microsoft.com; alexac@microsoft.com).

Digital Object Identifier 10.1109/TSA.2003.822627

a formal framework to overcome this deficiency, and can be viewed as generalized, probabilistic noise removal. When all the signals are represented in the cepstral or log domain, such generalized "noise removal" that takes into account the prior information is achieved via optimal statistical estimation using a nonlinear environment model. This process has been formalized as an efficient feature enhancement algorithm and will be presented in detail in this paper.

There are numerous speech waveform or feature enhancement techniques in the literature, including our own earlier work, that also heavily rely on the use of prior information for the clean speech statistics [1], [3], [4], [16]–[19], [24], [28]–[30]. The new technique described in this paper either generalizes or differentiates from these cited earlier algorithms in three key aspects. The first aspect concerns the nature of the prior information used for characterizing the clean speech statistics. The prior speech models used for speech enhancement described in [1], [3], [4], [19], [24], and [30] use no dynamic or trajectory properties of speech. Only the spectral shape information derived from each individual frame of speech is exploited. On the other hand, the prior speech models described in [17], [18], [28] make use of very weak dynamic properties of speech via an ergodic HMM. It has been well known that the HMM captures only the global, loosely specified temporal information of speech, and not the strong, locally defined, trajectory property of speech [7], [10], [26]. We believe that the latter, strong dynamic property is more important and desirable as the prior information for speech enhancement for the following reason. When viewed as generalized spectral "subtraction," statistical enhancement techniques all operate by denoising while (optimally) "verifying" the "subtraction residual" via probabilistic matching with the prior clean speech model. If the prior model is equipped with the dynamic-matching mechanism that permits the matching not only at the level of individual (static) frames but simultaneously at the level of a local sequence (dynamic) of frames, then the "verifying" performance—the principal role of the prior model for denoising—will be greatly enhanced. Instead of exploiting complex (locally) dynamic models as speech prior [10], [26], the technique described in this paper capitalizes on the simplest kind of such local dynamic information—frame-differential dynamic parameters—in constructing the speech prior model. Given the success of such dynamic parameters in speech recognition and given the motivation provided above for using the locally dynamic properties of speech, we expect a desirable balance between performance gain in speech feature enhancement and a low degree of algorithm complexity.

The second aspect in which the current work generalizes or differentiates from the earlier statistical techniques for speech enhancement concerns the specific domain in which speech and noise are parameterized. We believe that the cepstral or log domain is most desirable if the purpose of speech enhancement is for robust speech recognition, since this is the domain as close as possible to the back end of the recognizer. The work in [3], [4], [17], [18], [28], [29] developed the enhancement techniques mainly in the linear domain, either in the time-sample or frame-bounded (linear) spectral domain. While the speech feature enhancement techniques reported in [19], [24], and [30] pertain to the same log domain as in the current work, the current work generalizes them in providing both the static and dynamic prior information rather than only the static prior (or virtually no prior). We further note that the early work reported in [16] developed the speech enhancement technique also in the log domain. But the reported technique in [16] makes use of neither static nor dynamic prior information, thereby falling into the non-Bayesian framework, in contrast to the Bayesian framework where our new technique belongs and where the prior speech information has been exploited more heavily than all previously reported techniques.

The third important and unique aspect of the current work is the novel approximation technique developed in dealing with the nonlinearity in the acoustic distortion model. The Vector-Taylor-Series algorithm introduced and evaluated in [24], [25], [30] is a highly simplified and special case of our algorithm, and as will be shown in Section VI, it gave considerably lower performance in robust speech recognition compared with the full implementation of our algorithm. The approximation technique described in this paper is also different from that developed in the recent work reported in [19]. In addition to incorporating a dynamic prior, our new algorithm uses the rigorous full posterior to compute the estimator, subject only to the approximation introduced by truncated Taylor series expansion based on the original work of [24], [25].

The organization of this paper is as follows. In Section II, we establish a statistical model for the acoustic environment which relates the log-spectral vectors of clean speech, noise, and noisy speech in a nonlinear manner. This model provides the mechanism for observation likelihood computation, and serves as the basis for the Bayesian approach to solving the clean speech estimation problem. In Section III, we describe "prior" models for both clean speech and noise, supplying the prior information for estimating clean speech. The prior model for clean speech consists of joint static and frame-differential dynamic cepstral components. The prior information for noise is a deterministic, time varying noise estimate obtained via a sequential algorithm that effectively tracks nonstationarity of the noise. Combining the prior information and a linearized version of the statistical model for approximating the nonlinear acoustic environment, where linearization is carried out via truncated Taylor series, we use Bayes rule to derive the conditional minimum mean squared estimate (MMSE) for the clean speech cepstra. The derivation is presented in Section IV in detail. Section V addresses several key implementation issues, including the choice of the Taylor series expansion point, and the use of an iterative technique aimed to successively improve the Taylor-series approximation accuracy. In Section VI, comprehensive experimental results are reported that demonstrate the effectiveness of the new Bayesian approach, and in particular, of the use of dynamic cepstral features in the prior model.

II. Statistical Model for Acoustic Distortion

We first establish a statistical model for the log-spectral-domain acoustic distortion, which allows the computation of the conditional likelihood for the noisy speech observation, in the same domain, given all relevant information.

Following the standard discrete-time, linear system model for the acoustic distortion in the time and frequency domain [1], [24], we have the relationship between the noisy speech (y, Y), clean speech (x, X), additive noise (n, N), and channel impulse response h with corresponding transfer function H

$$y[t] = x[t] * h[t] + n[t]$$

and

$$Y[k] = X[k]H[k] + N[k] \tag{1}$$

respectively, where $*$ denotes circular convolution, and k is the frequency-bin index in DFT for a fixed-length time window.

Power spectra of the noisy speech can then be obtained from the DFT in (1) by

$$|Y[k]|^2 = |X[k]|^2|H[k]|^2 + |N[k]|^2 + 2|X[k]||H[k]||N[k]|\cos\theta_k$$

where θ_k denotes the (random) angle between the two complex variables $N[k]$ and $(X[k]H[k])$.

We now apply a set of Mel-scale filters (L in total) to power spectra $|Y[k]|^2$ in the frequency domain, where the l^{th} filter is characterized by the transfer function $W_k^{(l)} \geq 0$. This produces L channel (Mel-filter bank) energies of

$$\sum_k W_k^{(l)}|Y[k]|^2 = \sum_k W_k^{(l)}|X[k]|^2|H[k]|^2 + \sum_k W_k^{(l)}|N[k]|^2 + 2\sum_k W_k^{(l)}|X[k]||H[k]||N[k]|\cos\theta_k \tag{2}$$

with $l = 1, 2, \ldots, L$.

Denoting the various channel energies in (2) by

$$|\tilde{Y}^{(l)}|^2 = \sum_k W_k^{(l)}|Y[k]|^2,\ |\tilde{X}^{(l)}|^2 = \sum_k W_k^{(l)}|X[k]|^2,\ |\tilde{N}^{(l)}|^2 = \sum_k W_k^{(l)}|N[k]|^2$$

and

$$|\tilde{H}^{(l)}|^2 = \frac{\sum_k W_k^{(l)}|X[k]|^2|H[k]|^2}{|\tilde{X}^{(l)}|^2}$$

we simplify (2) to

$$|\tilde{Y}^{(l)}|^2 = |\tilde{X}^{(l)}|^2|\tilde{H}^{(l)}|^2 + |\tilde{N}^{(l)}|^2 + 2\lambda^{(l)}|\tilde{X}^{(l)}||\tilde{H}^{(l)}||\tilde{N}^{(l)}| \tag{3}$$

where we define

$$\lambda^{(l)} \equiv \frac{\sum_k W_k^{(l)}(\tilde{X}[k]\tilde{H}[k])\tilde{N}[k]\cos\theta_k}{|\tilde{X}^{(l)}||\tilde{H}^{(l)}||\tilde{N}^{(l)}|}.$$

$\lambda^{(l)}$ is a scalar, which can be shown to have its value between -1 and 1.

Define the log channel energy vectors

$$\mathbf{y} = \begin{bmatrix} \log|\tilde{Y}^{(1)}|^2 \\ \log|\tilde{Y}^{(2)}|^2 \\ \cdots \\ \log|\tilde{Y}^{(l)}|^2 \\ \cdots \\ \log|\tilde{Y}^{(L)}|^2 \end{bmatrix},\ \mathbf{x} = \begin{bmatrix} \log|\tilde{X}^{(1)}|^2 \\ \log|\tilde{X}^{(2)}|^2 \\ \cdots \\ \log|\tilde{X}^{(l)}|^2 \\ \cdots \\ \log|\tilde{X}^{(L)}|^2 \end{bmatrix},$$

$$\mathbf{n} = \begin{bmatrix} \log|\tilde{N}^{(1)}|^2 \\ \log|\tilde{N}^{(2)}|^2 \\ \cdots \\ \log|\tilde{N}^{(l)}|^2 \\ \cdots \\ \log|\tilde{N}^{(L)}|^2 \end{bmatrix},\ \mathbf{h} = \begin{bmatrix} \log|\tilde{H}^{(1)}|^2 \\ \log|\tilde{H}^{(2)}|^2 \\ \cdots \\ \log|\tilde{H}^{(l)}|^2 \\ \cdots \\ \log|\tilde{H}^{(L)}|^2 \end{bmatrix} \tag{4}$$

and define the vector

$$\boldsymbol{\lambda} = \begin{bmatrix} \lambda^{(1)} \\ \lambda^{(2)} \\ \cdots \\ \lambda^{(l)} \\ \cdots \\ \lambda^{(L)} \end{bmatrix}.$$

Equation (3) can now be written as

$$e^{\mathbf{y}} = e^{\mathbf{x}} \bullet e^{\mathbf{h}} + e^{\mathbf{n}} + 2\boldsymbol{\lambda} \bullet e^{\mathbf{x}/2} \bullet e^{\mathbf{h}/2} \bullet e^{\mathbf{n}/2} = e^{\mathbf{x}+\mathbf{h}} + e^{\mathbf{n}} + 2\boldsymbol{\lambda} \bullet e^{(\mathbf{x}+\mathbf{h}+\mathbf{n})/2} \tag{5}$$

where the $\bullet$ operation for two vectors denotes element-wise product, and each exponentiation of a vector above is also an element-wise operation.

To obtain the log channel energy for noisy speech, we apply the log operation on both sides of (5)

$$\begin{aligned} \mathbf{y} &= \log\left[e^{\mathbf{x}+\mathbf{h}} \bullet (\mathbf{1} + e^{\mathbf{n}-\mathbf{x}-\mathbf{h}} + 2\,\boldsymbol{\lambda} \bullet e^{(\mathbf{x}+\mathbf{h}+\mathbf{n})/2-\mathbf{x}-\mathbf{h}})\right] \\ &= \mathbf{x} + \mathbf{h} + \log[\mathbf{1} + e^{\mathbf{n}-\mathbf{x}-\mathbf{h}} + 2\,\boldsymbol{\lambda} \bullet e^{(\mathbf{n}-\mathbf{x}-\mathbf{h})/2}]. \end{aligned} \tag{6}$$

This can be further simplified to

$$\begin{aligned} \mathbf{y} &= \mathbf{x}+\mathbf{h}+\log\left[(\mathbf{1}+e^{\mathbf{n}-\mathbf{h}-\mathbf{x}}) \bullet [\mathbf{1}+2\boldsymbol{\lambda} \bullet e^{(\mathbf{n}-\mathbf{x}-\mathbf{h})/2} \bullet / (\mathbf{1}+e^{\mathbf{n}-\mathbf{x}-\mathbf{h}})]\right] \\ &= \mathbf{x}+\mathbf{h}+\log(\mathbf{1}+e^{\mathbf{n}-\mathbf{h}-\mathbf{x}})+\log[\mathbf{1}+2\boldsymbol{\lambda} \bullet e^{(\mathbf{n}-\mathbf{x}-\mathbf{h})/2} \bullet / (\mathbf{1}+e^{\mathbf{n}-\mathbf{x}-\mathbf{h}})] \\ &= \mathbf{x}+\mathbf{h}+\log(\mathbf{1}+e^{\mathbf{n}-\mathbf{h}-\mathbf{x}})+\log\left[\mathbf{1}+\boldsymbol{\lambda} \bullet / \cos\left(\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2}\right)\right] \\ &\approx \mathbf{x}+\mathbf{h}+\log(\mathbf{1}+e^{\mathbf{n}-\mathbf{h}-\mathbf{x}})+\boldsymbol{\lambda} \bullet / \cos\left(\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2}\right) \end{aligned} \tag{7}$$

where "$\bullet/$" denotes element-wise vector division in the above, and the last step of approximation uses the assumption[1] that $\boldsymbol{\lambda} \ll \cosh((\mathbf{n}-\mathbf{x}-\mathbf{h})/2)$.

