Dynamic Compensation of HMM Variances Using the Feature Enhancement Uncertainty Computed From a Parametric Model of Speech Distortion

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

***Abstract*—This paper presents a new technique for dynamic, frame-by-frame compensation of the Gaussian variances in the hidden Markov model (HMM), exploiting the feature variance or uncertainty estimated during the speech feature enhancement process, to improve noise-robust speech recognition. The new technique provides an alternative to the Bayesian predictive classification decision rule by carrying out an integration over the feature space instead of over the model-parameter space, offering a much simpler system implementation, lower computational cost, and dynamic compensation capabilities at the frame level. The computation of the feature enhancement variances is carried out using a probabilistic and parametric model of speech distortion, free from the use of any stereo training data. Dynamic compensation of the Gaussian variances in the HMM recognizer is derived, which is simply enlarging the HMM Gaussian variances by the feature enhancement variances. Experimental evaluation using the full Aurora2 test data sets demonstrates a significant digit error rate reduction, averaged over all noisy and signal-to-noise-ratio conditions, compared with the baseline that did not exploit the enhancement variance information. When the true enhancement variances are used, further dramatic error rate reduction is observed, indicating the strong potential for the new technique and the strong need for high accuracy in estimating the variances associated with feature enhancement. All the results, using either the true variances of the enhanced features or the estimated ones, show that the greatest contribution to recognizer's performance improvement is due to the use of the uncertainty for the static features, next due to the delta features, and the least due to the delta–delta features.**

***Index Terms*—Dynamic variance compensation, hidden Markov model (HMM) variance, noise-robust automatic speech recognition (ASR), parametric environment model, speech feature enhancement, uncertainty in feature enhancement.**

I. Introduction

EFFECTIVE exploitation of variances or uncertainty is a key essence in nearly all branches of statistical pattern recognition. In the already successful applications of hidden Markov model (HMM) based robust speech recognition, uncertainty in the HMM parameter values has been represented by their statistical distributions [9], [11]. The motivation of such model-space Bayesian approaches has been the widely varied speech properties due to many possible sources of differences, including speakers and acoustic environments, across and possibly within training and test data. In order to take advantage of the model parameter uncertainty, the decision rule for recognition or decoding has been improved from the conventional MAP rule to Bayesian predictive classification (BPC) rule [8]. The former, MAP rule is described by

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} p(\mathbf{x}|\Lambda, \mathbf{W}) P(\mathbf{W}) \tag{1}$$

where $P(\mathbf{W})$ is the prior probability that the speaker utters a word sequence $\mathbf{W}$, and $P(\mathbf{x}|\Lambda, \mathbf{W})$ is the probability that the speaker produces the acoustic feature sequence, $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots \mathbf{x}_T]$, when $\mathbf{W}$ is the intended word sequence. Computation of the probability $P(\mathbf{x}|\Lambda, \mathbf{W})$ uses deterministic parameters, denoted by Λ, in the speech model.

When the parameters Λ of the speech model are made random to take account their uncertainty, the improved BPC rule requires integration over all possible parameter values [8]

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} \left[\int_{\Lambda \in \Omega} p(\mathbf{x}|\Lambda, \mathbf{W}) p(\Lambda|\phi, \mathbf{W}) d\Lambda \right] P(\mathbf{W}) \tag{2}$$

where ϕ is the (deterministic) hyper-parameters characterizing the distribution of the random model parameters, Ω denotes all possible values that the feature vector sequence $\mathbf{x}$ can take, and the integral becomes the desired acoustic score.

An alternative, which we will explore in depth in this paper, to the model-space characterization of uncertainty (e.g., BPC discussed above) is to represent the uncertainty by integrating over the feature space instead of over the model parameters. In addition to offering a much simpler system implementation and lower computational cost, accounting for uncertainty in the feature space (versus in the model space) has the added advantage of dynamic compensation at as fine a level as the individual feature frame. The uncertainty in the feature space can be established during a statistical feature enhancement or extraction process. While most of the feature enhancement algorithms developed in the past discard the uncertainty information [3]–[5], such side information available from most of these algorithms can be effectively taken advantage of to improve the recognition decision rule. More detailed motivations for making use of the feature-space uncertainty, called "uncertainty decoding," can be found in our recent work [6], where positive results were reported based on a specific, stereo-based feature enhancement al-

Manuscript received November 22, 2002; revised July 5, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Maurizio Omologo.

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.2005.845814

1063-6676/$20.00 © 2005 IEEE

gorithm (SPLICE [4], [5]) under a matched training and testing condition.[1]

To relax the matched condition required of the SPLICE for effective uncertainty decoding, we in this paper present a new uncertainty decoding technique based on a statistical enhancement algorithm developed using a probabilistic and parametric model of speech distortion. This makes the development of the system free from any stereo training data, as was required by the uncertainty decoding approach described in [6]. In this new technique, dynamic compensation for the Gaussian variances in the HMM recognizer will be shown to be simply enlarging them by the feature enhancement variances estimated based on the parametric model. While the closest approach to our new technique for HMM-based uncertainty decoding appears to be that of [1], key differences exist. The most important difference is that in [1], the estimated uncertainty is used to modify the HMM parameters that are intended to match the noisy, unprocessed speech data. In contrast, our technique modifies the HMM parameters so that they match the enhanced speech data. In addition, the technique for estimating uncertainty in feature enhancement in [1] is very different from the estimation approach presented in this paper.

The organization of this paper is as follows. In Section II, we introduce the new decision rule that exploits the variances associated with feature enhancement computed dynamically using the parametric model of speech distortion. We show how dynamic compensation of the Gaussian variances in the HMM can be easily accomplished once the feature enhancement variances are estimated. Detailed computation for the feature enhancement variances, as well as the expectations, is presented in Section III. Main derivation steps are provided for such computation in the cases of using the prior clean-speech models for the static features alone and for the joint static and dynamic features. Comprehensive results obtained using the complete Aurora2 task is reported in Section IV. They demonstrate the effectiveness of the feature-space uncertainty decoding for noise-robust speech recognition under the full range of noisy and signal-to-noise-ratio (SNR) conditions supplied by the Aurora2 database. In particular, we show that when the true enhancement variances are used, further dramatic error rate reduction can be achieved, indicating the strong potential for the new technique and the strong need for high accuracy in estimating the feature enhancement variances. Finally, a summary, discussion, and conclusion of the work are provided in Section V.

II. New Decision Rule Exploiting Variance in Feature Enhancement

As discussed in the Introduction section, uncertainty decoding based on the feature-space variance information can be made highly dynamic, and it provides greater simplicity compared with the model-space uncertainty decoding strategy exemplified by the BPC decision rule of (2). The counterpart of the BPC rule in the feature space requires an integration over the uncertainty in the feature domain $\mathbf{x}$ (rather than over that in the model parameter Λ)

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} \left[\int_{\mathbf{x}\in\Upsilon} p(\mathbf{x}|\Lambda, \mathbf{W}) p(\mathbf{x}|\theta) d\mathbf{x} \right] P(\mathbf{W}) \quad (3)$$

where Λ is the fixed model parameters (no uncertainty), θ is the parameters characterizing the distribution, $p(\mathbf{x}|\theta)$, of the speech features determined by a statistical feature extraction algorithm, and Υ represents all possible values that the feature vector sequence $\mathbf{x}$ may take. Note that, unlike the model-domain uncertainty characterization by $p(\Lambda|\phi, \mathbf{W})$ in (2), $p(\mathbf{x}|\theta)$ in (3) can be reasonably assumed to be independent of the word identities $\mathbf{W}$ (and hence independent of model parameters Λ). Later in this section, we will show that the integral in (3) indeed corresponds to the desirable acoustic score after taking into account the feature-domain uncertainty.

The need for the use of the new decision rule (3) is based on our acceptance that no noise reduction or feature enhancement algorithm is perfect. Use of an estimated degree of the imperfection according to the distribution $p(\mathbf{x}|\theta)$ provides a mechanism to effectively mask some undesirable distortion effects. For example, the frames with a negative instantaneous SNR which are difficult to enhance can be automatically discounted when the variance in $p(\mathbf{x}|\theta)$ for these frames is sufficiently large. This mechanism may also effectively extend the HMM uncertainty to cover the gap between the true clean speech features and the estimated clean speech features.

