Enhancement of Log Mel Power Spectra of Speech Using a Phase-Sensitive Model of the Acoustic Environment and Sequential Estimation of the Corrupting Noise

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

***Abstract*—This paper presents a novel speech feature enhancement technique based on a probabilistic, nonlinear acoustic environment model that effectively incorporates the phase relationship (hence phase sensitive) between the clean speech and the corrupting noise in the acoustic distortion process. The core of the enhancement algorithm is the MMSE (minimum mean square error) estimator for the log Mel power spectra of clean speech based on the phase-sensitive environment model, using highly efficient single-point, second-order Taylor series expansion to approximate the joint probability of clean and noisy speech modeled as a multivariate Gaussian. Since a noise estimate is required by the MMSE estimator, a high-quality, sequential noise estimation algorithm is also developed and presented. Both the noise estimation and speech feature enhancement algorithms are evaluated on the Aurora2 task of connected digit recognition. Noise-robust speech recognition results demonstrate that the new acoustic environment model which takes into account the relative phase in speech and noise mixing is superior to the earlier environment model which discards the phase under otherwise identical experimental conditions. The results also show that the sequential MAP (maximum a posteriori) learning for noise estimation is better than the sequential ML (maximum likelihood) learning, both evaluated under the identical phase-sensitive MMSE enhancement condition.**

***Index Terms*—Noise estimate, noise-robust ASR, phase-sensitive acoustic environment model, sequential algorithm, speech feature enhancement.**

I. Introduction

THIS paper addresses the problem of speech feature enhancement, and the associated problem of noise feature estimation, when the noisy speech features alone are available as the observational information. Enhancement of speech waveforms and features for improved auditory perception and for robust machine speech recognition has been an outstanding and difficult problem in speech processing for many years [1], [12], [13], [17], [25]. The problem is becoming increasingly important recently due to emerging commercial deployment of speech recognition technology which demands a high degree of noise robustness [24], [21]. Toward high-performance solutions to robust speech feature enhancement and accurate noise estimation, we recently developed a series of enhancement techniques capitalizing on the availability of stereo training data [5], [6], [11] or on a simple nonlinear model of the acoustic environment [20], [2], [8], [9], [14], [7]. The latter approach discards the phase relationship between the clean speech and the additive noise during the speech signal corruption process. The former approach is an end-to-end system, and due to its use of the stereo data, takes phase errors into consideration but only in an implicit manner. To overcome some weaknesses of these earlier techniques, such as the difficulty of acquiring well-matched stereo training data and the performance limit due to loss of the phase information, we have more recently developed a new feature enhancement technique which requires no stereo training data. This new technique explicitly exploits the novel concept of phase sensitivity in the acoustic environment model to derive the MMSE estimator for clean speech. In addition, in order to compute the MMSE estimator as an explicit function of the log-spectrum noise feature (and of the noisy speech feature), the technique employs a new sequential point estimator for nonstationary noise based on the MAP (maximum a posteriori) principle. Both of these aspects of the new technique form the core material of this paper.

Exploitation of the phase between speech and noise for speech enhancement has in the past been limited only to nonstatistical techniques, largely related to the framework known as nonlinear spectral subtraction [18], [25]. An insightful analysis was provided in [25], which links the SNR-dependent subtraction factor in nonlinear spectral subtraction to the missing phase information in the conventional linear spectral subtraction. A deterministic, empirical technique based on this analysis and on an approximate numerical solution was also proposed in [25]. The technique proposed in this paper which capitalizes on the same essential phase information, on the other hand, is established using a very different framework based on MMSE statistical estimation.[1] In addition, the phase-sensitive statistical model for the acoustic environment presented in this paper includes not only the additive noise case (as in [25]) but also simultaneously the case for convolutional distortion.

This paper is organized as follows. In Section II, we will describe the new phase-sensitive nonlinear model for the

Manuscript received September 10, 2002; revised August 28, 2003. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dirk van Compernolle.