[1]Justifications for this assumption are provided here. From $\cosh(z) = (e^z + e^{-z})/2$, it is clear that as z moves away from zero, where the minimum of $\cosh(0) = 1$ lies, the function $\cosh(z)$ rises quickly above zero at an exponential rate. For example, $\cosh(2) = \cosh(-2) = 3.76 > 1$, $\cosh(3) = \cosh(-3) = 10.07 \gg 1$, etc. The corresponding situation where $|(\mathbf{n}-\mathbf{x}-\mathbf{h})/2|$ is between 2 to 3 is common in the data. In the meantime, while $\boldsymbol{\lambda}$ in principle ranges from -1 to 1, we find empirically that it is mostly within the range of -0.3 to $+0.3$.

Due to the generally small values of the last term in (7), this acoustic environment model can be interpreted as a predictive mechanism for $\mathbf{y}$ where the predictor is

$$\hat{\mathbf{y}} = \mathbf{x} + \mathbf{h} + \mathbf{g}(\mathbf{n} - \mathbf{x} - \mathbf{h})$$

in which

$$\mathbf{g}(\mathbf{z}) = \log(\mathbf{1} + e^{\mathbf{z}}).$$

The small prediction residual in (7):

$$\mathbf{r} = \boldsymbol{\lambda} \bullet / \cosh\left(\frac{\mathbf{n} - \mathbf{x} - \mathbf{h}}{2}\right) \tag{8}$$

is complicated to evaluate and to model. It is therefore represented by an "ignorance" model as a Gaussian random vector. Using the stereo speech data consisting of about 10 000 digit sequences in the training set of the Aurora2 database and using (7), we have empirically verified that the average value of $\boldsymbol{\lambda} \bullet / \cosh((\mathbf{n} - \mathbf{x} - \mathbf{h})/2)$ is very close to zero for each vector element. We also observed empirically that the distribution of $\boldsymbol{\lambda}$ has Gaussian shapes (subject to the truncation above $+1$ and below -1). Therefore, as a reasonable choice, the zero mean vector is fixed in the Gaussian distribution as an approximate model for the prediction residual.

The covariance matrix of the modeled Gaussian random vector for the prediction residual (8) is clearly a function of the (instantaneous) SNR. Empirical analysis verifies this SNR dependency, and further shows that the covariance matrices for all SNR levels are strongly diagonally dominant. Diagonality of the covariance matrix is therefore assumed in the implementation.

The work presented in this paper avoids the implementation complexity associated with the SNR dependency. Instead, as an approximation, we estimate one fixed diagonal covariance matrix, by pooling all available SNR's including clean speech, of the residual noise using the Aurora2's full multi-condition training set. Since the true noise is available in the Aurora2 database, errors in the model can be computed precisely for each frame in the training set. And the sample covariance matrix is computed as its estimate. Assuming a fixed (diagonal) covariance matrix $\boldsymbol{\Psi}$, the statistical model for the acoustic environment is thus established as

$$\mathbf{y} = \underbrace{\mathbf{x} + \mathbf{h} + \mathbf{g}(\mathbf{n} - \mathbf{x} - \mathbf{h})}_{\hat{\mathbf{y}}} + \mathbf{r} \tag{9}$$

with $\mathbf{r} \sim \mathcal{N}(\mathbf{r}; \mathbf{0}, \boldsymbol{\Psi})$.

Another simplification in the implementation work described in this paper is to take account of additive noise only. The channel distortion is handled via a separate process of cepstral mean normalization. This further simplifies the model of (9) into

$$\mathbf{y} = \underbrace{\mathbf{x} + \mathbf{g}(\mathbf{n} - \mathbf{x})}_{\hat{\mathbf{y}}} + \mathbf{r}. \tag{10}$$

The Gaussian assumption for the residual $\mathbf{r}$ in model of (10) allows straightforward computation of the likelihood for the noisy speech observation according to

$$p(\mathbf{y}|\mathbf{x}, \mathbf{n}) = \mathcal{N}(\mathbf{y}; \mathbf{x} + \mathbf{g}(\mathbf{n} - \mathbf{x}), \boldsymbol{\Psi}) = \mathcal{N}(\mathbf{y}; \hat{\mathbf{y}}, \boldsymbol{\Psi}). \tag{11}$$

This likelihood model is one key component in the Bayesian framework for speech enhancement, which will be described in Section IV.

III. Prior Models

In addition to the acoustic environment model for the likelihood evaluation of $\mathbf{y}$ as derived in the preceding section, the Bayesian estimation framework which we adopt here also requires "prior" models for the statistical behavior of clean speech features and of noise features. Both the speech and noise features are known to be nonstationary. The mechanisms we have designed to capture the nonstationarity in the prior model for clean speech are: 1) using dynamic features which take the local, time difference of static features; and 2) using multiple modes (mixtures) in the probability distribution, allowing the mode to switch freely at different times in an unstructured manner. In contrast, the mechanism designed to embrace the nonstationarity in the prior model for noise is to directly represent the prior properties of the noise features in an explicitly time-indexed fashion. Both of these design considerations make use of the well established results of our earlier research [8], [11], thereby facilitating the implementation of the new speech enhancement algorithm considerably.

A. Model for Clean Speech Incorporating Dynamic Features

The prior model exploited in this work takes into account both the static and dynamic properties of clean speech, in the domain of log Mel-channel energy (or equivalently in the domain of cepstrum via a fixed, linear transformation). One simple way of capturing the dynamic property is to use the frame-differential, or "delta" feature, defined by

$$\Delta\mathbf{x}_t \equiv \mathbf{x}_t - \mathbf{x}_{t-1}$$

where a one-step, backward time (frame) difference is used in this work.

The functional form of the probability distribution for both the static and delta features of clean speech is chosen, motivated by simplicity in the algorithm implementation, as a mixture of multivariate Gaussians, where in each Gaussian component the static and delta features are assumed to be uncorrelated with each other. This gives the joint PDF:

$$p(\mathbf{x}_t, \Delta\mathbf{x}_t) = \sum_{m=1}^{M} c_m \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_m^x, \boldsymbol{\Sigma}_m^x) \mathcal{N}(\Delta\mathbf{x}_t; \boldsymbol{\mu}_m^{\Delta\mathbf{x}}, \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}). \tag{12}$$

In our speech feature enhancement system implementation, a standard EM algorithm is used to train the mean and covariance parameters $\boldsymbol{\mu}_m^x$, $\boldsymbol{\Sigma}_m^x$, $\boldsymbol{\mu}_m^{\Delta\mathbf{x}}$, and $\boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}$ in the cepstral domain. Then the mean vectors in the log Mel-channel energy domain are obtained via the linear transform using the inverse cosine transformation matrix. The two covariance matrices in the log Mel-channel energy domain are computed also from those in the

cepstral domain, using the inverse cosine transformation matrix and its transpose. After this training and the transformations, we now assume that all parameters in (12) are known in the log Mel-channel energy domain.

Note that due to the inclusion of the delta feature in (12), the speech frame $\mathbf{x}_t$ is no longer independent of its previous frame. This allows the trajectory information of speech to be captured as part of the prior information. Compared with the conventional approaches which exploit only the static features in the Gaussian mixture model, the additional information source provided by (12) is the new dynamic parameters $\boldsymbol{\mu}_m^{\Delta\mathbf{x}}$ and $\boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}$ which cannot be inferred from the static parameters $\boldsymbol{\mu}_m^{\mathbf{x}}$ and $\boldsymbol{\Sigma}_m^{\mathbf{x}}$. These orthogonal sources of information permit more accurate characterization of the prior statistical properties of the clean-speech sequence.

B. Model for Noise Features Using Time-Varying, Fixed-Point Estimate

In principle, in the Bayesian framework adopted in this work, it is also desirable to provide a prior distribution for the noise parameter $\mathbf{n}$. Due to the fast changing nature of the noise in the database (Aurora2) which we evaluate our algorithm on, the noise distribution would need to be nonstationary or time-varying; that is, the noise distribution be a function of time frame t. Given only a limited amount of noisy speech training data available, even assuming a simple Gaussian model for the noise feature with a time-varying mean and variance, accurate estimation of these parameters is still very difficult. We in this work use the results from our earlier research where the noise feature is assumed to be deterministic and is tracked sequentially directly from the individual noisy test utterance.[2] This is equivalent to assuming a nonstationary (degenerated) Gaussian model as the prior for noise, where the mean vector indexed separately for each time frame t is known and where the covariance matrix is fixed to be zero. That is, the prior probability distribution for noise is reduced to a time-varying, vector-valued, delta function

$$p(\mathbf{n}_t) = \delta(\mathbf{n}_t - \bar{\mathbf{n}}_t). \tag{13}$$

The above method of dealing with noise nonstationarity can be considered as a "nonparametric" technique, where the noise variable $\bar{\mathbf{n}}_t$ is explicitly indexed by each time t, rather than being drawn from a parametric distribution. This is in contrast to the use of the time-invariant mixture model as a parametric method for capturing speech nonstationarity as described earlier in this section. The speech nonstationarity is implicitly embedded via the possibility of mode (component) switching in the mixture model. This parametric technique is appropriate for modeling clean speech, since the training set used to estimate the parameters in the mixture model can often easily cover the acoustic space of the "hidden" clean speech responsible for generating the noisy speech observations.[3] However, this same parametric technique would not be appropriate for modeling noise nonstationarity. This is because the noise types and levels are too numerous, and they are too difficult to predict in advance for training a mixture model with a full coverage of the acoustic space for the time-varying noise embedded in the test data. Adaptation to the test data is necessary, and the very small amount of adaptation data in each changing test utterance makes parametric techniques ineffective. The nonparametric adaptive tracking technique we developed in [11] produces explicitly time-varying parameters $\bar{\mathbf{n}}_t$ in (13) that are used in the current work. Here we briefly explain the computation of this noise tracking algorithm in principle. It uses the iterative stochastic approximation to improve piecewise linear approximation to a nonlinear acoustic distortion model, and it uses a recursive EM algorithm to compute the (locally) optimal on-line estimate of the noise at each frame taking into account the exponentially decaying effects of the past history. In the algorithm implementation, three iterations are used for each new noisy speech frame. In each iteration, the posterior probability for each mixture component of the Gaussian-mixture clean speech model is computed first. Then the first and second-order derivatives of the E-step objective function are computed using this posterior in an efficient, recursive manner. Finally, the noise estimate is updated using both orders of derivatives.

IV. Bayesian Approach to Speech Feature Enhancement

Given the prior models for clean speech (12) and for noise (13), and given the likelihood model (11), an application of Bayes rule in principle would give the posterior probability for the clean speech conditioned on the noisy speech observations. This computation, however, is highly complex, since it would require very expensive nonlinear techniques. The computation is made feasible in this work, as described in this section, in two ways. First, linearization on the nonlinear predictor, $\hat{\mathbf{y}}$, in the likelihood model (11) is made. The approximation accuracy is improved via an iterative technique in nonlinear signal processing [23], which was previously successfully applied to speech enhancement in [19] and in spontaneous speech recognition in [12]. Second, while computing the entire posterior probability would be desirable for an integrated system for signal processing and speech recognition, speech feature enhancement as front-end signal processing of primary concern to this work does not require the complete posterior probability. In this section, we describe the estimator used in this work that can be computed via Bayes rule from the likelihood model (11) and prior models (12) and (13). We then derive the estimation formulas with the prior speech model for static features only and for joint static/dynamic features, respectively, using linear approximation to the nonlinear predictor, $\hat{\mathbf{y}}$, in the likelihood model (11).