There are two key issues concerning the use of the new decision rule (3) that exploits variances in statistical feature extraction or enhancement for improving noise-robust speech recognition. The first issue is: Given an estimate of the uncertainty in a feature enhancement algorithm, how to incorporate it into the recognizer's decision rule? We now first address this issue below.

Consider an HMM system with Gaussians as the state (s)-dependent output distribution: $p(\mathbf{x}|\Lambda_s) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s)$, where $\mathbf{x}$ denotes a clean speech feature vector. We now denote by $\hat{\mathbf{x}}$ the enhanced speech feature vector from $\mathbf{x}$, and denote the estimation error in the enhancement process as $\mathbf{e}$, where

$$\mathbf{x} = \hat{\mathbf{x}} + \mathbf{e}. \quad (4)$$

We further assume that the estimation error is a zero-mean Gaussian random variable

$$\mathbf{e} \sim \mathcal{N}(\mathbf{e}; \mathbf{0}, \boldsymbol{\Sigma}_{\hat{\mathbf{x}}})$$

where $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}}$ is the covariance matrix associated with feature enhancement, and is the key quantity studied in this paper.

Under these assumptions, we can easily compute the integral in (3) as

$$\begin{aligned} \int_{\mathbf{x}\in\Upsilon_1} p(\mathbf{x}|\Lambda) p(\mathbf{x}|\theta) d\mathbf{x} &= \int_{\mathbf{x}} p(\mathbf{x}|\Lambda_s) p(\mathbf{x}|\boldsymbol{\Sigma}_{\hat{\mathbf{x}}}) d\mathbf{x} \\ &= \int_{\mathbf{x}} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) \mathcal{N}(\mathbf{x}; \hat{\mathbf{x}}, \boldsymbol{\Sigma}_{\hat{\mathbf{x}}}) d\mathbf{x} \\ &= \mathcal{N}(\hat{\mathbf{x}}; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s + \boldsymbol{\Sigma}_{\hat{\mathbf{x}}}) \end{aligned} \quad (5)$$

[1] A similar motivation also appeared recently for HMM-based speaker recognition in [14], [15], and for HMM-based speech recognition in [1], [10].

where $p(\mathbf{x}|\mathbf{\Sigma}_{\hat{\mathbf{x}}}) = \mathcal{N}(\mathbf{x}; \hat{\mathbf{x}}, \mathbf{\Sigma}_{\hat{\mathbf{x}}})$ according to (4), and we used the well known equality

$$\int \mathcal{N}\left(x; \mu_1, \sigma_1^2\right) \mathcal{N}\left(x; \mu_2, \sigma_2^2\right) dx = \mathcal{N}\left(\mu_1; \mu_2, \sigma_1^2 + \sigma_2^2\right).$$

In (5), Υ_1 is used to represent all possible values that the feature vector at each time frame may take.

On the other hand, the desirable acoustic score for each state s when taking into account the feature-domain uncertainty expressed by variance $\mathbf{\Sigma}_{\hat{\mathbf{x}}}$ can be determined by

$$\begin{aligned} p(\mathbf{x}|s) &= \int_{\mathbf{e}} p(\mathbf{x}, \mathbf{e}|s) d\mathbf{e} \\ &= \int_{\mathbf{e}} p(\mathbf{x}|\mathbf{e}, s) p(\mathbf{e}) d\mathbf{e} \\ &= \int_{\mathbf{e}} \mathcal{N}(\mathbf{e}; \boldsymbol{\mu}_s - \hat{\mathbf{x}}, \mathbf{\Sigma}_s) \mathcal{N}(\mathbf{e}; \mathbf{0}, \mathbf{\Sigma}_{\hat{\mathbf{x}}}) d\mathbf{e} \\ &= \mathcal{N}(\hat{\mathbf{x}}; \boldsymbol{\mu}_s, \mathbf{\Sigma}_s + \mathbf{\Sigma}_{\hat{\mathbf{x}}}). \end{aligned} \tag{6}$$

The second line in (6) was obtained by using the following result:

$$\begin{aligned} p(\mathbf{x}|\mathbf{e}, s) &= p(\hat{\mathbf{x}} + \mathbf{e}|s) \\ &= \frac{1}{(2\pi)^{\frac{D}{2}} |\mathbf{\Sigma}_{\mathbf{s}}|^{\frac{1}{2}}} \\ &\quad \times \exp\left[-\frac{1}{2} \left[(\hat{\mathbf{x}} + \mathbf{e}) - \boldsymbol{\mu}_s \right]^{\mathrm{Tr}} \mathbf{\Sigma}_s^{-1} \left[(\hat{\mathbf{x}} + \mathbf{e}) - \boldsymbol{\mu}_s \right] \right] \\ &= \frac{1}{(2\pi)^{\frac{D}{2}} |\mathbf{\Sigma}_{\mathbf{s}}|^{\frac{1}{2}}} \\ &\quad \times \exp\left[-\frac{1}{2} \left[\mathbf{e} - (\boldsymbol{\mu}_s - \hat{\mathbf{x}}) \right]^{\mathrm{Tr}} \mathbf{\Sigma}_s^{-1} \left[\mathbf{e} - (\boldsymbol{\mu}_s - \hat{\mathbf{x}}) \right] \right] \\ &= \mathcal{N}(\mathbf{e}; \boldsymbol{\mu}_s - \hat{\mathbf{x}}, \mathbf{\Sigma}_s). \end{aligned} \tag{7}$$

The identical result in (6) and (5) shows that the integral $\int_{\mathbf{x} \in \Upsilon_1} p(\mathbf{x}|\Lambda) p(\mathbf{x}|\theta) d\mathbf{x}$ in (3) is indeed the desirable acoustic score $p(\mathbf{x}|s)$ after taking account the feature-domain uncertainty. Most importantly, this identical result for the acoustic score obtained by integrating over the feature-domain uncertainty is Gaussian for the enhanced speech feature vector $\hat{\mathbf{x}}$. It has the same mean vector $\boldsymbol{\mu}_s$ as in the Gaussian associated with the clean speech HMM state s. However, it has its variance that is increased by summing the variance $\mathbf{\Sigma}_s$ of the Gaussian of the clean speech HMM and the variance $\mathbf{\Sigma}_{\hat{\mathbf{x}}}$ associated with the uncertainty in feature enhancement.

Note that when the feature-enhancement variance can be computed on a frame-by-frame basis (giving rise to $\mathbf{\Sigma}_{\hat{\mathbf{x}}_t}$), then the result of (6) permits highly desirable dynamic compensation of HMM Gaussian variances on a frame-by-frame basis. This also is significantly simpler to implement than the model-space integration in (2).

The second issue concerning the use of the new decision rule (3) is: How to estimate the uncertainty in statistical feature enhancement? We address this issue in the next section in the context of a specific feature enhancement algorithm based on a specific parametric model of speech distortion.

III. Computing Uncertainty Based on a Parametric Model of Speech Distortion

A. Overview of a Parametric Model of Speech Distortion

The parametric model of speech distortion, similar to the ones described earlier in [3] and [7], is briefly reviewed here. This serves as the basis for robust feature extraction, from which the uncertainty [i.e., the Gaussian variance $\mathbf{\Sigma}_{\hat{\mathbf{x}}}$ in (6)] is computed. Let $\mathbf{y}$, $\mathbf{x}$, and $\mathbf{n}$ be single-frame vectors of log Mel-filter energies for the noisy speech, clean speech, and additive noise, respectively. These quantities can be shown to be governed by the following relationship (see the Appendix for a detailed derivation):

$$\begin{aligned} \mathbf{y} &= \mathbf{x} + \log\left[(\mathbf{1} + e^{\mathbf{n}-\mathbf{x}}) \left[\mathbf{1} + \frac{2\boldsymbol{\lambda} e^{\frac{\mathbf{n}-\mathbf{x}}{2}}}{(\mathbf{1} + e^{\mathbf{n}-\mathbf{x}})} \right] \right] \\ &\approx \mathbf{x} + \log(\mathbf{1} + e^{\mathbf{n}-\mathbf{x}}) + \frac{\boldsymbol{\lambda}}{\cosh\left(\frac{\mathbf{n}-\mathbf{x}}{2}\right)} \end{aligned} \tag{8}$$

where $\boldsymbol{\lambda}$ is the inner product between the clean speech and noise vectors of Mel-filter energies in the linear domain, and the last step of approximation uses the assumption that $\boldsymbol{\lambda} \ll \cosh(\mathbf{n} - \mathbf{x}/2)$.