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.2003.820201

[1]The MMSE estimation technique for speech enhancement was initially established in [12].

acoustic environment, which characterizes in statistical terms the acoustic distortion from clean speech features in the (Mel-scaled) log-spectrum domain to their noisy counterpart due to the additive noise corruption in the linear waveform domain. The MMSE estimator for noise removal based on this model will be derived in Section III. The novel MAP noise tracking algorithm will be presented in Section IV, which supplies an essential quantity required by the MMSE estimator. In Section V, we will provide experimental evidence on the Aurora2 task for the superiority of the phase-sensitive MMSE estimator and of the MAP noise tracker over their respective baselines.

II. Probabilistic, Phase-Sensitive Modeling for the Acoustic Environment

A. Relationship Among the Phase Factor and the Log-Spectra of Noise, Channel, Clean and Distorted Speech

Using the discrete-time, linear system model for the acoustic distortion in the time domain [1], [20], we have the well-known relationship among the noisy speech ($y(t)$), clean speech ($x(t)$), additive noise ($n(t)$), and the impulse response of the linear distortion channel ($h(t)$)

$$y(t) = x(t) * h(t) + n(t).$$

In the frequency domain, the equivalent relationship is

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

where k is the frequency-bin index in DFT given a fixed-length time window, and $H(k)$ is the (frequency-domain) transfer function of the linear channel.

The power spectrum of the noisy speech can then be obtained from the DFT in (1) by

$$\begin{aligned}|Y[k]|^2 =& |X[k]H[k] + N[k]|^2 \\ =& |X[k]|^2|H[k]|^2 + |N[k]|^2 + (X[k]H[k])(N[k])^* \\ &+ (X[k]H[k])^* N[k] \\ =& |X[k]|^2|H[k]|^2 + |N[k]|^2 + 2|X[k]||H[k]||N[k]|\cos\theta_k\end{aligned} \tag{2}$$

where θ_k denotes the (random) angle between the two complex variables $N[k]$ and $(X[k]H[k])$. Equation (2) incorporates the phase relationship between the (linearly filtered) clean speech and the additive corrupting noise in the speech distortion process, which will be shown to be important for improving the performance of speech feature enhancement. It is noted that in the traditional models for acoustic distortion [1], [20], [22], the last term in (2) has been assumed to be zero. This is correct only in expected sense. The phase-sensitive model presented in this paper based on (2) with nonzero instantaneous values in the last term removes such a commonly made but un-realistic assumption.

The new (2) leads to the following relationship among the phase factor $\boldsymbol{\alpha}$ (which is related to the angle between Mel-filter vectors of clean speech and noise, and is defined precisely in (28) in the Appendix) and the log Mel power spectra of noise $\mathbf{n}$, of channel $\mathbf{h}$, of clean speech $\mathbf{x}$, and of distorted speech $\mathbf{y}$

$$\begin{aligned}\mathbf{y} =& \mathbf{x} + \mathbf{h} + \log[\mathbf{1} + e^{\mathbf{n-x-h}} + 2\boldsymbol{\alpha} \bullet e^{(\mathbf{n-x-h})/2}] \\ \equiv& \mathbf{y}(\mathbf{x}, \mathbf{n}, \mathbf{h}, \boldsymbol{\alpha}).\end{aligned} \tag{3}$$

See a detailed derivation of (3) in the Appendix .

From (3), the phase factor (vector) $\boldsymbol{\alpha}$ can be solved as a function of the remaining variables

$$\begin{aligned}\boldsymbol{\alpha} =& \frac{e^{\mathbf{y-x-h}} - e^{\mathbf{n-x-h}} - \mathbf{1}}{2e^{(\mathbf{n-x-h})/2}} \\ =& 0.5(e^{\mathbf{y-(n+x+h)}/2} - e^{(\mathbf{n-x-h})/2} - e^{-(\mathbf{n-x-h})/2}) \\ \equiv& \boldsymbol{\alpha}(\mathbf{x}, \mathbf{n}, \mathbf{h}, \mathbf{y}).\end{aligned} \tag{4}$$