A. Minimum Mean Square Error (MMSE) Estimator

Given the observation vector $\mathbf{y}$, the minimum mean square error (MMSE) estimator $\hat{\mathbf{x}}$ for the random vector $\mathbf{x}$ is one that minimizes the MSE distortion measure of

$$MSE \equiv E[(\mathbf{x} - \hat{\mathbf{x}})^T(\mathbf{x} - \hat{\mathbf{x}})]$$

[2]Details of this sequential estimation algorithm for highly nonstationary noise can be found in [11], which has been based on the theoretical framework published in [31].

[3]This statement would not be correct if Lombard effect were prominent in the production of distorted speech. Our evaluation data from Aurora2 database do not fall into this situation.

or

$$\hat{\mathbf{x}} = \arg\min_{\hat{\mathbf{x}}} MSE = \arg\min_{\hat{\mathbf{x}}} E[(\mathbf{x}-\hat{\mathbf{x}})^T(\mathbf{x}-\hat{\mathbf{x}})].$$

From the fundamental theorem of estimation theory (cf., [23, pp. 175–176]), the MMSE estimator is shown to be the following conditional expectation, which is the expected value of the posterior probability $p(\mathbf{x}|\mathbf{y})$:

$$\hat{\mathbf{x}} = E[\mathbf{x}|\mathbf{y}] = \int \mathbf{x}p(\mathbf{x}|\mathbf{y})d\mathbf{x}. \tag{14}$$

This becomes

$$\hat{\mathbf{x}} = \frac{\int \mathbf{x}p(\mathbf{y}|\mathbf{x})p(\mathbf{x})d\mathbf{x}}{p(\mathbf{y})} \tag{15}$$

after using Bayes rule.

While the MMSE estimator is generally more difficult to derive than some other estimator for random parameters such as the maximum a posteriori (MAP) estimator for speech waveforms or features, we choose the MMSE estimator in this work for two reasons. First, in much of the past work on speech enhancement using HMMs, the MMSE estimator has in practice exhibited consistently superior enhancement performance over the (approximate) MAP estimator for speech waveforms or features (cf. [17] and [28]). The second reason is a theoretical one. Although the MMSE estimator is defined for the MSE distortion measure, its optimality also extends over to a large class of other distortion measures (under only some mild conditions). This property does not hold for the MAP estimator. Because the perceptually significant distortion measure for speech is unknown, the wide coverage of the distortion classes by the MMSE estimator with the same optimality is highly desirable.

B. Estimation With Prior Speech Model for Static Features Only

To facilitate the derivation of the MMSE estimator with the prior speech model for joint static and dynamic features, we in this subsection first derive the estimator with prior speech model for static features only. The result will be extended to the desired case in the next subsection.

In this derivation, the prior model for clean speech is a simplified version of model (12), and frame index t is dropped since the model is independent of t

$$p(\mathbf{x}) = \sum_{m=1}^{M} c_m \underbrace{\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_m^x,\boldsymbol{\Sigma}_m^x)}_{p(\mathbf{x}|m)}. \tag{16}$$

The derivation starts from (15), from which we use

$$p(\mathbf{y}|\mathbf{x}) = \int p(\mathbf{n})p(\mathbf{y}|\mathbf{x},\mathbf{n})d\mathbf{n} \text{ and } p(\mathbf{x}) = \sum_{m=1}^{M} c_m p(\mathbf{x}|m)$$

to obtain

$$\hat{\mathbf{x}} = \frac{\sum_{m=1}^{M} c_m \int\int \mathbf{x}p(\mathbf{n})p(\mathbf{x}|m)p(\mathbf{y}|\mathbf{x},\mathbf{n})d\mathbf{x}d\mathbf{n}}{p(\mathbf{y})}. \tag{17}$$

Using the deterministic prior noise model (13), (17) is simplified to

$$\hat{\mathbf{x}} = \frac{\sum_{m=1}^{M} c_m \int \mathbf{x}p(\mathbf{x}|m)p(\mathbf{y}|\mathbf{x},\bar{\mathbf{n}})d\mathbf{x}}{p(\mathbf{y})}. \tag{18}$$

Using the likelihood model (11), we now evaluate the integral in (18) as

$$\begin{aligned}\mathbf{I}_m &= \int \mathbf{x}p(\mathbf{x}|m)p(\mathbf{y}|\mathbf{x},\bar{\mathbf{n}})d\mathbf{x} \\ &= \int \mathbf{x}\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_m^x,\boldsymbol{\Sigma}_m^x)\mathcal{N}(\mathbf{y};\mathbf{x}+\mathbf{g}(\bar{\mathbf{n}}-\mathbf{x}),\boldsymbol{\Psi})d\mathbf{x}\end{aligned} \tag{19}$$

where $\mathbf{y}$ and $\bar{\mathbf{n}}$ are treated as constants. This integral, unfortunately, does not have a closed-form result due to the nonlinear function of $\mathbf{x}$ in $\mathbf{g}(\bar{\mathbf{n}}-\mathbf{x})$. To overcome this, we linearize the nonlinearity using truncated Taylor series. The first-order Taylor series has the form of

$$\mathbf{y} \approx \mathbf{x} + \mathbf{g}(\bar{\mathbf{n}}-\mathbf{x}_0) + \mathbf{g}'(\mathbf{x}_0)(\mathbf{x}-\mathbf{x}_0) + \mathbf{r}$$

where $\mathbf{x}_0$ is the fixed expansion point,[4] and $\mathbf{g}'(\mathbf{x}_0)$ is the gradient of function $\mathbf{g}(\bullet)$ evaluated at $\mathbf{x}_0$. The zero-th order Taylor series expansion on $\mathbf{g}(\bullet)$ has a much simpler form

$$\mathbf{y} \approx \mathbf{x} + \mathbf{g}(\bar{\mathbf{n}}-\mathbf{x}_0) + \mathbf{r}.$$

This approximation simplifies the likelihood model (11) to

$$p(\mathbf{y}|\mathbf{x},\mathbf{n}) = \mathcal{N}(\mathbf{y};\underbrace{\mathbf{x}+\mathbf{g}(\bar{\mathbf{n}}-\mathbf{x}_0)}_{\hat{\mathbf{y}}},\boldsymbol{\Psi}) \tag{20}$$

which will be used in the remaining derivation in this paper.

Now, the integral of (19) becomes

$$\begin{aligned}\mathbf{I}_m &= \int \mathbf{x}\mathcal{N}(\mathbf{x};\boldsymbol{\mu}_m^x,\boldsymbol{\Sigma}_m^x)\mathcal{N}(\mathbf{y};\mathbf{x}+\mathbf{g}_0,\boldsymbol{\Psi})d\mathbf{x} \\ &\propto \int \mathbf{x}e^{-1/2[(\mathbf{x}-\boldsymbol{\mu}_m^x)^T(\boldsymbol{\Sigma}_m^x)^{-1}(\mathbf{x}-\boldsymbol{\mu}_m^x)+(\mathbf{y}-\mathbf{x}-\mathbf{g}_0)^T\boldsymbol{\Psi}^{-1}(\mathbf{y}-\mathbf{x}-\mathbf{g}_0)]}d\mathbf{x}.\end{aligned} \tag{21}$$

where $\mathbf{g}_0 = \mathbf{g}(\bar{\mathbf{n}}-\mathbf{x}_0)$, which can be treated as a constant now. After fitting the exponent in (21) into a standard quadratic form in $\mathbf{x}$, and using $\int \mathbf{x}\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{x} = \boldsymbol{\mu}$, a closed-form result is obtained as

$$\begin{aligned}\mathbf{I}_m &= (\boldsymbol{\Sigma}_m^x+\boldsymbol{\Psi})^{-1}[\boldsymbol{\Psi}\boldsymbol{\mu}_m^x+\boldsymbol{\Sigma}_m^x(\mathbf{y}-\mathbf{g}_0)]\mathcal{N}(\mathbf{y};\boldsymbol{\mu}_m^x+\mathbf{g}_0,\boldsymbol{\Sigma}_m^x+\boldsymbol{\Psi}) \\ &= [\mathbf{W}_1(m)\boldsymbol{\mu}_m^x+\mathbf{W}_2(m)(\mathbf{y}-\mathbf{g}_0)]N_m(\mathbf{y})\end{aligned} \tag{22}$$

where we introduced the weighting matrices $\mathbf{W}_1(m) = (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}$ and $\mathbf{W}_2(m) = \mathbf{I} - \mathbf{W}_1(m)$, and where

$$N_m(\mathbf{y}) = \mathcal{N}(\mathbf{y};\boldsymbol{\mu}_m^x + \mathbf{g}_0, \boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})$$

can be easily shown to be the likelihood of observation $\mathbf{y}$ given the m-th component in the clean speech model and under the zero-th order approximation made in (20). That is,

$$p(\mathbf{y}|m) \approx N_m(\mathbf{y}).$$

[4]Selection of this expansion point is crucial for the success of speech recognition applications. This issue will be discussed in Section V.

The denominator in (18)

$$p(\mathbf{y}) = \sum_{m=1}^{M} c_m \int p(\mathbf{x}|m)p(\mathbf{y}|\mathbf{x}, \bar{\mathbf{n}})d\mathbf{x}$$

can be easily computed in closed form also under the zero-th order approximation of (20)

$$p(\mathbf{y}) = \sum_{m=1}^{M} c_m p(\mathbf{y}|m) = \sum_{m=1}^{M} c_m N_m(\mathbf{y}). \tag{23}$$

Now, substituting (22) and (23) into (18), we obtain the final closed-form MMSE estimator

$$\begin{aligned}\hat{\mathbf{x}} &= \frac{\sum_{m=1}^{M} c_m N_m(\mathbf{y})[\mathbf{W}_1(m)\boldsymbol{\mu}_m^x + \mathbf{W}_2(m)(\mathbf{y} - \mathbf{g}_0)]}{\sum_{m=1}^{M} c_m N_m(\mathbf{y})} \\ &= \sum_{m=1}^{M} \gamma_m[\mathbf{W}_1(m)\boldsymbol{\mu}_m^x + \mathbf{W}_2(m)(\mathbf{y} - \mathbf{g}_0)] \end{aligned} \tag{24}$$

where

$$\gamma_m(\mathbf{y}) = \frac{c_m N_m(\mathbf{y})}{\sum_{m=1}^{M} c_m N_m(\mathbf{y})}$$

is the posterior probability $p(m|\mathbf{y})$ for the mixture component.

The MMSE estimator for clean speech in (24) has a clear interpretation. The component in the first term, $\boldsymbol{\mu}_m^x$, is the prior mean vector in the clean speech model. The component in the second term

$$\mathbf{y} - \mathbf{g}_0 \approx (\mathbf{x} + \mathbf{g}_0 + \mathbf{r}) - \mathbf{g}_0 = \mathbf{x} + \mathbf{r}$$

is the true clean speech vector perturbed by a small zero-mean residual, and can be interpreted as the prediction of clean speech when no prior information on speech statistics is available. After the prior information is being made available, each summand in the estimator of (24) is a weighted sum of these two terms (for each mixture component), where the weights are determined by the relative sizes of the variances, $\bar{\boldsymbol{\Psi}}$ and $\bar{\boldsymbol{\Sigma}}_m^x$, in the likelihood prediction model and in the prior speech model, respectively. The final MMSE estimator is another, outer-loop, weighted sum of this combined prediction with each weight being the posterior probability for each mixture component.

C. Estimation With Prior Speech Model for Static and Dynamic Features

We now derive the (conditional) MMSE estimator using a more complex prior speech model (12) with dynamic features, instead of model (16) with static features only.