In order to avoid complicated evaluation of the small prediction residual (8) of

$$\mathbf{r} = \frac{\boldsymbol{\lambda}}{\cosh(\frac{\mathbf{n}-\mathbf{x}}{2})} \tag{9}$$

we represent it by an "ignorance" model as a zero-mean, Gaussian random vector. This, thus, gives a probabilistic parametric model of

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

where $\mathbf{g}(\mathbf{z}) = \log(\mathbf{1} + e^{\mathbf{z}})$, and $\mathbf{r} \sim \mathcal{N}(\mathbf{r}; \mathbf{0}, \mathbf{\Psi})$.

The Gaussian assumption for the residual $\mathbf{r}$ in (9) allows straightforward computation of the conditional likelihood for the noisy speech vector according to

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

B. Computing Expectations of Enhanced Speech Features

We now discuss the computation of the expectations of enhanced speech features as the minimum mean square error (MMSE) estimates of clean speech given the speech distortion model of (11).

1) Prior Clean-Speech Model for Static Features: We first present a technique for computing the expectation of enhanced speech features using a prior clean-speech model for the static feature $\mathbf{x}_t$ alone. The following Gaussian-mixture distribution is assumed as the prior PDF:

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

In the experiments reported in this paper, we used $M = 256$, and the standard EM algorithm was used to train all model parameters.

For simplicity purposes, the prior model for noise is assumed to be a time-varying Dirac delta function

$$p(\mathbf{n}_t) = \delta(\mathbf{n}_t - \bar{\mathbf{n}}_t) \tag{12}$$

where $\bar{\mathbf{n}}_t$ is computed by a noise tracking algorithm described in [2] and is assumed to be known in the following description of the iterative MMSE estimation for the clean speech vectors.

Some derivation steps for the MMSE estimate described in this subsection have been given in [3], which will be briefly outlined here. Given the noisy speech observation vector $\mathbf{y}$, the MMSE estimate $\hat{\mathbf{x}}$ for the random vector $\mathbf{x}$ is one that minimizes the MSE distortion measure of $MSE \equiv E[(\mathbf{x}-\hat{\mathbf{x}})^T(\mathbf{x}-\hat{\mathbf{x}})]$, or

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

It is well known that the MMSE estimate is the expected value of the posterior probability $p(\mathbf{x}|\mathbf{y})$

$$\hat{\mathbf{x}} = E[\mathbf{x}|\mathbf{y}] = \int \mathbf{x} p(\mathbf{x}|\mathbf{y}) d\mathbf{x}. \tag{13}$$

Using Bayes rule and using the prior speech and noise models just described, this MMSE estimate becomes

$$\begin{aligned}\hat{\mathbf{x}} &= \frac{\int \mathbf{x} p(\mathbf{y}|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}}{p(\mathbf{y})} \\ &= \frac{\sum_{m=1}^{M} c_m \int\int \mathbf{x} p(\mathbf{n}) p(\mathbf{x}|m) p(\mathbf{y}|\mathbf{x},\mathbf{n}) d\mathbf{x} d\mathbf{n}}{p(\mathbf{y})} \\ &= \frac{\sum_{m=1}^{M} c_m \int \mathbf{x} p(\mathbf{x}|m) p(\mathbf{y}|\mathbf{x},\bar{\mathbf{n}}) d\mathbf{x}}{p(\mathbf{y})}.\end{aligned} \tag{14}$$

Substituting the parametric acoustic distortion model of (11) into (14) and carrying out the needed integration in an analytical form via the use of iterative Taylor series approximation (truncation to the first order), we have approximated the evaluation of the MMSE estimate in (14) using the following iterative procedure. First, train and fix all parameters in the clean speech model: c_m, $\boldsymbol{\mu}_m^x$, and $\boldsymbol{\Sigma}_m^x$. Then, compute the noise estimate, $\bar{\mathbf{n}}_t$, which has been described elsewhere [2], and compute the weighting matrices

$$\begin{aligned}\mathbf{W}_1(m) &= (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi}, \\ \mathbf{W}_2(m) &= \mathbf{I} - \mathbf{W}_1(m).\end{aligned} \tag{15}$$

Next, fix the total number, J, of intra-frame iterations. (Iterations are used to approximate, in an increasingly accurate manner, the nonlinear function $\mathbf{g}(\mathbf{n}-\mathbf{x})$ in (10) using truncated Taylor series expansion.[2]) 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 estimate by

$$\hat{\mathbf{x}}_t^{(1)} = \arg\max_{\boldsymbol{\mu}_m^x} \mathcal{N}\left[\mathbf{y}_t; \boldsymbol{\mu}_m^x + \mathbf{g}\left(\bar{\mathbf{n}}_t - \boldsymbol{\mu}_m^x\right), \boldsymbol{\Psi}\right]. \tag{16}$$

Then, execute the following steps for each time frame (and then sequentially over time frames).

- Step 1: Compute

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

where $\mathbf{g}^{(j)} = \log(\mathbf{1} + e^{\bar{\mathbf{n}}_t - \hat{\mathbf{x}}_t^{(j)}})$.

- Step 2: Update the MMSE estimate

$$\hat{\mathbf{x}}_t^{(j+1)} = \sum_m \gamma_t^{(j)}(m) \left[\mathbf{W}_1(m)\boldsymbol{\mu}_m^x + \mathbf{W}_2(m)\left(\mathbf{y}_t - \mathbf{g}^{(j)}\right)\right]. \tag{17}$$

- 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 $t = T$.

The expectation of the enhanced speech feature vector is obtained as the final iteration of the estimate above for each time frame

$$\boldsymbol{\mu}_{\hat{\mathbf{x}}_t} = \hat{\mathbf{x}}_t^{(J)}. \tag{18}$$

2) Prior Clean-Speech Model for Joint Static and Dynamic Features: We now discuss the computation of the MMSE estimate using a prior clean-speech model for the joint static feature $\mathbf{x}_t$ and delta feature $\Delta\mathbf{x}_t \equiv \mathbf{x}_t - \mathbf{x}_{t-1}$ according to the following Gaussian-mixture distribution:

$$p(\mathbf{x}_t, \Delta\mathbf{x}_t) = \sum_{m=1}^{M} c_m \mathcal{N}\left(\mathbf{x}_t; \boldsymbol{\mu}_m^x, \boldsymbol{\Sigma}_m^x\right) \mathcal{N}\left(\Delta\mathbf{x}_t; \boldsymbol{\mu}_m^{\Delta\mathbf{x}}, \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}\right) \tag{19}$$

where independence between the static and delta features is assumed.

For the joint prior case here, we have similar steps to the static-prior case above. First, train and fix all parameters in the clean speech model: c_m, $\boldsymbol{\mu}_m^x$, $\boldsymbol{\mu}_m^{\Delta\mathbf{x}}$, $\boldsymbol{\Sigma}_m^x$, and $\boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}$. The noise estimate, $\bar{\mathbf{n}}_t$, is the same as before. More complex weighting matrices are computed now due to the new delta-feature component in the prior speech model

$$\begin{aligned}\mathbf{V}_1(m) &= (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi} \left(\boldsymbol{\Sigma}_m^x + \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}\right)^{-1} \left(\boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}\right), \\ \mathbf{V}_2(m) &= (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi} \left(\boldsymbol{\Sigma}_m^x + \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}\right)^{-1} \boldsymbol{\Sigma}_m^x, \\ \mathbf{V}_3(m) &= (\boldsymbol{\Sigma}_m^x + \boldsymbol{\Psi})^{-1} \boldsymbol{\Sigma}_m^x.\end{aligned}$$

Next, similar iterative steps to the static-prior case above are taken, using the same initialization of the clean-speech estimate of (16). Due to the introduction of the new delta-feature component in the prior speech model, Step 2 for updating the MMSE estimate is changed from (17) to the following one that incorporates the contribution from the new delta-feature component:

$$\begin{aligned}\hat{\mathbf{x}}_t^{(j+1)} = &\sum_m \gamma_t^{(j)}(m) \left[\mathbf{V}_1(m)\boldsymbol{\mu}_m^x + \mathbf{V}_2(m)\boldsymbol{\mu}_m^{\Delta\mathbf{x}}\right] \\ &+ \left[\sum_m \gamma_t^{(j)}(m)\mathbf{V}_2(m)\right] \hat{\mathbf{x}}_{t-1}^{(J)} \\ &+ \left[\sum_m \gamma_t^{(j)}(m)\mathbf{V}_3(m)\right] \left(\mathbf{y}_t - \mathbf{g}\left(\bar{\mathbf{n}}_t - \hat{\mathbf{x}}_t^{(j)}\right)\right).\end{aligned} \tag{20}$$

Again, the expectation of the enhanced speech feature vector is obtained as the final iteration of the estimate for each time frame according to (18).