B. Probabilistic, Phase-Sensitive Modeling of the Acoustic Environment

We now use the nonlinear relationship among the phase factor $\boldsymbol{\alpha}$ and the log-domain signal quantities of $\mathbf{x}$, $\mathbf{n}$, $\mathbf{h}$, and $\mathbf{y}$, shown in (3) or (4), as the basis to develop a probabilistic phase-sensitive model for the acoustic environment. The outcome of a probabilistic model for the acoustic environment is explicit determination of the conditional probability, $p(\mathbf{y}|\mathbf{x}, \mathbf{n}, \mathbf{h})$, of the noisy speech observation ($\mathbf{y}$) given all other acoustic variables $\mathbf{x}$, $\mathbf{n}$, and $\mathbf{h}$. This conditional probability will be required for deriving an optimal estimate of clean speech, as will be presented in the next section.

To determine the form of $p(\mathbf{y}|\mathbf{x}, \mathbf{n}, \mathbf{h})$, we first need to assume a form of the statistical distribution for the phase factor $\boldsymbol{\alpha} = \{\alpha^{(l)}, l = 1, 2, \ldots, L\}$. To accomplish this, we first note that the angle θ_k between the complex variables of $N[k]$ and $(X[k]H[k])$ is uniformly distributed over $(-\pi, \pi)$. This amounts to the maximal degree of randomness in mixing speech and noise, and has been empirically observed to be correct.

Then, from the definition of $\alpha^{(l)}$ in (28) (Appendix), it can be shown that the phase factor $\alpha^{(l)}$ for each Mel-filter l can be approximated by a (weighted) sum of a number of independent, zero-mean random variables distributed (nonuniformly) over $(-1, 1)$, where the total number of terms equals the number of DFT bins (with a nonzero gain) allocated to the Mel-filter. When the number of terms becomes large, as is typical for high-frequency filters, the central limit theorem postulates that $\alpha^{(l)}$ will be approximately Gaussian. Law of large numbers further postulates that the Gaussian will have zero mean since each term of $\cos(\theta_k)$ has a zero mean.

Thus, the statistical distribution for the phase factor can be reasonably assumed to be a zero-mean Gaussian

$$p(\alpha^{(l)}) = \mathcal{N}(\alpha^{(l)}; 0, \Sigma_\alpha^{(l)})$$

where the filter-dependent variance $\Sigma_\alpha^{(l)}$ is estimated from a set of training data (see details in Section V-D). Since noise and (channel-distorted) clean speech are mixed independently for each DFT bin, we can also reasonably assume that the different components of the phase factor $\boldsymbol{\alpha}$ are uncorrelated. Thus, we have the multivariate Gaussian distribution of

$$p(\boldsymbol{\alpha}) = \mathcal{N}(\boldsymbol{\alpha}; \mathbf{0}, \boldsymbol{\Sigma}_\alpha) \tag{5}$$

where $\boldsymbol{\Sigma}_\alpha$ is a diagonal covariance matrix.

Given $p(\boldsymbol{\alpha})$, we are now in a position to derive an appropriate form for $p_y(\mathbf{y}|\mathbf{x},\mathbf{n},\mathbf{h})$. To do so, we first fix the values of $\mathbf{x}$, $\mathbf{n}$, and $\mathbf{h}$, treating them as constants. We then view (3) as a (monotonic) nonlinear transformation from random variables $\boldsymbol{\alpha}$ to $\mathbf{y}$. Using the well-known result from probability theory on determining the PDF for functions of random variables, we have

$$p_y(\mathbf{y}|\mathbf{x},\mathbf{n},\mathbf{h}) = |J_{\boldsymbol{\alpha}}(\mathbf{y})|\, p_\alpha(\boldsymbol{\alpha}|\mathbf{x},\mathbf{n},\mathbf{h}) \tag{6}$$

where $J_{\boldsymbol{\alpha}}(\mathbf{y}) = 1/(\partial\mathbf{y}/\partial\boldsymbol{\alpha})$ is the Jacobian[2] of the nonlinear transformation.