Given the estimated clean speech feature in the immediately past frame, $\hat{\mathbf{x}}_{t-1}$, the conditional MMSE estimator for the current frame t becomes

$$\hat{\mathbf{x}}_{t|t-1} \equiv E[\mathbf{x}|\mathbf{y}, \hat{\mathbf{x}}_{t-1}].$$

Following a similar derivation for (18), its counterpart result is

$$\begin{aligned}\hat{\mathbf{x}}_{t|t-1} &= \frac{\sum_{m=1}^{M} c_m \int \mathbf{x}_t p(\mathbf{x}_t|m, \hat{\mathbf{x}}_{t-1})p(\mathbf{y}_t|\mathbf{x}_t, \bar{\mathbf{n}}_t)d\mathbf{x}_t}{p(\mathbf{y}_t)} \\ &\approx \frac{\sum_{m=1}^{M} c_m \int \mathbf{x}_t p(\mathbf{x}_t|m, \mathbf{x}_{t-1})p(\mathbf{y}_t|\mathbf{x}_t, \bar{\mathbf{n}}_t)d\mathbf{x}_t}{p(\mathbf{y}_t)} \end{aligned} \tag{25}$$

where we used the approximation

$$\begin{aligned} p(\mathbf{x}_t|m, \hat{\mathbf{x}}_{t-1}) &= \int p(\mathbf{x}_t, \mathbf{x}_{t-1}|m, \hat{\mathbf{x}}_{t-1})d\mathbf{x}_{t-1} \\ &= \int p(\mathbf{x}_t|\mathbf{x}_{t-1}, m, \hat{\mathbf{x}}_{t-1})p(\mathbf{x}_{t-1}|\hat{\mathbf{x}}_{t-1})d\mathbf{x}_{t-1} \\ &= \int p(\mathbf{x}_t|\mathbf{x}_{t-1}, m)p(\mathbf{x}_{t-1}|\hat{\mathbf{x}}_{t-1})d\mathbf{x}_{t-1} \\ &= \int p(\mathbf{x}_t|\mathbf{x}_{t-1}, m)\mathcal{N}(\mathbf{x}_{t-1}; \hat{\mathbf{x}}_{t-1}, \boldsymbol{\Sigma}^{\hat{\mathbf{x}}_{t-1}})d\mathbf{x}_{t-1} \\ &\approx p(\mathbf{x}_t|m, \mathbf{x}_{t-1}). \end{aligned} \tag{26}$$

This approximation dramatically simplifies the MMSE estimator, which would otherwise require dynamic programming or a solution of the inverse of a large tridiagonal matrix. Either would incur a much larger computational cost than the approximate solution presented in this section. The approximation above can be justified if we assume a zero variance, i.e., $\boldsymbol{\Sigma}^{\hat{\mathbf{x}}_{t-1}} = \mathbf{0}$, in the presumed Gaussian for $p(\mathbf{x}_{t-1}|\hat{\mathbf{x}}_{t-1})$. That is, the MMSE estimator $\hat{\mathbf{x}}_{t-1}$ is assumed to have incurred no error: $p(\mathbf{x}_{t-1}|\hat{\mathbf{x}}_{t-1}) = \delta(\mathbf{x}_{t-1} - \hat{\mathbf{x}}_{t-1})$. This assumption is less erroneous when the history in the past is short than when this history is longer, and hence we choose $t-1$ instead of the frame further back as in most speech recognition systems.

To compute the integral in the above (25), we first evaluate the conditional prior of

$$\begin{aligned} &p(\mathbf{x}_t|m, \mathbf{x}_{t-1}) \\ &\propto p(\mathbf{x}_t, \mathbf{x}_t - \mathbf{x}_{t-1}|m) = \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_m^x, \boldsymbol{\Sigma}_m^x)\mathcal{N}(\Delta\mathbf{x}_t; \boldsymbol{\mu}_m^{\Delta\mathbf{x}}, \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}) \\ &\propto e^{-1/2[(\mathbf{x}_t - \boldsymbol{\mu}_m^x)^T(\boldsymbol{\Sigma}_m^x)^{-1}(\mathbf{x}_t - \boldsymbol{\mu}_m^x) + (\mathbf{x}_t - \mathbf{x}_{t-1} - \boldsymbol{\mu}_m^{\Delta\mathbf{x}})^T(\boldsymbol{\Sigma}_m^{\Delta\mathbf{x}})^{-1}(\mathbf{x}_t - \mathbf{x}_{t-1} - \boldsymbol{\mu}_m^{\Delta\mathbf{x}})}. \end{aligned} \tag{27}$$

Fitting the exponent in (27) into the standard quadratic form in $\mathbf{x}_t$, we have

$$p(\mathbf{x}_t|m, \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{28}$$

where

$$\boldsymbol{\mu}_m = (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}})^{-1}\boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}\boldsymbol{\mu}_m^x + (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}})^{-1}\boldsymbol{\Sigma}_m^x(\mathbf{x}_{t-1} + \boldsymbol{\mu}_m^{\Delta\mathbf{x}}) \tag{29}$$

and

$$\boldsymbol{\Sigma}_m = (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}})^{-1}\boldsymbol{\Sigma}_m^x\boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}. \tag{30}$$

Using the same zero-th order approximation, (20), to the nonlinear function in the likelihood model, and substituting

(28)–(30) into (25), we obtain the final result for the conditional MMSE estimator

$$\hat{\mathbf{x}}_{t|t-1} \approx \sum_{m=1}^{M} \gamma_m [\mathbf{V}_1(m)\boldsymbol{\mu}_m^x + \mathbf{V}_2(m)(\hat{\mathbf{x}}_{t-1} + \boldsymbol{\mu}_m^{\Delta\mathbf{x}}) + \mathbf{V}_3(m)(\mathbf{y} - \mathbf{g}(\bar{\mathbf{n}}_t - \mathbf{x}_0))] \tag{31}$$

where

$$\mathbf{V}_1(m) = \underbrace{(\boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}}_{\mathbf{W}_1}(\boldsymbol{\Sigma}_m^x + \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}})^{-1}\boldsymbol{\Sigma}_m^{\Delta\mathbf{x}},$$
$$\mathbf{V}_2(m) = \underbrace{(\boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})^{-1}\boldsymbol{\Psi}}_{\mathbf{W}_1}(\boldsymbol{\Sigma}_m^x + \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}})^{-1}\boldsymbol{\Sigma}_m^x$$

and

$$\mathbf{V}_3(m) = \mathbf{W}_2(m) = (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})^{-1}\boldsymbol{\Sigma}_m^x \tag{32}$$

and where $\mathbf{x}_{t-1} \approx \hat{\mathbf{x}}_{t-1}$ is used.

Note that

$$\mathbf{V}_1(m) + \mathbf{V}_2(m) + \mathbf{V}_3(m) = \mathbf{I}. \quad \forall\, m.$$

This thus also provides a clear interpretation of (31), generalizing from (24) given earlier. That is, each summand in (31), as a mixture-component (m) specific contribution to the final estimator, is a weighted sum of three terms. The unweighted first two terms are derived from the static and dynamic elements in the prior clean speech model, respectively. The unweighted third term is derived from the predictive mechanism based on the linearized acoustic distortion model in absence of any prior information.

Note also that under the limiting case where $\boldsymbol{\Sigma}_m^{\Delta\mathbf{x}} \to \infty$, we have

$$\mathbf{V}_1(m) \to \mathbf{W}_1(m), \text{ and } \mathbf{V}_2(m) \to \mathbf{0}.$$

Then the conditional MMSE estimator (31) reverts to the MMSE estimator (24) when no prior for dynamic speech features is exploited. This shows a desirable property of (31) since when $\boldsymbol{\Sigma}_m^{\Delta\mathbf{x}} \to \infty$ the effect of using the prior for dynamic features should indeed be diminishing to null.

As the opposite limiting case, let $\boldsymbol{\Sigma}_m^{\Delta\mathbf{x}} \to \mathbf{0}$. We then have

$$\mathbf{V}_1(m) \to \mathbf{0}, \text{ and } \mathbf{V}_2(m) \to \mathbf{W}_1(m).$$

That is, only the prior information for the dynamic speech features is used for speech feature enhancement.

V. Key Issues in Algorithm Implementation

In this section, we provide some key implementation details for the speech feature enhancement algorithm (31) derived in the preceding section, and then give the algorithm execution steps.

A. Initialization and Iterative Refinement of Taylor Series Expansion Point

While deriving the conditional MMSE estimator (31), as well as its limiting case (24), we left untouched the key issue of how to choose the Taylor series expansion point, $\mathbf{x}_0$, in approximating the likelihood model in (20). This crucial issue for the algorithm implementation is resolved in two ways. First, the following crude but reasonable estimate for the "clean" speech is used to initialize the Taylor series expansion point:

$$\mathbf{x}_0 = \arg\max_{\boldsymbol{\mu}_m^x} p(\mathbf{y}|m) \approx \arg\max_{\boldsymbol{\mu}_m^x} \mathcal{N}[\mathbf{y}; \boldsymbol{\mu}_m^x + \mathbf{g}(\bar{\mathbf{n}} - \boldsymbol{\mu}_m^x), \boldsymbol{\Psi}].$$

Second, this initial estimate is refined successively via iterations using the conditional MMSE estimator (31). This turns the algorithm (31) into an iterative one, which will be formalized shortly.

The motivation for the use of iterations is our simple recognition that the accuracy of the truncated Taylor series approximation to a nonlinear function is determined largely by the accuracy of the expansion point to the true variable value of the function's argument (given the fixed expansion order), and that (31) is simply the "best" available estimate of that true variable value. Therefore, the successive refinement on this estimate should improve the Taylor series approximation accuracy and hence the new estimator's quality. The use of iterations is also motivated by its success in the work of [19], where a large number of Taylor series expansion points are used. This contrasts the one single-point expansion used in this work that considerably cuts down the computational load in our algorithm.

The convergence property of our iterative algorithm has not been systematically explored. However, under the special case where $\mathbf{W}_1 = \mathbf{0}$ (using no prior information), (31) becomes the well known fixed-point iterative solution of solving for $\mathbf{x}$ in the nonlinear equation

$$\mathbf{x} = \mathbf{y} - \mathbf{g}(\bar{\mathbf{n}} - \mathbf{x}) = \mathbf{y} - \log\left(\mathbf{1} + e^{\bar{\mathbf{n}} - \mathbf{x}}\right) = \mathbf{h}(\mathbf{x}). \tag{33}$$

The convergence property for this special case of the algorithm can be found in standard numerical analysis textbooks (e.g., [22]). Since the gradient of the right hand side of (33) is always less than one

$$\nabla\mathbf{h}(\mathbf{x}) = \mathbf{1} - \frac{\mathbf{1}}{\mathbf{1} + e^{\bar{\mathbf{n}} - \mathbf{x}}} < \mathbf{1}$$

the iterative solution to (33) is guaranteed to converge, as has been observed in our experimental work also.

B. Variance Scaling

Another important issue we explored in the algorithm implementation is the variance weighting aimed to balance the contributions of the static and dynamic feature priors to the overall qualities of denoising and of speech recognition performance. We found that the use of the variances in the clean speech model estimated from the training data set alone does not give optimal performance (see details in Section VI-C). This suggests that the information provided by the static cepstral means and that by the differential cepstral means in the clean speech model do not consistently complement each other in enhancing the accuracy of the model, based on the simple parametric form of PDF given by (12), in representing the true, underlying dynamics of clean speech features.[5]

[5]This has been a well known problem in statistical modeling of speech-feature dynamics. A search for solutions to this problem has produced a number of advanced statistical models beyond the conventional HMM [7], [26], [10].

An additional reason why the use of the variances in the clean speech model estimated from the training data set is not most desirable may be attributed to the approximation made in (26): $\mathbf{\Sigma}^{\hat{\mathbf{x}}} = \mathbf{0}$. Due to the necessarily imperfect estimator, this variance can not be zero. Not accounting for such a variance is a source of the above problem.