[2]In the experiments reported in this paper, we used $J = 3$ based on empirical convergence properties and computation considerations.

C. Computing Variances of Enhanced Speech Features

We now describe the techniques for computing variances of enhanced speech features separately, as with the computation of the expectations just presented, in the cases of using prior clean-speech models for the static features alone and for the joint static and dynamic features.

1) Prior Clean-Speech Model for Static Features: Given the expectation for the enhanced speech feature computed as described in Section III-B-1, the variance of the enhanced speech feature can now be computed according to

$$\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t} = E\left[\mathbf{x}_t\mathbf{x}_t^T|\mathbf{y}_t\right] - \boldsymbol{\mu}_{\hat{\mathbf{x}}_t}\boldsymbol{\mu}_{\hat{\mathbf{x}}_t}^T \tag{21}$$

where the second-order moment is

$$\begin{aligned} &E\left[\mathbf{x}_t\mathbf{x}_t^T|\mathbf{y}_t\right] \\ &= \int \mathbf{x}_t\mathbf{x}_t^T p(\mathbf{x}_t|\mathbf{y}_t,\bar{\mathbf{n}}_t)d\mathbf{x}_t \\ &= \frac{\int \mathbf{x}_t\mathbf{x}_t^T p(\mathbf{x}_t)p(\mathbf{y}_t|\mathbf{x}_t,\bar{\mathbf{n}}_t)d\mathbf{x}_t}{p(\mathbf{y}_t)} \\ &= \frac{\sum_{m=1}^M c_m \overbrace{\int \mathbf{x}_t\mathbf{x}_t^T p(\mathbf{x}_t|m)p(\mathbf{y}_t|\mathbf{x}_t,\bar{\mathbf{n}}_t)d\mathbf{x}_t}^{I_m(\mathbf{y}_t)}}{p(\mathbf{y}_t)}. \end{aligned} \tag{22}$$

After using the zeroth order Taylor series to approximate[3] the nonlinear function $\mathbf{g}(\bar{\mathbf{n}}_t - \mathbf{x}_t)$ (contained in $p(\mathbf{y}_t|\mathbf{x}_t,\bar{\mathbf{n}}_t)$; rf. (11)) by $\mathbf{g}_0(\bar{\mathbf{n}}_t - \mathbf{x}_0)$ (denoted below by $\mathbf{g}_0$ for short), the integral in (22) becomes

$$\begin{aligned} I_m(\mathbf{y}_t) &\approx \int \mathbf{x}_t\mathbf{x}_t^T \mathcal{N}(\mathbf{x}_t;\boldsymbol{\mu}_m^x,\boldsymbol{\Sigma}_m^x)\mathcal{N}(\mathbf{y}_t;\mathbf{x}_t+\mathbf{g}_0,\boldsymbol{\Psi})d\mathbf{x}_t \\ &= \int \mathbf{x}_t\mathbf{x}_t^T \mathcal{N}(\mathbf{x}_t;\boldsymbol{\mu}_m^x,\boldsymbol{\Sigma}_m^x)\mathcal{N}(\mathbf{x}_t;\mathbf{y}_t-\mathbf{g}_0,\boldsymbol{\Psi})d\mathbf{x}_t \\ &= \int \mathbf{x}_t\mathbf{x}_t^T \mathcal{N}\left[\mathbf{x}_t;\boldsymbol{\theta}_m(t),(\boldsymbol{\Sigma}_m^x+\boldsymbol{\Psi})^{-1}\boldsymbol{\Sigma}_m^x\boldsymbol{\Psi}\right]d\mathbf{x}_t \\ &\quad \times N_m(\mathbf{y}_t) \\ &= \left[(\boldsymbol{\Sigma}_m^x+\boldsymbol{\Psi})^{-1}\boldsymbol{\Sigma}_m^x\boldsymbol{\Psi} + \boldsymbol{\theta}_m\boldsymbol{\theta}_m^T\right] \times N_m(\mathbf{y}_t) \end{aligned} \tag{23}$$

where

$$\begin{aligned} \boldsymbol{\theta}_m(t) &= (\boldsymbol{\Sigma}_m^x+\boldsymbol{\Psi})^{-1}\left[\boldsymbol{\Psi}\boldsymbol{\mu}_m^x + \boldsymbol{\Sigma}_m^x(\mathbf{y}_t-\mathbf{g}_0)\right], \\ N_m(\mathbf{y}_t) &= \mathcal{N}\left[\mathbf{y}_t-\mathbf{g}_0;\boldsymbol{\mu}_m^x,\boldsymbol{\Sigma}_m^x+\boldsymbol{\Psi}\right] \\ &= \mathcal{N}\left[\mathbf{y}_t;\boldsymbol{\mu}_m^x+\mathbf{g}_0,\boldsymbol{\Sigma}_m^x+\boldsymbol{\Psi}\right]. \end{aligned}$$

The third line in (23) above was obtained using the well established result in Gaussian computation (setting $a = b = 1$, $\mu_1 = \boldsymbol{\mu}_m^x$, $\mu_2 = \mathbf{y}_t - \mathbf{g}_0$, $\sigma_1^2 = \boldsymbol{\Sigma}_m^x$, $\sigma_2^2 = \boldsymbol{\Psi}$)

$$\begin{aligned} &\mathcal{N}\left(ax;\mu_1,\sigma_1^2\right)\mathcal{N}\left(bx;\mu_2,\sigma_2^2\right) \\ &\qquad = \mathcal{N}(x;\mu,\sigma^2)\mathcal{N}\left(a\mu_2;b\mu_1,a^2\sigma_2^2+b^2\sigma_1^2\right) \end{aligned}$$

where

$$\mu = \frac{a\mu_1\sigma_2^2 + b\mu_2\sigma_1^2}{a^2\sigma_2^2+b^2\sigma_1^2}; \quad \sigma^2 = \frac{\sigma_1^2\sigma_2^2}{a^2\sigma_2^2+b^2\sigma_1^2}.$$

[3]In this approximation, $\mathbf{x}_0$ is the Taylor series' expansion point, which is iteratively updated.

Substituting the result of (23) into (22), we obtain

$$E\left[\mathbf{x}_t\mathbf{x}_t^T|\mathbf{y}_t\right] = \sum_{m=1}^M \eta_m(\mathbf{y}_t)\left[(\boldsymbol{\Sigma}_m^x+\boldsymbol{\Psi})^{-1}\boldsymbol{\Sigma}_m^x\boldsymbol{\Psi} + \boldsymbol{\theta}_m(t)\boldsymbol{\theta}_m^T(t)\right] \tag{24}$$

where

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

and where we used the result that $p(\mathbf{y}_t) = \sum_{m=1}^M c_m N_m(\mathbf{y}_t)$ for the denominator.