The diagonal elements of the Jacobian can be computed, using (3) and then using (30) (see the Appendix), by

$$\begin{aligned}\operatorname{diag}\left(\frac{\partial\mathbf{y}}{\partial\boldsymbol{\alpha}}\right) &= \frac{2e^{(\mathbf{n}-\mathbf{x}-\mathbf{h})/2}}{\mathbf{1}+e^{\mathbf{n}-\mathbf{x}-\mathbf{h}}+2\boldsymbol{\alpha}\bullet e^{(\mathbf{n}-\mathbf{x}-\mathbf{h})/2}} \\ &= \frac{2e^{(\mathbf{n}+\mathbf{x}+\mathbf{h})/2}}{e^{\mathbf{x}+\mathbf{h}}+e^{\mathbf{n}}+2\boldsymbol{\alpha}\bullet e^{(\mathbf{n}+\mathbf{x}+\mathbf{h})/2}} = 2\, e^{(\mathbf{n}+\mathbf{x}+\mathbf{h})/2-\mathbf{y}}.\end{aligned} \tag{7}$$

The determinant of the diagonal matrix of (7) is then the product of all the diagonal elements. Also, the Gaussian assumption for $\boldsymbol{\alpha}$ gives

$$p(\boldsymbol{\alpha}|\mathbf{x},\mathbf{n},\mathbf{h}) = p\left[\boldsymbol{\alpha}(\mathbf{x},\mathbf{n},\mathbf{h},\mathbf{y})\right] = \mathcal{N}\left[\boldsymbol{\alpha}(\mathbf{x},\mathbf{n},\mathbf{h},\mathbf{y});\mathbf{0},\boldsymbol{\Sigma}_\alpha\right]. \tag{8}$$

Substituting (7) and (8) into (6), we establish the following probabilistic model of the acoustic environment:

$$\begin{aligned}p_y(\mathbf{y}|\mathbf{x},\mathbf{n},\mathbf{h}) = &\frac{1}{2}\left|\operatorname{diag}\left(e^{\mathbf{y}-(\mathbf{n}+\mathbf{x}+\mathbf{h})/2}\right)\right| \\ &\mathcal{N}\left[\frac{1}{2}(e^{\mathbf{y}-(\mathbf{n}+\mathbf{x}+\mathbf{h})/2}-e^{(\mathbf{n}-\mathbf{x}-\mathbf{h})/2}-e^{-(\mathbf{n}-\mathbf{x}-\mathbf{h})/2});\mathbf{0},\boldsymbol{\Sigma}_\alpha\right].\end{aligned} \tag{9}$$

Because $\boldsymbol{\alpha}$ is the inner product (proportional to cosine of the phase) between the Mel-filter vectors of noise and clean speech characterizing their phase relationship, a Gaussian distribution on it makes the environment model of (9) phase sensitive.[3]

For exposition simplicity, in the remaining of this paper we assume: 1) The log-domain noise vector $\mathbf{n} = \bar{\mathbf{n}}$ is deterministic, or $p(\mathbf{n}) = \delta(\mathbf{n}-\bar{\mathbf{n}})$ ($\bar{\mathbf{n}}$ can be obtained by point estimators as will be described in Section IV); and 2) The channel distortion can be ignored: $\mathbf{h} = \mathbf{0}$. Further, since the covariance matrix $\boldsymbol{\Sigma}_\alpha$ is assumed to be diagonal with nonzero elements denoted by $\Sigma_\alpha^{(l)}$, we will present the scalar rather than vector derivation, without loss of generality, for speech feature enhancement next.