Rather than providing a rigorous, expensive solution, we adopt a simple yet effective way of problem fixing which we describe below. That is, the contributions of the static and dynamic feature priors are adjusted by empirically scaling[6] the variance $\mathbf{\Sigma}_m^{\Delta\mathbf{x}}$ (including all diagonal elements) in (31). This keeps the estimation algorithm intact while slightly changing the weights in (31) to

$$\mathbf{V}_1(m,\rho)=(\mathbf{\Sigma}_m^x+\mathbf{\Psi})^{-1}\mathbf{\Psi}(\mathbf{\Sigma}_m^x+\rho\mathbf{\Sigma}_m^{\Delta\mathbf{x}})^{-1}(\rho\mathbf{\Sigma}_m^{\Delta\mathbf{x}}) \quad (34)$$

$$\mathbf{V}_2(m,\rho)=(\mathbf{\Sigma}_m^x+\mathbf{\Psi})^{-1}\mathbf{\Psi}(\mathbf{\Sigma}_m^x+\rho\mathbf{\Sigma}_m^{\Delta\mathbf{x}})^{-1}\mathbf{\Sigma}_m^x \quad (35)$$

where ρ is the variance scaling factor. The effects of choosing different ρ will be studied and reported in Section VI-C.

C. Algorithm Description

Summarizing the implementation considerations above, we now describe the complete execution steps for the speech feature algorithm below.

First, train and fix all parameters in the clean speech model: c_m, $\boldsymbol{\mu}_m^x$, $\boldsymbol{\mu}_m^{\Delta\mathbf{x}}$, $\mathbf{\Sigma}_m^x$, and $\mathbf{\Sigma}_m^{\Delta\mathbf{x}}$. Then, compute the noise estimates, $\bar{\mathbf{n}}_t$, for all frames of all test data based on the sequential tracking algorithm described in [11]. Further, precompute the weights $\mathbf{V}_1$, $\mathbf{V}_2$, and $\mathbf{V}_3$, which are dependent on only the known model parameters, according to (34), (35) and (32).

Next, fix the total number, J, of intra-frame iterations. For each frame $t = 2,3,\ldots,T$ in a noisy utterance $\mathbf{y}_t$, set iteration number $j = 1$, and initialize the clean speech estimator by

$$\hat{\mathbf{x}}_t^{(1)} = \arg\max_m \mathcal{N}[\mathbf{y}_t; \boldsymbol{\mu}_m^x + \mathbf{g}(\bar{\mathbf{n}}_t - \boldsymbol{\mu}_m^x), \mathbf{\Psi}] \quad (36)$$

where $\mathbf{\Psi}$ is the error covariance matrix in the acoustic distortion model described in Section II, and is estimated in advance from a set of training data.

Then, execute the following steps sequentially over time frames.

- Step 1: Compute

$$\gamma_t^{(j)}(m) = \frac{c_m\mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_m^x + \mathbf{g}^{(j)}, \mathbf{\Sigma}_m^x + \mathbf{\Psi})}{\sum_{m=1}^{M} c_m\mathcal{N}(\mathbf{y}_t; \boldsymbol{\mu}_m^x + \mathbf{g}^{(j)}, \mathbf{\Sigma}_m^x + \mathbf{\Psi})} \quad (37)$$

  where $\mathbf{g}^{(j)} = \log\left(\mathbf{1} + e^{\bar{\mathbf{n}}_t - \hat{\mathbf{x}}_t^{(j)}}\right)$.
- Step 2: Update the estimator:[7]

$$\hat{\mathbf{x}}_t^{(j+1)} = \sum_m \gamma_t^{(j)}(m)[\mathbf{V}_1(m,\rho)\boldsymbol{\mu}_m^x + \mathbf{V}_2(m,\rho)\boldsymbol{\mu}_m^{\Delta\mathbf{x}}] + \left[\sum_m \gamma_t^{(j)}(m)\mathbf{V}_3(m)\right]\hat{\mathbf{x}}_{t-1}^{(J)}$$

$$+ \left[\sum_m \gamma_t^{(j)}(m)\mathbf{V}_3(m)\right](\mathbf{y} - \mathbf{g}(\bar{\mathbf{n}}_t - \hat{\mathbf{x}}_t^{(j)})). \quad (38)$$

- Step 3: If $j < J$, increment j by one and continue the iteration by returning to Step 1. If $j = J$, then increment t by one and start the algorithm again by re-setting $j = 1$ to process the next time frame until the end of the utterance $t = T$.

[6]This process is similar to the empirical weighting of the language model score adopted by almost any speech recognition system.

[7]We set $\hat{\mathbf{x}}_1^{(J)}$ to be the right-hand side of (36) and update the estimator for $t > 1$.

VI. Speech Recognition Experiments

A. Database and Recognition Task

The iterative algorithm presented thus far for estimating clean speech feature vectors has been evaluated on the Aurora2 database, using the standard recognition tasks designed for this database [20]. The database consists of English connected digits recorded in clean environments. Three sets of digit utterances (sets A, B, and C) are prepared as the test material. These utterances are artificially contaminated by adding noise recorded under a number of conditions and for different noise levels (sets A, B, and C), and also by passing them through different distortion channels (for set C only).

The recognition system used in our evaluation experiments are based on continuous HMM's, and one HMM is trained for each digit under clean condition. Both training and recognition phases are performed using the HTK scripts provided by the Aurora2 database. The speech feature used for the reference experiments to evaluate the new denoising algorithm is the standard MFCC's. The new algorithm is used only as the front-end.

B. Results for Aurora2 Task

Table I summarizes the results for all three sets of the test data in the Aurora2 database. The HMM systems with four different front-ends are compared: 1) use of the iterative algorithm to implement the conditional MMSE estimator, as described in Section V, with the prior speech model consisting of both static and dynamic cepstra and with the optimal variance scaling factor; 2) use of the same estimator except with the prior speech model consisting of only the static cepstra; 3) use of the same estimator except with the prior speech model consisting of only the dynamic cepstra; and 4) use of the slightly modified, Aurora2-supplied standard reference MFCC's with no denoising.[8]

The HMMs used in the four systems are the same. They are trained using the same clean-speech training set supplied in the Aurora2 database. Note that the front-end (2) above is implemented by setting the variance scaling factor ρ in (34) and (35) to be a very large number (5000). And the front-end (3) above is implemented by setting the variance scaling factor ρ to be zero. These extreme values effectively nullify the contributions from the dynamic and static features, respectively, in the prior speech model.

Comparisons in Table I show that the conditional MMSE estimator that fully utilizes both the static and dynamic cepstral distributions [front-end (1)] performs significantly better than the same estimator which utilizes only the partial information [front-ends (2) and (3)]. In producing the results with

[8]The Aurora2-supplied MFCC's use the log-magnitude spectra, and we modified them to the log-magnitude squared spectra.

TABLE I
COMPARISONS OF AURORA2 RECOGNITION RATES (%) FOR THE HMM SYSTEMS USING FOUR DIFFERENT FRONT-ENDS FOR ALL SETS OF TEST DATA: 1) USING MIXTURE-OF-GAUSSIAN CLEAN-SPEECH MODEL UTILIZING BOTH STATIC AND DYNAMIC CEPSTRA; 2) USING MIXTURE-OF-GAUSSIAN CLEAN-SPEECH MODEL UTILIZING ONLY STATIC CEPSTRA; 3) USING MIXTURE-OF-GAUSSIAN CLEAN-SPEECH MODEL UTILIZING ONLY DYNAMIC CEPSTRA; 4) MFCC'S WITH NO DENOISING. THE SAME SINGLE SET OF HMMS USED IN THE FOUR SYSTEMS ARE TRAINED USING THE IDENTICAL CLEAN-SPEECH TRAINING SET SUPPLIED IN THE AURORA2 DATABASE AFTER CEPSTRAL MEAN NORMALIZATION. CEPSTRAL MEAN NORMALIZATION IS APPLIED TO THE TEST SETS ALSO, AFTER SPEECH FEATURE ENHANCEMENT

Front-ends	Set A	Set B	Set C	Overall
(1) Prior: Static-Dynamic Cepstra	86.72	87.03	81.70	**85.84**
(2) Prior: Static Cepstra only	84.74	85.19	78.87	83.74
(3) Prior: Dynamic Cepstra only	79.96	78.91	76.72	78.89
(4) No Denoising (reference)	61.34	55.75	66.14	60.06

TABLE II
DETAILED RECOGNITION RATES (%) USING THE CONDITIONAL MMSE ESTIMATOR FOR CLEAN SPEECH. FOUR NOISE CONDITIONS: SUBWAY, BABBLE, CAR, EXHIBITION-HALL NOISES; SNRS FROM 0 DB TO 20 DB IN 5-DB INCREMENT; SET-A RESULTS CLEAN SPEECH TRAINING

SNR	Subway	Babble	Car	Exhibition	Average
20 dB	97.94	98.46	98.54	98.06	98.25
15 dB	95.46	97.16	97.91	96.24	96.69
10 dB	89.99	92.84	95.71	92.32	92.72
5 dB	81.46	78.57	89.29	82.17	82.87
0 dB	64.75	51.81	70.92	64.76	63.06
Ave.	85.92	83.77	90.47	86.71	86.72

TABLE III
DETAILED RECOGNITION RATES (%) USING THE CONDITIONAL MMSE ESTIMATOR FOR CLEAN SPEECH. FOUR NOISE CONDITIONS: RESTAURANT, STREET, AIRPORT, AND TRAIN-STATION NOISES; SNRS FROM 0 DB TO 20 DB IN 5-DB INCREMENT; SET-B RESULTS WITH CLEAN SPEECH TRAINING

SNR	Restaurant	Street	Airport	Station	Average
20 dB	98.43	97.43	98.54	98.80	98.30
15 dB	96.81	96.07	97.58	97.53	97.00
10 dB	92.78	91.51	94.04	95.09	93.36
5 dB	80.84	81.56	86.01	86.98	83.85
0 dB	56.09	60.22	66.21	68.16	62.67
Ave.	84.99	85.36	88.48	89.31	87.03

TABLE IV
DETAILED RECOGNITION RATES (%) USING THE CONDITIONAL MMSE ESTIMATOR FOR CLEAN SPEECH. FOUR NOISE CONDITIONS: SUBWAY (AS IN SET-A) AND STREET NOISES (AS IN SET-B), AND BOTH ARE MODIFIED BY PASSING THE NOISY SPEECH THROUGH A DIFFERENT DISTORTION CHANNEL; SNRS FROM 0 DB TO 20 DB IN 5-DB INCREMENT; SET-C RESULTS WITH CLEAN SPEECH TRAINING

SNR	Subway-M	Street-M	Average
20 dB	96.99	97.73	97.36
15 dB	94.26	95.16	94.71
10 dB	88.00	88.57	88.29
5 dB	76.36	76.66	76.51
0 dB	54.28	48.97	51.63
Ave.	81.98	81.42	81.70

front-end (1), an optimal value of $\rho = 5.5$ is used (details of the optimization will be discussed later.) They are, however, all significantly and consistently better than the standard MFCC's supplied by the AURORA task using no robust preprocessing to enhance speech features [front-end (4)]. The relative word error rate reduction using front-end (1) is 64.54% compared with the results with standard MFCCs using no enhancement. These results are statistically significant, based on a total of $1001 \times 10 \times 5 = 50\,050$ test utterances from all set A, B, and C, among which there are 8008 distinct digit sequences corrupted under various distortion conditions.

In Table II, detailed recognition rates (%) for each of the four noise conditions and for each of the SNR's in Set-A using front-end (1) are provided. The same results for Set-B and Set-C are presented in Tables III and IV, respectively, with different noise types and distortion conditions.

C. Effects of Different Ways of Incorporating Joint Static and Dynamic Priors

As one major contribution of this work to speech enhancement, the frame-differential dynamic cepstral features (in the log-domain) are used, in conjunction with the static cepstra, in the prior speech model in the Bayesian statistical framework to remove additive noise in the linear domain. We have systematically carried out experiments to evaluate the effects of incorporating such a new source of prior in our speech feature enhancement applications. The comprehensive experimental results are shown in Table V, where the full Aurora2 test sets (A, B, and C) are used and HMMs are trained with clean speech.