Equation (21) then gives the estimate of the variance for the (static) enhanced feature. In our implementation, an iterative procedure similar to the computation of the expectation described in Section III-B is used to estimate the variance also, for the same purpose of reducing errors caused due to the approximation of $\mathbf{g}(\bar{\mathbf{n}} - \mathbf{x})$ by $\mathbf{g}_0(\bar{\mathbf{n}} - \mathbf{x}_0)$. For each iteration, the variance estimate takes the final form of

$$\begin{aligned} \boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t} &= \sum_{m=1}^M \eta_m(\mathbf{y}_t)\left[(\boldsymbol{\Sigma}_m^x+\boldsymbol{\Psi})^{-1}\boldsymbol{\Sigma}_m^x\boldsymbol{\Psi} + \boldsymbol{\theta}_m(t)\boldsymbol{\theta}_m^T(t)\right] \\ &\quad - \left[\sum_m \gamma_t(m)\left(\mathbf{W}_1(m)\boldsymbol{\mu}_m^x + \mathbf{W}_2(m)(\mathbf{y}_t-\mathbf{g}_0)\right)\right] \\ &\quad \times \left[\sum_m \gamma_t(m)\left(\mathbf{W}_1(m)\boldsymbol{\mu}_m^x + \mathbf{W}_2(m)(\mathbf{y}_t-\mathbf{g}_0)\right)\right]^T \end{aligned} \tag{25}$$

after combining (21), (24), and (17). Note that the weights $\gamma_t(m)$ above in the form of posterior probability are computed for each of the iterations.

2) Prior Clean-Speech Model for Joint Static and Dynamic Features: We now use the expectation for the enhanced speech feature as described in Section III-B2 to compute the variance estimate with the use of the prior clean-speech model for joint static and dynamic features.

With the use of the new speech prior PDF in (19), the second-order moment in (22) is changed to

$$\begin{aligned} &E\left[\mathbf{x}_t\mathbf{x}_t^T|\mathbf{y}_t,\mathbf{x}_{t-1}\right] \\ &\approx \frac{\sum_{m=1}^M c_m \overbrace{\int \mathbf{x}_t\mathbf{x}_t^T p(\mathbf{x}_t|m,\hat{\mathbf{x}}_{t-1})p(\mathbf{y}_t|\mathbf{x}_t,\bar{\mathbf{n}}_t)d\mathbf{x}_t}^{I_m(\mathbf{y}_t)}}{p(\mathbf{y}_t)} \end{aligned} \tag{26}$$

where the conditional PDF can be written as shown in (27) at the bottom of the next page. After completing the squares for the exponent of (27), we have

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

where

$$\begin{aligned} \boldsymbol{\mu}_m &= \underbrace{\left(\boldsymbol{\Sigma}_m^x + \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}\right)^{-1}\boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}}_{w}\boldsymbol{\mu}_m^x \\ &\quad + \underbrace{\left(\boldsymbol{\Sigma}_m^x + \boldsymbol{\Sigma}_m^{\Delta\mathbf{x}}\right)^{-1}\boldsymbol{\Sigma}_m^x}_{1-w}\left(\hat{\mathbf{x}}_{t-1} + \boldsymbol{\mu}_m^{\Delta\mathbf{x}}\right) \end{aligned} \tag{29}$$

is a weighted sum of the contribution of "static-prior" $\boldsymbol{\mu}_m^x$ and that of the "dynamic-prior" $\hat{\mathbf{x}}_{t-1}+\boldsymbol{\mu}_m^{\Delta\mathbf{x}}$, and

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

Substituting (28) into (26) and using the same zeroth order Taylor series to approximate the nonlinear function $\mathbf{g}(\bar{\mathbf{n}}_t-\mathbf{x}_t)$ by $\mathbf{g}_0(\bar{\mathbf{n}}_t-\mathbf{x}_0)$ as in the static-prior case, we obtain the integral in (26) as

$$\begin{aligned}I_m(\mathbf{y}_t)&=\int\mathbf{x}_t\mathbf{x}_t^T\mathcal{N}(\mathbf{x}_t;\boldsymbol{\mu}_m,\boldsymbol{\Sigma}_m)\mathcal{N}(\mathbf{y}_t;\mathbf{x}_t+\mathbf{g}_0,\boldsymbol{\Psi})d\mathbf{x}_t\\&=\int\mathbf{x}_t\mathbf{x}_t^T\mathcal{N}\left[\mathbf{x}_t;\boldsymbol{\Theta}_m(t),(\boldsymbol{\Sigma}_m+\boldsymbol{\Psi})^{-1}\boldsymbol{\Sigma}_m\boldsymbol{\Psi}\right]d\mathbf{x}_t\\&\quad\times\bar{N}_m(\mathbf{y}_t)\\&=\left[(\boldsymbol{\Sigma}_m+\boldsymbol{\Psi})^{-1}\boldsymbol{\Sigma}_m\boldsymbol{\Psi}+\boldsymbol{\Theta}_m\boldsymbol{\Theta}_m^T\right]\times\bar{N}_m(\mathbf{y}_t)\end{aligned} \tag{31}$$

where

$$\begin{aligned}\boldsymbol{\Theta}_m(t)&=(\boldsymbol{\Sigma}_m+\boldsymbol{\Psi})^{-1}\left[\boldsymbol{\Psi}\boldsymbol{\mu}_m+\boldsymbol{\Sigma}_m(\mathbf{y}_t-\mathbf{g}_0)\right],\\\bar{N}_m(\mathbf{y}_t)&=\mathcal{N}\left[\mathbf{y}_t;\boldsymbol{\mu}_m+\mathbf{g}_0,\boldsymbol{\Sigma}_m+\boldsymbol{\Psi}\right].\end{aligned}$$

Substituting the result of (31) into (26), we obtain

$$E\left[\mathbf{x}_t\mathbf{x}_t^T|\mathbf{y}_t\right]=\sum_{m=1}^{M}\zeta_m(\mathbf{y}_t)\left[(\boldsymbol{\Sigma}_m+\boldsymbol{\Psi})^{-1}\boldsymbol{\Sigma}_m\boldsymbol{\Psi}+\boldsymbol{\Theta}_m(t)\boldsymbol{\Theta}_m^T(t)\right] \tag{32}$$

where

$$\zeta_m(\mathbf{y}_t)=\frac{c_m\bar{N}_m(\mathbf{y}_t)}{\sum_{m=1}^{M}c_m\bar{N}_m(\mathbf{y}_t)}.$$

The estimation is again carried out iteratively in order to improve Taylor-series approximation. For each iteration, the final form of the variance estimation as has been implemented is

$$\begin{aligned}\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t|\hat{\mathbf{x}}_{t-1}}&=\sum_{m=1}^{M}\zeta_m(\mathbf{y}_t)\\&\quad\times\left[(\boldsymbol{\Sigma}_m+\boldsymbol{\Psi})^{-1}\boldsymbol{\Sigma}^x\boldsymbol{\Psi}+\boldsymbol{\Theta}_m(t)\boldsymbol{\Theta}_m^T(t)\right]-\boldsymbol{\mu}_{\hat{\mathbf{x}}_t}\boldsymbol{\mu}_{\hat{\mathbf{x}}_t}^T\end{aligned} \tag{33}$$

where

$$\begin{aligned}\boldsymbol{\mu}_{\hat{\mathbf{x}}_t}&=\sum_m\gamma_t(m)\left[\mathbf{V}_1(m)\boldsymbol{\mu}_m^x+\mathbf{V}_2(m)\boldsymbol{\mu}_m^{\Delta\mathbf{x}}\right]\\&+\left[\sum_m\gamma_t(m)\mathbf{V}_2(m)\right]\hat{\mathbf{x}}_{t-1}+\left[\sum_m\gamma_t(m)\mathbf{V}_3(m)\right](\mathbf{y}_t-\mathbf{g}_0)\end{aligned}$$

is based on (20).