III. MMSE Log Power Spectral Estimator of Clean Speech

A. Algorithm and its Derivation

Given the log (Mel) power spectra of the noisy speech observation y, the MMSE estimator $\hat{x}$ for clean speech x is the conditional expectation

$$\hat{x} = E[x|y] = \int x p(x|y)dx = \frac{\int x p_{\bar{n}}(y|x)p(x)dx}{p(y)} \tag{10}$$

where $p_{\bar{n}}(y|x) = p(y|x,\bar{n})$ is determined by the probabilistic environment model just presented. The prior model for clean speech, $p(x)$, in (10) is assumed to have the Gaussian mixture PDF

$$p(x) = \sum_{m=1}^{M} c_m \underbrace{\mathcal{N}(x;\mu_m,\sigma_m^2)}_{p(x|m)} \tag{11}$$

whose parameters are pre-trained from the log-spectral clean speech data. This allows us to write (10) as

$$\hat{x} = \frac{\sum_{m=1}^{M} c_m \int x \overbrace{p(x|m)p_y(y|x,\bar{n})}^{K_m(x,\bar{n},y)} dx}{p(y)}. \tag{12}$$

The main difficulty in computing $\hat{x}$ above is the non-Gaussian nature of $p(y|x,\bar{n})$ in (9). To overcome this difficulty, we use the truncated second-order Taylor series expansion to approximate the exponent of

$$\begin{aligned}K_m(x,\bar{n},y) &= \mathcal{N}(x;\mu_m,\sigma_m^2)\times\frac{\mathcal{N}(\alpha(x,\bar{n},y);0,\Sigma_\alpha)}{2\,e^{(\bar{n}+x)/2-y}} \\ &= \frac{C}{\sigma_m}e^{-0.5(x-\mu_m)^2/\sigma_m^2-0.5x-0.5\alpha^2(x,\bar{n},y)/\Sigma_\alpha}\end{aligned}$$

where C is independent of the mixture component m.[4] That is, we approximate the function

$$\begin{aligned}b_m(x,\bar{n},y) &= -\frac{0.5(x-\mu_m)^2}{\sigma_m^2}-0.5x-\frac{0.5\,\alpha^2(x,\bar{n},y)}{\Sigma_\alpha} \\ &= -\frac{0.5(x-\mu_m)^2}{\sigma_m^2}-0.5x-\frac{(e^y-e^{\bar{n}}-e^x)^2}{8\,e^{\bar{n}+x}\,\Sigma_\alpha}\end{aligned} \tag{13}$$

by

$$\begin{aligned}b_m(x,\bar{n},y) \approx\, & b_m^{(0)}(x_0,\bar{n},y)+b_m^{(1)}(x_0,\bar{n},y)(x-x_0) \\ &+\frac{b_m^{(2)}(x_0,\bar{n},y)}{2}(x-x_0)^2.\end{aligned} \tag{14}$$

In (14), we used a single expansion point x_0 (i.e., x_0 does not depend on the mixture component m) to have significantly improved computational efficiency, and x_0 is iteratively updated to increase its accuracy to the true value of clean speech x. The Taylor series expansion coefficients have the following closed forms:

$$\begin{aligned}b_m^{(0)}(x_0,\bar{n},y) &= b_m(x,\bar{n},y)\,|_{x=x_0} \\ &= -\frac{(x_0-\mu_m)^2}{2\sigma_m^2}-\frac{x_0}{2}-\frac{(e^y-e^{\bar{n}}-e^{x_0})^2}{8\,e^{\bar{n}+x_0}\,\Sigma_\alpha}, \\ b_m^{(1)}(x_0,\bar{n},y) &= \frac{\partial b_m(x,\bar{n},y)}{\partial x}\,|_{x=x_0} \\ &= -\frac{x_0-\mu_m}{\sigma_m^2}-\frac{1}{2}+\frac{e^{2y-\bar{n}-x_0}-2e^{y-x_0}+e^{\bar{n}-x_0}-e^{x_0-\bar{n}}}{8\Sigma_\alpha}, \\ b_m^{(2)}(x_0,\bar{n},y) &= \frac{\partial^2 b_m(x,\bar{n},y)}{\partial^2 x}\,|_{x=x_0} \\ &= -\frac{1}{\sigma_m^2}+\frac{-e^{2y-\bar{n}-x_0}+2e^{y-x_0}-e^{\bar{n}-x_0}-e^{x_0-\bar{n}}}{8\Sigma_\alpha}.\end{aligned}$$

[2]It can be easily shown that this Jacobian matrix is a diagonal one.