In Table V, the percent speech recognition accuracy is shown for a wide range of degree, signified by varying values of the variance scaling factor ρ in (34) and (35), to which the dynamic prior information is jointly used with the static prior information. When setting $\rho = 0$, we have $\mathbf{V}_1 = 0$ and $\mathbf{V}_2 = \mathbf{W}_1$. From (31), we see that the term associated with the use of the static prior is eliminated and the dynamic prior becomes the entire prior at work. At the other extreme, when $\rho \to \infty$ (set at a very large number of 5000 in the program), we have $\mathbf{V}_2 = 0$ and $\mathbf{V}_1 = \mathbf{W}_1$. Under this condition, the term associated with the use of the dynamic prior is eliminated and the static prior becomes the entire prior at work. We note that neither of these extreme conditions produces good performance. The peak performance is reached for the value of the variance scaling factor in the range between five and six. Based on the analysis provided in (26) of Section IV, this suggests that the variance of the conditional MMSE estimator, $\boldsymbol{\Sigma}^{\hat{\mathbf{x}}_{t-1}}$, would be on average in the range of four to five times of the estimated variance for the dynamic parameter using the clean-speech training set. This analysis elegantly accounts for the need for using variance scaling in order to achieve the optimally performance. In summarizing the above results and analysis, we conclude that the direct use of the estimated variances from clean speech (i.e., $\rho = 1$) without

TABLE V
SPEECH RECOGNITION ACCURACY (%) AS A FUNCTION OF VARIANCE SCALING FACTOR ρ IN (34) AND (35). ρ ADJUSTS THE RELATIVE CONTRIBUTIONS OF THE STATIC AND DYNAMIC FEATURE DISTRIBUTIONS IN THE PRIOR SPEECH MODEL TO SPEECH ENHANCEMENT. THE LAST COLUMN LISTS THE RECOGNITION ACCURACY WHEN NEITHER THE STATIC NOR THE DYNAMIC PRIOR INFORMATION IS USED DIRECTLY WHEN THE WEIGHT $\mathbf{W}_1$ IS SET TO ZERO. FULL TEST SETS (A, B, AND C) ARE USED. HMMS ARE TRAINED WITH CLEAN SPEECH. CEPSTRAL MEAN NORMALIZATION IS USED AFTER SPEECH FEATURE ENHANCEMENT

ρ	0	1	4.5	5	5.5	6	10	50	5000	$\mathbf{W}_1 = 0$
Rec.Acc(%)	78.89	83.98	85.32	85.82	85.84	85.80	85.15	84.00	83.74	77.08

taking account of the variance of the estimator does not lead to the optimal performance. The optimal balance for the joint use of the static and dynamic prior information depends on the quality of the estimator, which is measured by the variance of the estimator.

In the final column of Table V, we also list the recognition accuracy when the weight $\mathbf{W}_1$ is set to zero, or equivalently $\mathbf{V}_1 = \mathbf{V}_2 = 0$ and $\mathbf{V}_3 = 1$ in (31). That is, no prior information from either the static or dynamic portion is used directly. The prior information is only minimally and indirectly used in computing the outer-loop weight of posterior γ_m of (31) and in initializing the estimator in (36). This is a highly simplified, special case of our full algorithm of (31), which, as shown in Table V, has produced very poor performance. It is interesting to note that this special case corresponds to the algorithm described in [30], [24], where no similar $\mathbf{V}_1$ and $\mathbf{V}_2$ terms exist for direct and explicit use of the speech prior. Since $\mathbf{W}_1 = 0$ is equivalent to $\mathbf{\Psi} = 0$, our algorithm can be viewed as a principled way of generalizing that in [30], [24] by taking into account the errors in using the prediction $\hat{\mathbf{y}}$ to approximate the true observation $\mathbf{y}$ in (10). It is also interesting to note that the recognition accuracy reported in [30], using the identical training and test sets as we have used to obtain the results in Table V, is comparable to our special case under the same (poor) approximations. The somewhat minor difference, 78.37% in [30] vs. 77.08% in Table V, may be accounted for by their use of additional bandpass filtering which we did not use, and possibly by different implementations of the algorithm such as their distributed assignment of the nonlinear function's argument, versus our single-point initialization described in Section V-A.

D. Effects of Using Iterations for Estimator Refinement

In this subsection, we report the experimental results on the role of using iterations to refine the conditional MMSE estimator as described in Section V. While we gave theoretical motivations for using iterations in Section V.A, we show empirical evidence here now. In Table VI is shown the speech recognition accuracy for the full Aurora2 test sets as a function of the within-frame iteration number J in the iterative algorithm of Section V.C. Consistent improvement of recognition rates is achieved as the iteration number J increases from one to three, for both cases where the optimal balance is used between the static and dynamic priors (top row in Table VI) and where only the static prior is used (bottom row). After iteration three, the accuracy tapers off slowly. We note that most of the previously published, related algorithms correspond to using a single within-frame iteration $J = 1$. In this study, we show that more iterations improve the robust speech recognition performance due to the greater accuracy in the Taylor series approximation to the nonlinear acoustic distortion model. However, the optimal balance of the iteration number has only be explored empirically in this study.

TABLE VI
SPEECH RECOGNITION ACCURACY (%) AS A FUNCTION OF THE WITHIN-FRAME ITERATION NUMBER J USED IN THE ITERATIVE CONDITIONAL MMSE ENHANCEMENT ALGORITHM DESCRIBED IN SECTION V.C. CEPSTRAL MEAN NORMALIZATION IS USED AFTER SPEECH FEATURE ENHANCEMENT

Iteration number J	1	2	3	4	5
Rec.Acc (%) ($\rho = 5.5$)	83.94	85.69	85.84	85.66	85.57
Rec.Acc (%) ($\rho \to \infty$)	83.29	83.50	83.74	84.46	84.40

TABLE VII
THE SAME RECOGNITION ACCURACY RESULTS AS IN TABLE VI, EXCEPT NO CEPSTRAL MEAN NORMALIZATION IS USED

Iteration number J	1	2	3	4	5
Rec.Acc (%)($\rho = 5.5$)	83.13	84.00	84.80	84.65	84.55
Rec.Acc (%)($\rho \to \infty$)	81.99	82.50	83.00	83.41	83.40

TABLE VIII
THE SAME SUMMARY RECOGNITION ACCURACY RESULTS AS IN TABLE I, EXCEPT NO CEPSTRAL MEAN NORMALIZATION IS APPLIED

Methods	Set A	Set B	Set C	Overall
Prior: Static-Dynamic Cepstra	85.66	86.15	80.40	**84.80**
Prior: Static Cepstra only	84.20	84.72	77.17	83.00
Prior: Dynamic Cepstra only	78.07	76.29	76.74	77.09

E. Effects of Cepstral Mean Normalization

In the current implementation of the speech feature enhancement algorithm derived in Section IV, no channel distortion has been taken into account. Rather, the effects of channel distortion are handled via a separate process of cepstral mean normalization after the conditional MMSE estimator is computed. All the results presented so far include this latter process. To examine how effective this separate process is, we present in Table VII the speech recognition results analogous to Table VI, except by removing cepstral mean normalization in both training and testing. In all cases, the accuracy improvement using cepstral mean normalization is consistent but relatively minor.

We also present in Table VIII the summary recognition results analogous to Table I (except no cepstral mean normalization is used), where a breakdown is shown between the results of Sets A/B (additive noise only) and Set C (additive noise plus channel mismatch). Uniform degradation of the performance is observed.

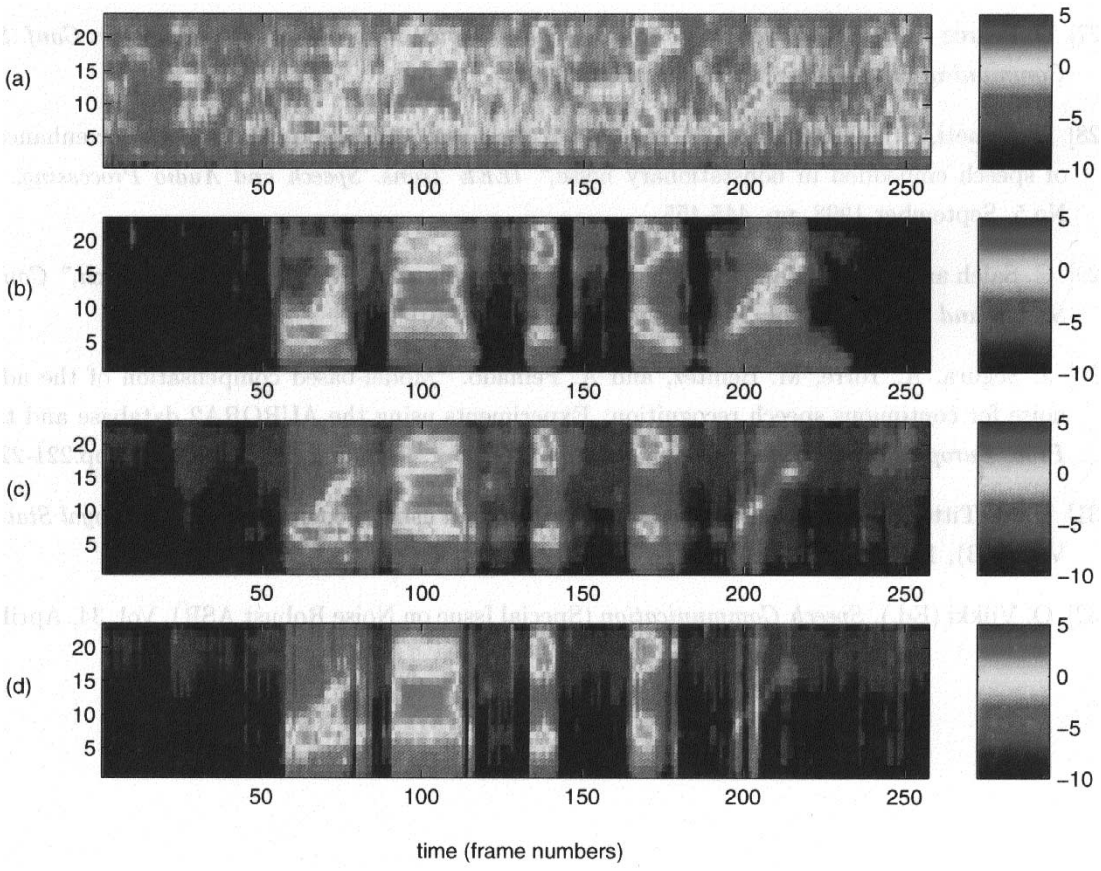

Fig. 1. Comparing spectrograms of two versions [(c) and (d)] of the enhanced speech with and without using the dynamic feature distribution in the prior speech model. From top to bottom: (a) noisy speech ($\mathrm{SNR} = 0$ dB); (b) true clean speech; (c) enhanced speech by the conditional MMSE estimator using joint static and dynamic feature distributions in the prior speech model; and (d) enhanced speech by the MMSE estimator using only the static feature distribution in the prior speech model.

F. Illustrations of the Effects of Incorporating Dynamic Prior

The above speech recognition results have consistently proved the benefits of incorporating the dynamic prior for the conditional MMSE estimator computation in the overall performance. We in this subsection further examine some fine-grained properties that illustrate the underlying reasons for the performance improvement. We do this by comparing the acoustic properties of the enhanced speech features using different prior information, and by comparing them with the target, true speech as well as with the initial distorted speech before the enhancement.

In Fig. 1 are four spectrograms,[9] with and without feature enhancement, for one Aurora2 test utterance corrupted by nonstationary babble noise with an average SNR of 0 dB. From the top panel to the bottom one are: (a) noisy speech before feature enhancement; (b) true clean speech as the target result for enhancement; (c) enhanced speech as the conditional MMSE estimate using joint static and dynamic feature distributions in the prior speech model; and (d) enhanced speech using only the static feature distribution in the model. Comparing the spectrograms of (c) and (d), we observe a greater degree of smoothness across frames, due to the use of the dynamic prior. In particular, during frames starting at 200, the smooth formant transition has been largely recovered in the enhanced features when the dynamic prior is used [comparing (c) with (b)]. In contrast, such a smooth formant transition in the clean speech has been mostly eliminated when no dynamic prior is used [panel (c)].