D. Computing Variances of Temporal Differences of the Enhanced Features

In our implementation, the temporal differences of the enhanced features, also referred to as the delta or dynamic features, are computed in the same manner as those for the clean speech features

$$\Delta\hat{\mathbf{x}}_t=\sum_{\tau=-K}^{K}w_\tau\hat{\mathbf{x}}_{t+\tau},\quad\Delta^2\hat{\mathbf{x}}_t=\sum_{\tau=-L}^{L}v_\tau\Delta\hat{\mathbf{x}}_{t+\tau} \tag{34}$$

where $K=3$, $L=2$, and the weights w_τ and v_τ are fixed. (The second-order delta feature in (34) is also called the acceleration feature.[4]) Under the assumptions of temporal independence and that the variances do not change over the time window from $-K$ to $+K$ and from $-L$ and $+L$, we can easily determine the variances for these delta features according to

$$\boldsymbol{\Sigma}_{\Delta\hat{\mathbf{x}}_t}=\left(\sum_{\tau=-K}^{K}w_\tau^2\right)\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t},\quad\boldsymbol{\Sigma}_{\Delta^2\hat{\mathbf{x}}_t}=\left(\sum_{\tau=-L}^{L}v_\tau^2\right)\boldsymbol{\Sigma}_{\Delta\hat{\mathbf{x}}_t} \tag{35}$$

where $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t}$ is already computed as described in Section III-C.

IV. Speech Recognition Experiments on the Aurora2 Task

We have described in Section III-B and Section III-C the expectation and variance estimates, which fully characterize the statistical distribution, assumed to be Gaussian, of the enhanced speech features. Given this distribution, the feature-space uncertainty decoding rule (3) can be used to perform speech recognition. We have evaluated this new decoding strategy on the Aurora2 database. The task is to recognize strings of connected English digits embedded in several types of artificially created distortion environments with a range of SNR's from 0 to 20 dB. Three sets of digit utterances (sets A, B, and C) are prepared as the test material. Set A is used for evaluating the system with matched training and testing additive noises, Set B with mismatched training and testing additive noises, and Set C with both channel and additive distortions.

In our current work, the decoding rule (3) is implemented in the digit HMM recognizer by adding $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t}$ or $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t|\hat{\mathbf{x}}_{t-1}}$ to the variances of all Gaussians in the HMM at each frame t, as described and justified in Section II, while using $\boldsymbol{\mu}_{\hat{\mathbf{x}}_t}$ as the observation vector. The HMM used as the backend of the recognizer is defined by the ETSI Aurora group [13]. Each digit HMM has 16 states with three Gaussians per state. The silence model has three states with six Gaussians per state. A one-state short-pause model is used and tied with the middle state of the silence model. Therefore, the additional computation due to the use of uncertainty decoding is 546 sums of diagonal covariance matrices, one for each of the HMM's Gaussians.

The original HMM's used for decoding (before adding the variance estimate $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t}$ or $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t|\hat{\mathbf{x}}_{t-1}}$) are trained using all clean speech files in the training set of the Aurora2 database. The

[4] $L=2$ gives a window of $(-2, 2)$ for delta features, which is equivalent to a window of $(-5, 5)$ for the original static features.

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

TABLE I
AURORA2 PERFORMANCE (PERCENT ACCURATE) EXPLOITING THE VARIANCES IN DIFFERENT SETS OF FEATURE STREAMS. UNCERTAINTY OR VARIANCES ARE COMPUTED USING THE ESTIMATION FORMULAS DESCRIBED IN SECTION III-B-1 AND SECTION III-C-1 BASED ON THE PRIOR CLEAN-SPEECH MODEL FOR STATIC FEATURES ALONE. AURORA2 REFERENCE RECOGNITION ACCURACY UNDER THE CONDITION OF CLEAN TRAINING IS LISTED IN THE LAST ROW

	set A	set B	set C	Ave.
I: MAP-rule (variances=0)	84.20	84.72	77.17	83.00
II: Static variance only	85.40	85.50	79.60	84.28
III: Static/Δ variances	86.11	85.90	80.20	84.84
IV: Static/Δ /Δ^2 variances	86.10	85.95	80.18	84.86
Aurora2 reference accuracy (clean training)	58.74	53.40	66.00	58.06

noise estimate used for computing both the expectations and variances of the enhanced features in the experiments below is based on the iterative stochastic approximation algorithm described in [2].

A. Results of Using Estimated Uncertainty in Different Sets of Feature Streams

The results of robust speech recognition using dynamic compensation of HMM variances, based on the feature enhancement uncertainty computed as described in Section III, are presented in this subsection. As with the description of computing feature enhancement uncertainty in Section III, we also present the results separately for the cases of using the prior clean-speech model for static features alone and for joint static and dynamic features.

1) Prior Clean-Speech Model for Static Features Alone: Table I presents the percent-accurate performance results on all three sets of the Aurora2 test data, averaged over all SNRs from 0 to 20 dB and over four (sets A/B) or two (set C) distortion conditions (each condition and SNR contains 1101 digit strings). Row I gives the baseline results using the conventional (plug-in) MAP rule (1) (i.e., "point" decoding), where the expectations of the enhanced speech feature vectors $\boldsymbol{\mu}_{\hat{\mathbf{x}}_t}$'s computed according to (18) described in Section III-B1 [jointly with $\Delta\hat{\mathbf{x}}_t$ and $\Delta^2\hat{\mathbf{x}}_t$ computed by (34)] are used as the observational feature vector sequence $\mathbf{x}$ in (1). The variances for all feature streams (static and dynamic) are set to zero: $\boldsymbol{\Sigma}_{\hat{\mathbf{x}}_t} = \boldsymbol{\Sigma}_{\Delta\hat{\mathbf{x}}_t} = \boldsymbol{\Sigma}_{\Delta^2\hat{\mathbf{x}}_t} = 0$.

Row II in Table I shows the recognizer's performance using the feature-space uncertainty decoding rule (3), where the variance of the static feature stream is computed according to (26) (using the prior clean-speech model for static features alone), while the variances of the dynamic feature streams are set to zero: $\boldsymbol{\Sigma}_{\Delta\hat{\mathbf{x}}_t} = \boldsymbol{\Sigma}_{\Delta^2\hat{\mathbf{x}}_t} = 0$. The overall improvement in the recognition accuracy from 83.00% to 84.28% corresponds to 7.5% digit error rate reduction. The error rate is further reduced, up to 10.9% reduction, when the variances ($\boldsymbol{\Sigma}_{\Delta\hat{\mathbf{x}}_t}$ and $\boldsymbol{\Sigma}_{\Delta^2\hat{\mathbf{x}}_t}$) of the dynamic feature streams are estimated by (35) rather than being set to zero (Rows III and IV). However, we observed that exploiting the variance of the acceleration feature stream ($\boldsymbol{\Sigma}_{\Delta^2\hat{\mathbf{x}}_t}$) has contributed to virtually no performance improvement once the variance of the delta feature stream has been exploited. One possible reason for this is the assumption made in computing variances for the dynamic feature streams (35) that these variances do not change over the time window. Since the time window $(-5, +5)$ for the acceleration feature stream is wider than that for the delta feature stream $(-3, +3)$, the assumption becomes less valid. Hence, the incorporation of the uncertainty for the acceleration feature stream is expected to become less effective.

The results in Table I (and in other tables) were obtained using the fixed iteration number $J = 3$. We found that an increased number of iterations has little influence on the performance but a reduced number ($J = 1$ or $J = 2$) degrades the performance appreciably. Also, all the results were obtained with the Gaussian mixture number fixed at $M = 256$. Decreasing M leads to slowly degraded recognition accuracy.

2) Prior Clean-Speech Model for Joint Static and Dynamic Features: In parallel with the results presented in Table I where the variances of feature enhancement are computed using the prior clean-speech model for static features alone, we now present the parallel results where the prior clean-speech model for joint static and dynamic features is used. Table II lists the percent-accurate performance results on the same three sets of the Aurora2 test data, where Row I is the baseline results using the MAP rule with the expectations of the enhanced speech feature vectors $\boldsymbol{\mu}_{\hat{\mathbf{x}}_t}$'s computed according to (20) described in Section III-B-2 (also appended by $\Delta\hat{\mathbf{x}}_t$ and $\Delta^2\hat{\mathbf{x}}_t$ computed from (34)). Row II shows the results using the feature-space uncertainty decoding rule when the variance of the static feature stream is used while the variances of the dynamic feature streams are set to zero. The overall performance in recognition accuracy is greater than the counterpart in Table I due to the use of the better prior model for computing the expectations.[5] The relative improvement via the use of feature-enhancement variances from 84.80% to 86.13% gives 8.8% digit error rate reduction, higher than the counterpart 7.5% in Table I. This may reflect the greater effectiveness when using the joint static and dynamic prior model for computing the feature enhancement variance also. Similar to the results in Table I, the error rate is also further reduced, totaling to 11.4% reduction, when the variances ($\boldsymbol{\Sigma}_{\Delta\hat{\mathbf{x}}_t}$ and $\boldsymbol{\Sigma}_{\Delta^2\hat{\mathbf{x}}_t}$) of the dynamic feature streams are estimated by (35) rather than being set to zero (Rows III and IV). Again, we observed that exploiting the variance of the acceleration feature stream ($\boldsymbol{\Sigma}_{\Delta^2\hat{\mathbf{x}}_t}$) has not contributed to performance improvement once the variance of the delta feature stream has been exploited.