[3]This contrasts our earlier model in [7] where an entire term of $\boldsymbol{\alpha}/\cosh((\mathbf{n}-\mathbf{x}-\mathbf{h})/2)$ was assumed to be a zero-mean Gaussian. Hence, the phase information has been seriously smeared due to elimination of the explicit dependency of the variances on the instantaneous SNR.

[4]C is a function of $\bar{n}$ and y, but will be cancelled out by the same quantity in the denominator to be discussed shortly.

Fitting (14) into a standard quadratic form, we obtain

$$b_m(x,\bar{n},y) \approx \frac{b_m^{(2)}(x_0,\bar{n},y)}{2}\left[x-\left(x_0-\frac{b_m^{(1)}(x_0,\bar{n},y)}{b_m^{(2)}(x_0,\bar{n},y)}\right)\right]^2 + w_m(x_0,\bar{n},y)$$

where

$$\begin{aligned} & w_m(x_0,\bar{n},y) \\ &= b_m^{(0)}(x_0,\bar{n},y) + \frac{b_m^{(2)}(x_0,\bar{n},y)}{2} \\ &\quad \cdot\left[x_0^2 - \frac{2b_m^{(1)}(x_0,\bar{n},y)}{b_m^{(2)}(x_0,\bar{n},y)}x_0 - \left(x_0-\frac{b_m^{(1)}(x_0,\bar{n},y)}{b_m^{(2)}(x_0,\bar{n},y)}\right)^2\right] \\ &= b_m^{(0)}(x_0,\bar{n},y) - \frac{[b_m^{(1)}(x_0,\bar{n},y)]^2}{2b_m^{(2)}(x_0,\bar{n},y)}. \end{aligned}$$

This then allows us to compute the integral of (12) in a closed form

$$\begin{aligned} & I_m(x_0,\bar{n},y) \\ &= \int xK_m(x,\bar{n},y)dx = \frac{C}{\sigma_m}\int xe^{b_m(x,\bar{n},y)}dx \\ &\approx \frac{C}{\sigma_m} \\ &\quad \cdot\int xe^{(b_m^{(2)}(x_0,\bar{n},y))/2\left[x-(x_0-(b_m^{(1)}(x_0,\bar{n},y)/b_m^{(2)}(x_0,\bar{n},y)))\right]^2+w_m(x_0,\bar{n},y)}dx \\ &= \frac{C}{\sigma_m}\, e^{w_m(x_0,\bar{n},y)} \\ &\quad \cdot\int xe^{(b_m^{(2)}(x_0,\bar{n},y))/2\left[x-(x_0-(b_m^{(1)}(x_0,\bar{n},y)/b_m^{(2)}(x_0,\bar{n},y)))\right]^2}dx \\ &= \frac{\sqrt{2\pi}C}{\sigma_m\sqrt{-b_m^{(2)}(x_0,\bar{n},y)}}\, e^{w_m(x_0,\bar{n},y)} \\ &\quad \cdot\int x\mathcal{N}\left(x;x_0-\frac{b_m^{(1)}(x_0,\bar{n},y)}{b_m^{(2)}(x_0,\bar{n},y)},\frac{-1}{b_m^{(2)}(x_0,\bar{n},y)}\right)dx \\ &= \frac{\sqrt{2\pi}C}{\sigma_m\sqrt{-b_m^{(2)}(x_0,\bar{n},y)}}\, e^{w_m(x_0,\bar{n},y)}\times\left(x_0-\frac{b_m^{(1)}(x_0,\bar{n},y)}{b_m^{(2)}(x_0,\bar{n},y)}\right). \end{aligned} \tag{15}$$

The last step above used the fact that the integral in the preceding step is the mean of the normal distribution.