Fig. 2 shows the same kind of spectrogram comparisons as in Fig. 1, but using a new test utterance and with a different type of additive noise (the same average SNR of 0 dB). The same smoothness across frames in the enhanced features of panel (c) can be seen, contrasting sharply with the frequent abruptness across frames in the spectrogram shown in panel (d).

Fig. 3 shows yet another example of the spectrogram comparisons for a new utterance that contains locally negative dB SNR during the last digit with frames 110–140 ($\mathrm{SNR} = 0$ dB averaged over the utterance as in Figs. 1 and 2). By comparing panels (b), (c), and (d), we observe that the both enhancement algorithms cut off some major clean speech features in the negative-SNR region, but the algorithm incorporating the dynamic prior (c) is still doing a lot better than the one without the dynamic prior (d).

When the SNR is increased from 0 dB to 5 dB, spectrogram comparisons presented in Fig. 4 demonstrate significantly reduced differences between using and without using the dynamic prior. In fact, most of the speech recognition performance improvement we observed in the experiments reported in this section comes from the $\mathrm{SNR} = 0$ dB case, consistent with the general observation illustrated in Figs. 1–4.

[9]The spectrograms are computed by multiplying the inverse Cosine transformation matrix to each of the cepstral vectors in the speech utterance on a frame-by-frame basis.

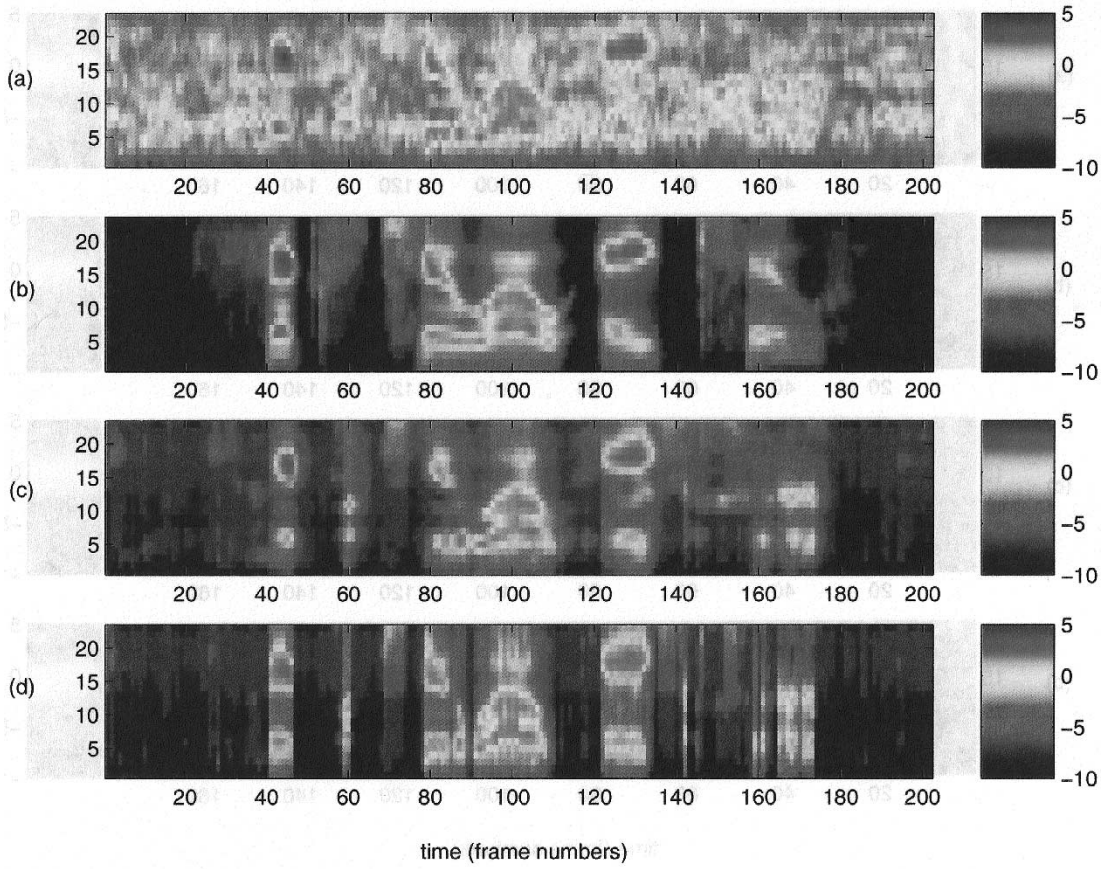

Fig. 2. The same spectrogram comparisons as in Fig. 1 using a new test utterance and with a different type of additive noise ($\mathrm{SNR} = 0$ dB).

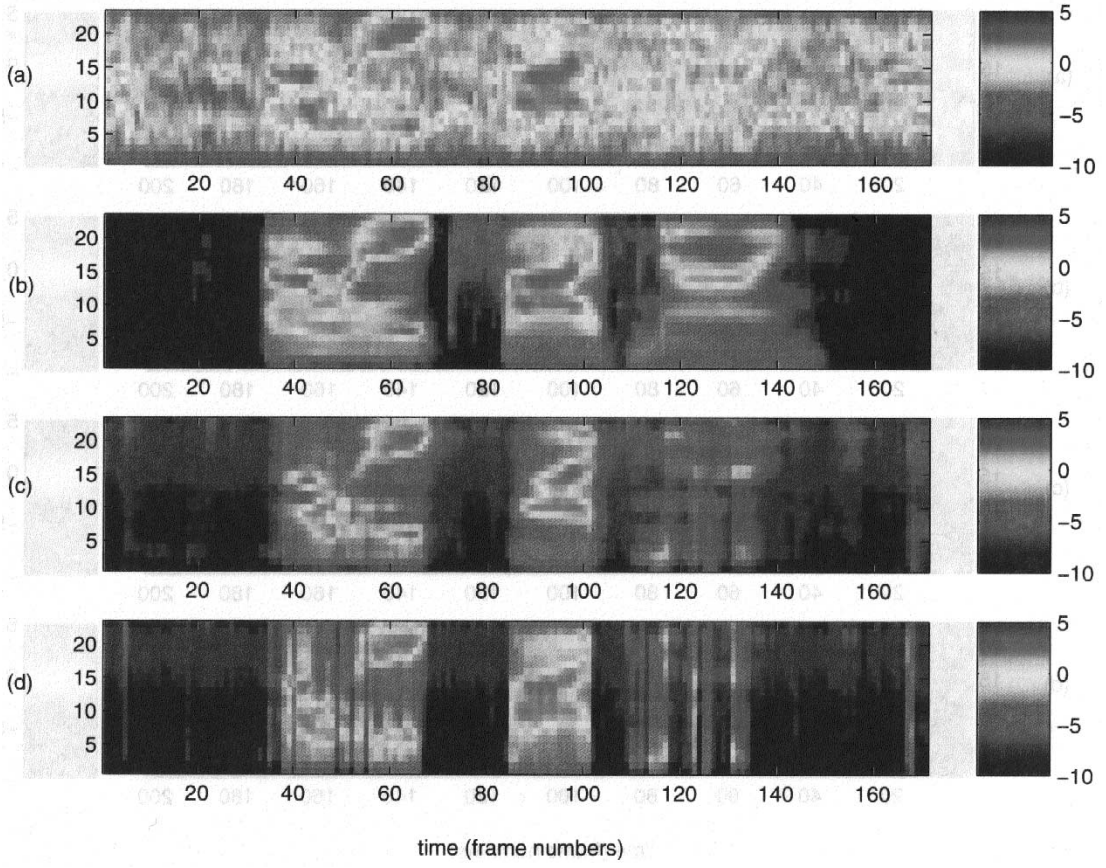

Fig. 3. Spectrogram comparisons with an example test utterance showing locally negative dB SNR ($\mathrm{SNR} = 0$ dB averaged over the utterance as in Figs. 1 and 2).

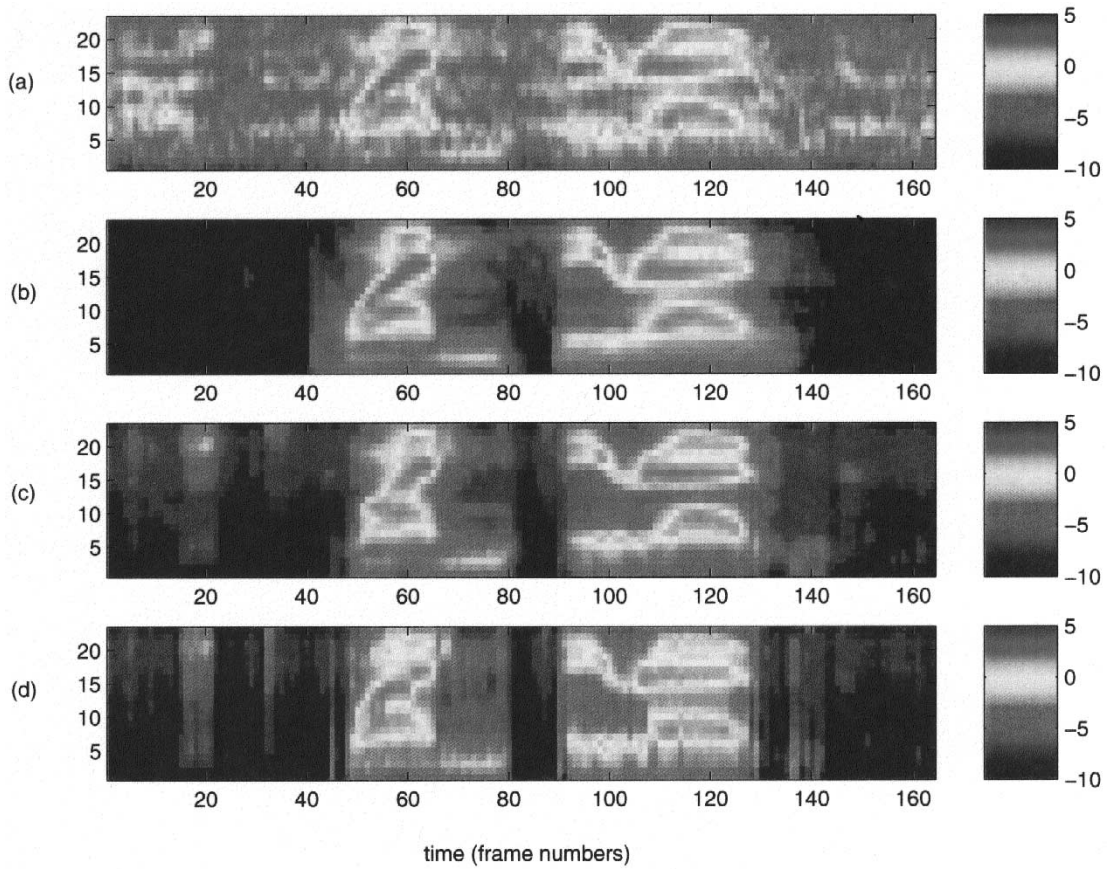

Fig. 4. Spectrogram comparisons where the average SNR is 5 dB. The difference is reduced between using and without using the dynamic feature distribution in the prior speech model.

VII. Summary and Discussions

In this paper, a novel algorithm with its detailed derivation, implementation, and evaluation is presented for statistical speech feature enhancement in the cepstral domain. It incorporates the joint static and dynamic cepstral features in the prior speech model in the Bayesian framework for optimal estimation of the clean speech features. The estimator is based on the full posterior computation and it elegantly integrates the predictive information from a statistical nonlinear acoustic distortion model, the prior information based on the static prior, and the prior based on the frame-differential dynamic prior. We have efficiently implemented this algorithm, which is used in the Aurora2 noise-robust speech recognition under the clean training condition. The results demonstrate significant improvement in the recognition accuracy by incorporating the joint static/dynamic prior, compared with using only the static or dynamic prior and with using no prior.