It is worth pointing out that the use of the dynamic feature prior (in addition to the static feature prior) is proved to be im-

[5]This has been demonstrated in [3] and will not be elaborated here.

TABLE II
AURORA2 PERFORMANCE (PERCENT ACCURATE) EXPLOITING THE VARIANCES IN DIFFERENT SETS OF FEATURE STREAMS. UNCERTAINTY OR VARIANCES ARE COMPUTED USING THE ESTIMATION FORMULAS DESCRIBED IN SECTION III-B-2 AND SECTION III-C-2 BASED ON THE PRIOR CLEAN-SPEECH MODEL FOR JOINT STATIC AND DYNAMIC FEATURES

	set A	set B	set C	Ave.
I: MAP-rule (variances=0)	85.66	86.15	80.40	84.80
II: Static variance only	86.95	87.56	81.62	86.13
III: Static/Δ variances	87.38	87.74	82.44	86.54
IV: Static/Δ/Δ^2 variances	87.34	87.79	82.45	86.54

TABLE III
AURORA2 PERFORMANCE (PERCENT ACCURATE) USING THE VARIANCES DETERMINED BY SQUARING THE DIFFERENCES BETWEEN THE ESTIMATED AND TRUE CLEAN SPEECH FEATURES. THIS ELIMINATES BIASES IN THE VARIANCE ESTIMATION

	set A	set B	set C	Ave.
Static variance only	90.31	91.12	84.70	89.51
Static/Δ variances	93.80	94.00	89.50	93.02
Static/Δ/Δ^2 variances	94.87	95.49	90.75	94.29

portant for improving the uncertainty decoding technique proposed in this paper. This can be seen from the improvement of recognition accuracy from Table I to Table II across the board.

B. Results on the Performance Limit of Uncertainty Decoding

To investigate the upper limit of possible performance improvement by exploiting variances for feature-space uncertainty decoding, we desire to eliminate biases in the variance estimation based on (33) and (35). To achieve this, we conducted diagnostic experiments where the "true" variances are computed by squaring the differences between the estimated and true clean speech features. The true clean speech features are computed from the clean speech waveforms available from the Aurora2 database, and the estimated clean speech features (expectations) are computed as described in Section III-B2. The performance results of Table III are significantly better than those in Tables I and II. In particular, we observe that the exploitation of the variances of both the static and the dynamic feature streams cuts the error rate by about half compared with using the variance for the static feature stream only (see the accuracy difference 89.51% vs. 94.29% in Table III). In contrast, the corresponding performance difference is much smaller when the estimated variances (as opposed to the true ones) are used. These results suggest that the biases of the variance estimates that produced the results of Tables I and II are undesirably large, and that better variance estimates developed in future research will have the potential to drastically improve the recognition performance from those shown in Tables I and II toward those in Table III.

V. CONCLUSION

The research described in this paper extends our earlier work in speech feature enhancement and noise-robust recognition on two fronts. First, it extends the Bayesian technique for the point-estimate-based speech feature enhancement [3] by exploiting the distribution of the enhanced feature via integration over the feature space, leading to the new recognition decision rule which capitalizes on the uncertainty information in the enhancement process discarded by the previous enhancement technique. Second, it extends the uncertainty decoding technique [6] by using a new approach based on a parametric model of speech distortion to statistical feature enhancement free from the use of any stereo training data.

The new recognition decision rule developed in this work provides an alternative to the BPC decision rule by carrying out an integration over the feature space instead of over the model-parameter space. This offers a much simpler system implementation and lower computational cost. Most importantly, it allows for dynamic (at the frame level) compensation of the Gaussian variance parameters in the HMM, which has been impossible to accomplish by the BPC technique.

In this paper, we provide detailed computational steps and their derivation for the variances associated with speech feature enhancement as required by dynamic HMM variance compensation. Two novel algorithms for estimating the variance of enhanced speech features, with the use of the clean-speech prior models for static features only and for joint static and dynamic features, respectively, are presented in detail. The essence of the algorithms is to make use of a parametric model of speech distortion for computing the second-order moment of the clean speech under its posterior PDF, and analytical solutions have been successfully developed. This novelty differentiates our algorithms from all other techniques in the literature [1], [6], [10], [12], [14], [15] which also exploited feature uncertainty in robust speech recognition or speaker recognition. In [1], in order to determine feature uncertainty, a training database had to be created in which noise was artificially added to clean speech and a third-order polynomial was used to approximate the mapping function. This kind of "stereo" training data was also needed in our earlier work [6], which uses the SPLICE technique [4], [5] to compute the variance associated with speech feature enhancement. In [10], while using no stereo training data, a special technique (i.e., Algonquin [7]) motivated from machine learning was used to determine the entire distribution of the enhanced features.[6] This distribution was then integrated into a rather sophisticated decoding rule. In [15] where the feature uncertainty was applied to speaker verification, an empirical noise model was established to enable the computation of the feature variances. Numerical integration was required for the computation, and approximated expressions were also provided under high SNR conditions with some empirical rules. Finally, in [14], the feature variances were determined empirically by first using trend fitting to the features and then taking sample variances. In contrast, the variance estimation technique presented in this paper is free from either stereo data or from any special conditions and empirical rules, and it is very simple to integrate into the decoding rule. The only approximation used in our computation of the feature variances is Taylor series truncation, and the approximation accuracy has been increasingly improved via iterations.

One principal advantage of the feature-domain uncertainty decoding technique presented in this paper over the model-domain BPC technique is the significantly lower computation cost. With our technique, there is no change in the HMM decoding, and the main computational load is in the estimation of the feature enhancement uncertainty as presented in Section III-C.

[6]The enhanced features were also computed by the same Algonquin technique.

These computations are expressed by (26) and (33), where the estimation formula for one of a few iterations (typically three) are given explicitly. Compared with the extra decoding requirement for BPC, our feature-domain variance estimation is much less intensive in computation, even with the matrix operations and iterations. In practice, in our implementation of (26) and (33), all the matrices have been assumed to be diagonal. This further substantially cuts down the computational load.

The effectiveness of our new estimation algorithms for the variances of speech feature enhancement and their use in dynamic HMM variance compensation (uncertainty decoding) has been experimentally evaluated using the full Aurora2 test data sets. We have obtained consistent results with the use of the variance estimates computed with either the prior clean-speech model for static features alone or the model for joint static and delta features. For the former, dynamic HMM variance compensation reduces the digit recognition error rate by 7.5% and 10.9%, respectively, when the static feature stream and both static/dynamic feature streams are subject to the variance compensation. For the latter, the error rate reduction is 8.8% and 11.4%, respectively. All such performance improvement is compared with the baseline system of the decoding MAP rule, which was the best result reported in [3] that did not exploit the variance information. Finally, all the results obtained show consistently that the greatest contribution to recognizer's performance improvement is derived from the use of the uncertainty in feature enhancement for the static features, next from the delta features, and the least from the delta-delta acceleration features.

We also reported the results from a set of diagnostic experiments where the "true" variance of the enhanced speech features is provided to the uncertainty decoding rule for dynamic HMM variance compensation so that the gap between the true and the estimated clean speech features is fully covered. More than 50% of the digit errors, committed when the estimated variance is used, have been corrected. This provides a clear direction of our future research on improving the quality of uncertainty estimation within the uncertainty decoding framework presented in this paper. Also, we recognize that one potential drawback of increasing the model variances, as in the feature-space uncertainty decoding presented in this paper and in the BPC rule, is the possibility to increase model overlap and hence to decrease model discrimination. For the small vocabulary task of Aurora2 that we have worked on, such increased model overlap due to the HMM variance enlargement may not be a problem since the phonetic space of the recognized objects (digits) is relatively sparse. It remains our future research to examine the effectiveness of the proposed dynamic HMM variance compensation strategy for large vocabulary tasks and to improve such a strategy when the phonetic space becomes more crowded.