The denominator of (12) is computed according to

$$\begin{aligned} p(y) &= \sum_{m=1}^{M} c_m\int K_m(x,\bar{n},y)dx = \sum_{m=1}^{M} c_m\frac{C}{\sigma_m}\int e^{b_m(x,\bar{n},y)}dx \\ &\approx \sum_{m=1}^{M} c_m\frac{\sqrt{2\pi}C}{\sigma_m\sqrt{-b_m^{(2)}(x_0,\bar{n},y)}}e^{w_m(x_0,\bar{n},y)} \\ &\quad \cdot\int\mathcal{N}\left(x;x_0-\frac{b_m^{(1)}(x_0,\bar{n},y)}{b_m^{(2)}(x_0,\bar{n},y)},\frac{-1}{b_m^{(2)}(x_0,\bar{n},y)}\right)dx \\ &= \sum_{m=1}^{M} c_m\frac{\sqrt{2\pi}C}{\sigma_m\sqrt{-b_m^{(2)}(x_0,\bar{n},y)}}e^{w_m(x_0,\bar{n},y)} \end{aligned} \tag{16}$$

where the integration over the normal distribution becomes one.

Substituting (15) and (16) into (12), we obtain the final MMSE estimator

$$\hat{x} \approx \sum_{m=1}^{M}\gamma_m(x_0,\bar{n},y)\left(x_0-\frac{b_m^{(1)}(x_0,\bar{n},y)}{b_m^{(2)}(x_0,\bar{n},y)}\right) \tag{17}$$

where the weighting factors are

$$\gamma_m(x_0,\bar{n},y) = \frac{\frac{c_m}{\sigma_m\sqrt{-b_m^{(2)}(x_0,\bar{n},y)}}e^{w_m(x_0,\bar{n},y)}}{\sum_{m=1}^{M}\frac{c_m}{\sigma_m\sqrt{-b_m^{(2)}(x_0,\bar{n},y)}}\, e^{w_m(x_0,\bar{n},y)}}.$$

Note that $\gamma_m(x_0,\bar{n},y)$, $b_m^{(1)}(x_0,\bar{n},y)$, and $b_m^{(2)}(x_0,\bar{n},y)$ in (17) are all dependent on the noise estimate $\bar{n}$.

In applying the MMSE estimator (17) to perform speech feature enhancement, we first use the result of another enhancement algorithm (published in [7]) to initialize x_0 at the right hand side of (17). (In the MMSE estimator implementation, we have explored various ways of initializing x_0 and found empirically that the performance of the estimator is rather sensitive to the initial value of x_0. The output of the enhancement system in [7] (where the joint static and dynamic prior of clean speech was used) gives a reasonably good quality of x_0 for initializing the current MMSE estimator). The estimated clean speech $\hat{x}$ is then used to update x_0 and the iteration continues until a fixed number of iterations is reached or convergence occurs. The use of the iteration is to overcome approximation errors introduced by truncated Taylor series expansion of (14). This iterative technique, commonly used in nonlinear signal processing [19], was previously successfully applied to speech enhancement in [14] and in spontaneous speech recognition in [10].

In the current implementation of the iterative MMSE estimation algorithm based on (17), the noisy-speech frames are processed independently of each other. That is, no dynamic constraints on the speech properties have been exploited. A fixed number of iterations are completed for one noisy-speech frame before starting new iterations to process the next noisy-speech frame.