While the noisy speech data provided by Aurora2 is not real-life data such as Aurora3, the success of our approach should in principle be able to generalize as long as the noise can be accurately estimated. Some earlier related denoising algorithms developed in our lab have been generalized from Aurora2 to Aurora3 and other internal data with real-life acoustic distortion without any difficulty [15] using the noise tracking algorithm developed in [11] (also described at the end of Section III of this paper). This verifies the effectiveness of the noise tracking algorithm for real-life noisy speech data, and also suggests that the new denoising algorithm presented in this paper based on the same noise estimate can also generalize well.

While a strong, log-domain speech prior model is exploited and described in this paper, one specific limitation of the current work is its use of a rather weak noise prior model. The prior information about noise has been made deterministic in terms of its point estimate in this work, rather than probabilistic by incorporating the noise variance estimate as well. Conventional spectral subtraction techniques may thus use the same noise estimate to remove the noise by converting it back from the log domain to the linear spectral domain. However, in addition to the need for setting several heuristic, error-prone parameters (such as spectral floor and overestimation factor, etc.), direct noise subtraction does not make use of the prior information about speech. That is, after obtaining the subtraction residual, which gives the spectrum estimate for clean speech, there are no principled ways of verifying that the residual looks like clean speech. If the residual does not look like the clean speech, it would be highly desirable to reject the spectral subtraction result and re-try the subtraction procedure, perhaps with different setups for the heuristic parameters. The technique presented in this paper can be viewed as a formal framework to achieve just that. The measure for how the "residual"[10] looks like clean speech

[10]This becomes the conditional MMSE estimate in our statistical framework, in place of the spectral difference between the noisy speech and noise estimate in spectral subtraction.

is taken to be how likely it is to have scored the prior model that characterizes the statistical property of clean speech. One major innovation introduced in this work is to exploit both the static and dynamic features in establishing this prior model so that when examining to what degree the "residual" fits with the clean speech statistics, not only the spectral shape of each individual frame is taken into account, but also the trajectory of these shapes across time frames is used as a measure for the "fitting". All the above intuitively appealing aspects of noise reduction are gracefully integrated into a consistent statistical framework as presented in this paper. This framework can thus be viewed as comprehensive and probabilistic "noise removal", based on any fixed noise estimate either in the linear spectral domain (as in spectral subtraction), or in the log spectral domain related to the linear-domain variables via a nonlinear environment model (as has been explored in this paper).

In addition to improving the prior model for the additive corrupting noise, further work will also include estimation of channel distortion and its optimal use in an extended version of the Bayesian enhancement framework described in Section IV. The results presented in Section VI.E demonstrated that the simple use of cepstral mean normalization does not accomplish the desired goal of compensating the channel distortion.

While the "ignorance" modeling approach adopted to quantitatively represent the prediction residual in the statistical model of (9) for the nonlinear acoustic distortion (Section II) has been shown to be reasonably successful, for greater success it is desirable to provide a more accurate, "mechanistic" model for the prediction residual. We are currently working toward such a model and the related, new Bayesian estimation approach to speech feature enhancement.

Finally, the optimal estimator presented in Section IV can be easily extended to include the conditional variance estimation, in addition to the conditional mean (point) estimation derived in this paper. Given both the mean and variance estimates for the enhanced speech features, the heuristic variance scaling procedure discussed in Section V-B can be eliminated, and our future work will also be able to aim at a tight integration between the front-end denoising and the back-end speech recognition.

REFERENCES

[1] A. Acero, *Acoustic and Environmental Robustness in Automatic Speech Recognition*. Norwell, MA: Kluwer, 1993.

[2] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "HMM adaptation using vector Taylor series for noisy speech recognition," in *Proc. Int. Conf. Spoken Language Proc.*, vol. 3, 2000, pp. 869–872.

[3] H. Attias, J. Platt, A. Acero, and L. Deng, "Speech denoising and dereverberation using probabilistic models," *Advances in Neural Information Processing Systems (NIPS)*, vol. 13, pp. 758–764, 2000.

[4] H. Attias, L. Deng, A. Acero, and J. Platt, "A new method for speech denoising and robust speech recognition using probabilistic models for clean speech and for noise," in *Proc. Eur. Conf. Speech Communication*, 2001, pp. 1903–1906.

[5] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, 1979, pp. 208–211.

[6] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, pp. 113–120, 1980.

[7] L. Deng, "A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal," *Signal Process.*, vol. 27, pp. 65–78, 1992.

[8] L. Deng, A. Acero, M. Plumpe, and X. D. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Proc. Int. Conf. Spoken Language Processing*, vol. 3, 2000, pp. 806–809.

[9] L. Deng, A. Acero, L. Jiang, J. Droppo, and X. D. Huang, "High-performance robust speech recognition using stereo training data," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, 2001, pp. 301–304.

[10] L. Deng, M. Aksmanovic, D. Sun, and J. Wu, "Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 507–520, 1994.

[11] L. Deng, J. Droppo, and A. Acero, "Recursive estimation of nonstationary noise using a nonlinear model with iterative stochastic approximation," in *Proc. Automatic Speech Recognition and Understanding*, Dec. 2001, p. 4.

[12] L. Deng and J. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the hidden vocal-tract-resonance dynamics," *J. Acoust. Soc. Amer.*, vol. 108, no. 6, pp. 3036–3048, Dec. 2000.

[13] L. Deng, K. Wang, A. Acero, H. Hon, J. Droppo, C. Boulis, Y. Wang, D. Jacoby, M. Mahajan, C. Chelba, and X. D. Huang, "Distributed speech processing in MiPad's multimodal user interface," *IEEE Trans. Speech Audio Processing*, vol. 10, pp. 605–619, Nov. 2002.

[14] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora2 database," in *Proc. Eur. Conf. Speech Communication*, Aalborg, Denmark, Sept. 2001, pp. 217–220.

[15] J. Droppo, A. Acero, and L. Deng, "Evaluation of SPLICE on the Aurora2 and Aurora3 tasks," in *Proc. Int. Conf. Spoken Language Proc.*, Denver, CO, Sept. 2002, pp. 121–124.

[16] Y. Ephraim, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 33, pp. 443–445, 1985.

[17] ——, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, pp. 1526–1555, Oct. 1992.

[18] ——, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Processing*, vol. 40, pp. 725–735, 1992.

[19] B. Frey, L. Deng, A. Acero, and T. Kristjansson, "ALGONQUIN: Iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition," in *Proc. Eur. Conf. Speech Communication*, Aalborg, Denmark, Sept. 2001, pp. 901–904.

[20] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR2000 on Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, Sept. 2000.

[21] A. H. Jazwinski, *Stochastic Processes and Filtering Theory*. New York: Academic, 1970.

[22] J. H. Mathews and K. D. Fink, *Numerical Methods — Using MATLAB*, 3rd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[23] J. Mendel, *Lessons in Estimation Theory for Signal Processing, Communications and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[24] P. Moreno, "Speech Recognition in Noisy Environments," Ph.D. dissertation, CMU, 1996.

[25] P. Moreno, B. Raj, and R. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, vol. 1, 1996, pp. 733–736.

[26] M. Ostendorf, V. Digalakis, and J. Rohlicek, "From HMM's to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 4, pp. 360–378, 1996.

[27] "ESE2 special sessions on noise robust recognition," in *Proc. Eur. Conf. Speech Communication*, D. Pearce, Ed., Aalborg, Denmark, Sept. 2001.

[28] H. Sameti, H. Sheikhzadeh, L. Deng, and R. Brennan, "HMM-based strategies for enhancement of speech embedded in nonstationary noise," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 445–455, Sept. 1998.

[29] G. Saleh and M. Niranjan, "Speech enhancement using a Bayesian evidence approach," *Comput. Speech Lang.*, vol. 15, no. 2, pp. 101–125, Apr. 2001.

[30] J. Segura, A. Torre, M. Benitez, and A. Peinado, "Model-based compensation of the additive noise for continuous speech recognition: Experiments using the AURORA2 database and tasks," in *Proc. Eur. Conf. Speech Communication*, Aalborg, Denmark, Sept. 2001, pp. 221–224.

[31] D. M. Titterington, "Recursive parameter estimation using incomplete data," *J. R. Statist. Soc. B*, vol. 46, pp. 257–267, 1984.

[32] *Speech Commun.*, vol. 34, 2001.

Li Deng (S'83–M'86–SM'91) received the B.S. degree from University of Science and Technology of China in 1982, the M.S. degree from the University of Wisconsin-Madison in 1984, and the Ph.D. degree from the University of Wisconsin-Madison in 1986.

He worked on large-vocabulary automatic speech recognition in Montreal, QC, Canada, from 1986 to 1989. In 1989, he joined Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, as Assistant Professor; he became Full Professor in 1996. From 1992 to 1993, he conducted sabbatical research at Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, and from 1997 to 1998, at ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan. In 1999, he joined Microsoft Research, Redmond, WA, as Senior Researcher, and is currently a Principal Investigator in the DARPA-EARS Program and Affiliate Professor of electrical engineering at University of Washington, Seattle. His research interests include acoustic–phonetic modeling of speech, speech and speaker recognition, speech synthesis and enhancement, speech production and perception, auditory speech processing, noise robust speech processing, statistical methods and machine learning, nonlinear signal processing, spoken language systems, multimedia signal processing, and multimodal human–computer interaction. In these areas, he has published over 200 technical papers and book chapters, and has given keynote, tutorial, and other invited lectures. He recently completed the book *Speech Processing—A Dynamic and Optimization-Oriented Approach* (New York: Marcel Dekker, 2003).

Dr. Deng served on Education Committee and Speech Processing Technical Committee of the IEEE Signal Processing Society during 1996–2000, and is currently serving as Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING.

Jasha Droppo received the B.S. degree in electrical engineering (cum laude, with honors) from Gonzaga University in 1994. He received the M.S. degree in electrical engineering and the Ph.D degree in electrical engineering from the University of Washington, Seattle, under Les Atlas in 1996 and 2000, respectively. At the University of Washington, he helped to develop and promote a discrete theory for time-frequency representations of audio signals, with a focus on speech recognition.

He joined the Speech Technology Group at Microsoft Research, Redmond, WA, in 2000. His academic interests include noise robustness and feature normalization for speech recognition, compression, and time-frequency signal representations.

Alex Acero (S'83–M'90–SM'00–F'03) received an engineering degree from the Polytechnic University of Madrid, Spain, in 1985, an M.S. degree from Rice University, Houston, TX, in 1987, and the Ph.D. degree from Carnegie Mellon University, Pittsburgh, PA, in 1990, all in electrical engineering.

He was a Senior Voice Engineer at Apple Computer (1990–1991) and Manager of the Speech Technology Group at Telefonica Investigacion y Desarrollo (1991–1993). He joined Microsoft Research, Redmond, WA, in 1994, where he is currently Manager of the Speech Group. He is also Affiliate Professor at the University of Washington, Seattle. He is author of the books *Spoken Language Processing* (Englewood Cliffs, NJ: Prentice-Hall, 2000) and *Acoustical and Environmental Robustness in Automatic Speech Recognition* (Norwell, MA: Kluwer, 1993). He also has written chapters in three edited books, has eight patents, and over 80 technical publications. His research interests include noise robustness, signal processing, acoustic modeling, statistical language modeling, spoken language processing, speech-centric multimodal interfaces, and machine learning. He is associate editor of *Computer Speech and Language*.

Dr. Acero has had several positions within the IEEE Signal Processing Society, including Member-at-Large of the Board of Governors, associate editor of IEEE SIGNAL PROCESSING LETTERS, and as Member (1996–2000) and Chair (2000–2002) of the Speech Technical Committee. He was General Co-Chair of the 2001 IEEE Workshop on Automatic Speech Recognition and Understanding, Sponsorship Chair of the 1999 IEEE Workshop on Automatic Speech Recognition and Understanding, and Publications Chair of ICASSP '98.