APPENDIX
DERIVATION OF (8)

We start with the additive noise model in the discrete-time (t) domain

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

where y, x, and n are (scalar) noisy speech, clean speech, and noise samples, respectively. Taking DFT on both sides, we have

$$Y[k] = X[k] + N[k] \tag{36}$$

where k is the frequency-bin index in DFT for a fixed-length time window. We then obtain the power spectra of the noisy speech from the DFT in (36)

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

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

A set of Mel-scale filters (L in total) are now applied to the 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 will produce L channel (Mel-filter bank) energies of

$$\sum_k W_k^{(l)}\,|Y[k]|^2 = \sum_k W_k^{(l)}\,|X[k]|^2 + \sum_k W_k^{(l)}\,|N[k]|^2 + 2\sum_k W_k^{(l)}\,|X[k]|\,|N[k]|\cos\theta_k \tag{37}$$

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

After denoting the various channel energies in (37) by

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

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

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

(37) can be simplified to

$$\left|\tilde{Y}^{(l)}\right|^2 = \left|\tilde{X}^{(l)}\right|^2 + \left|\tilde{N}^{(l)}\right|^2 + 2\lambda^{(l)}\left|\tilde{X}^{(l)}\right|\left|\tilde{N}^{(l)}\right| \tag{38}$$

where we define

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

We now define the log-channel energy vectors

$$\mathbf{y} = \begin{bmatrix} \log\left|\tilde{Y}^{(1)}\right|^2 \\ \log\left|\tilde{Y}^{(2)}\right|^2 \\ \cdots \\ \log\left|\tilde{Y}^{(l)}\right|^2 \\ \cdots \\ \log\left|\tilde{Y}^{(L)}\right|^2 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} \log\left|\tilde{X}^{(1)}\right|^2 \\ \log\left|\tilde{X}^{(2)}\right|^2 \\ \cdots \\ \log\left|\tilde{X}^{(l)}\right|^2 \\ \cdots \\ \log\left|\tilde{X}^{(L)}\right|^2 \end{bmatrix}, \quad \mathbf{n} = \begin{bmatrix} \log\left|\tilde{N}^{(1)}\right|^2 \\ \log\left|\tilde{N}^{(2)}\right|^2 \\ \cdots \\ \log\left|\tilde{N}^{(l)}\right|^2 \\ \cdots \\ \log\left|\tilde{N}^{(L)}\right|^2 \end{bmatrix} \tag{39}$$

and define the vector

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

After this, we rewrite (38) as

$$e^{\mathbf{y}} = e^{\mathbf{x}} + e^{\mathbf{n}} + 2\boldsymbol{\lambda} e^{\frac{\mathbf{x}}{2}} e^{\frac{\mathbf{n}}{2}} = e^{\mathbf{x}} + e^{\mathbf{n}} + 2\boldsymbol{\lambda} e^{\frac{\mathbf{x}+\mathbf{n}}{2}}. \tag{40}$$

Finally, we apply the log operation on both sides of (40) to obtain

$$\begin{aligned}\mathbf{y} &= \log\left[e^{\mathbf{x}}\left(\mathbf{1}+e^{\mathbf{n}-\mathbf{x}}+2\boldsymbol{\lambda}e^{\frac{\mathbf{x}+\mathbf{n}}{2}-\mathbf{x}}\right)\right] \\ &= \mathbf{x}+\log\left[\mathbf{1}+e^{\mathbf{n}-\mathbf{x}}+2\boldsymbol{\lambda}e^{\frac{\mathbf{n}-\mathbf{x}}{2}}\right] \end{aligned} \tag{41}$$

which directly leads to (8).

REFERENCES

[1] J. Arrowwood and M. Clements, "Using observation uncertainty in HMM decoding," in *Proc. ICSLP*, vol. III, Denver, CO, Sep. 2002, pp. 1561–1564.

[2] L. Deng, J. Droppo, and A. Acero, "Recursive noise estimation using iterative stochastic approximation for stereo-based robust speech recognition," in *Proc. ASRU Workshop*, Trento, Italy, Dec. 2001.

[3] ——, "A Bayesian approach to speech feature enhancement using the dynamic cepstral prior," in *Proc. ICASSP*, vol. I, Orlando, FL, May 2002, pp. 829–832.

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

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

[6] J. Droppo, A. Acero, and L. Deng, "Uncertainty decoding with SPLICE for noise robust speech recognition," in *Proc. ICASSP*, vol. I, Orlando, FL, May 2002, pp. 57–60.

[7] 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. Eurospeech*, 2001, pp. 901–904.

[8] Q. Huo and C. Lee, "A Bayesian predictive approach to robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 8, pp. 200–204, Nov. 2000.

[9] H. Jiang and L. Deng, "A robust compensation strategy against extraneous acoustic variations in spontaneous speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 1, pp. 9–17, Jan. 2002.

[10] T. Kristjansson and B. Frey, "Accounting for uncertainty in observations: A new paradigm for robust speech recognition," in *Proc. ICASSP*, vol. I, Orlando, FL, May 2002, pp. 61–64.

[11] C. Lee, C. Lin, and B. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. Signal Process.*, vol. 39, no. 4, pp. 806–814, Apr. 1991.

[12] A. Morris, J. Barker, and H. Bourlard, "From missing data to maybe useful data: soft data modeling for noise robust ASR," in *Proc. Workshop Innovation Speech Processing*, Stratford-upon-Avon, U.K., 2001.

[13] 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 Automatic Speech Recognition: Challenges for the Next Millennium*, Paris, France, Sep. 2000.

[14] M. Roch and R. Hurtig, "The integral decode: a smoothing technique for robust HMM-based speaker recognition," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 315–324, Jul. 2002.

[15] N. 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.

Li Deng (M'86–SM'91–F'04) received the B.S. degree from the University of Science and Technology of China, Beijing, in 1982, and the M.S. and Ph.D. degrees from the University of Wisconsin-Madison in 1984 and 1986, respectively.

He worked on large vocabulary automatic speech recognition in Montreal, QC, Canada, from 1986 to 1989. In 1989, he joined the Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada, as an Assistant Professor, where he became a tenured Full Professor in 1996. From 1992 to 1993, he conducted sabbatical research at the Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA, and from 1997 to 1998, at ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan. In 1999, he joined Microsoft Research, Redmond, WA, as a Senior Researcher and as an Affiliate Full Professor in electrical engineering at the 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 is inventor and co-inventor of numerous patents. He is principal author of *Speech Processing—A Dynamic and Optimization-Oriented Approach* (New York: Marcel Dekker, 2003).

Dr. Deng is a Fellow of the Acoustical Society of America. He served on Education Committee and Speech Processing Technical Committee, IEEE Signal Processing Society, from 1996 to 2000, and was Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, from 2002 to 2005. He currently serves on Multimedia Signal Processing Technical Committee. He was a Technical Chair of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP04).

Jasha Droppo (M'03) received the B.S. degree in electrical engineering (with honors) from Gonzaga University, Spokane, WA, in 1994, and the M.S. and Ph.D. degrees in electrical engineering from the University of Washington, Seattle, 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, 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 the B.S. degree from the Polytechnic University of Madrid, Madrid, Spain, in 1985, the 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 to 1991) and Manager of the Speech Technology Group, Telefonica Investigacion y Desarrollo (1991 to 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 Associate Editor of *Computer Speech and Language*. He is author of *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 speech recognition, synthesis and enhancement, speech denoising, language modeling, spoken language systems, statistical methods and machine learning, multimedia signal processing, and multimodal human-computer interaction.

Dr. Acero is an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS and of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He has held several positions within the IEEE Signal Processing Society, including Member-at-Large of the Board of Governors, and as Member (1996 to 2000) and Chair (2000 to 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.