B. Computational Analysis

In this subsection, we provide a brief analysis on the computational and memory requirements for the iterative MMSE speech feature enhancement algorithm just presented. The estimation formula of (17) describes the computation for each frame of noisy speech and for each algorithm iteration. Let T be the total number of frames and R be the fixed number of iterations. Then the computational load is proportional to $T\times R\times M$, where the proportion constant is related to the remaining quantities in (17) that are all computed rapidly in closed forms. The initial Taylor series expansion point x_0 in (17) is computed using a phase-insensitive enhancement algorithm described in [7]. It has the computational complexity about the same as the computation in (17) excluding the computations for the initial expansion point x_0 and for the sequential noise estimate of $\bar{n}$. Determining this noise estimate (described in the next section) also has a similar computational complexity. Therefore, considering computations for both x_0 and $\bar{n}$, the full computation for (17) becomes roughly three times of that excluding x_0 and $\bar{n}$.

Since the algorithm is executed frame by frame for each separate iterations, the memory requirement for the algorithm becomes proportional to M only, instead of to $T \times R \times M$ as for the computational requirement.

In our experiments described in Section V, the number of mixture components M for the clean speech model is set to be 256, and the number of algorithm iterations R ranges from one to 12. This choice made the overall algorithm (coded in Matlab 6.1) run about 10 to 100 times of real time in a Pentium III 547-MHz machine.

IV. Sequential Estimation of Nonstationary Noise

In this section, we present sequential trackers for estimating the log spectrum of nonstationary noise $\bar{n}$, which is needed in computing quantities $\gamma_m(x_0, \bar{n}, y)$, $b_m^{(1)}(x_0, \bar{n}, y)$, and $b_m^{(2)}(x_0, \bar{n}, y)$ in the iterative MMSE estimation of clean speech according to (17). This algorithm is generalized from the earlier ML estimator within the same recursive-EM framework presented in [8] and [9], based on a relatively simple phase-insensitive acoustic distortion model presented also in [8] and [9].

A. E-Step

In the E-step, we compute the MAP auxiliary function [4] of

$$Q_{MAP}(n_t) = Q_{ML}(n_t) + \rho \log p(n_t) \tag{18}$$

where $p(n_t)$ is the fixed prior distribution of Gaussian for noise n_t, ρ is the variance scaling factor, and the ML auxiliary function [4] is the following conditional expectation:

$$\begin{aligned} Q_{ML}(n_t) &= E\left[\log p(y_1^t, \mathcal{M}_1^t | n_t) | y_1^t, n_1^{t-1}\right] \\ &= \sum_{\tau=1}^{t} \sum_{m=1}^{M} \xi_\tau(m) \log\ p(y_\tau \mid m, n_t). \end{aligned} \tag{19}$$

In (19), $\mathcal{M}_1^t = m_1, m_2, \ldots, m_t$ is the sequence of (hidden) mixture components in the clean speech model (11) up to time frame t, and similarly we have the observation sequence of $y_1^t = y_1, y_2, \ldots, y_t$. The expectation in (19) is carried out with respect to the conditional distribution $p(\mathcal{M}_1^t | y_1^t, n_1^{t-1})$. Note that the objective function of (19) in the current sequential EM algorithm differs from the one in the conventional batch-EM in that $Q(n_t)$ in (19) is time indexed and the observation sequence is used up to that time as denoted by y_1^t.

After introducing the forgetting factor ϵ, $Q_{ML}(n_t)$ is modified to

$$\begin{aligned} Q_{ML}(n_t) &= \sum_{\tau=1}^{t} \epsilon^{t-\tau} \sum_{m=1}^{M} \xi_\tau(m) \log\ p(y_\tau \mid m, n_t) \\ &= -\sum_{\tau=1}^{t} \epsilon^{t-\tau} \sum_{m=1}^{M} \xi_\tau(m) \frac{(y_\tau - \mu_m^y)^2}{2\,\Sigma_m^y} + Const. \end{aligned} \tag{20}$$

In (20), the forgetting factor ϵ controls the balance between the ability of the algorithm to track noise nonstationarity and the reliability of the noise estimate,[5] and $\xi_\tau(m) = p(m | y_\tau, n_{\tau-1})$ is the posterior probability for the hidden mixture component. The posterior probability is computed using Bayes rule

$$\xi_\tau(m) = \frac{c_m p(y_\tau | m, n_{\tau-1})}{\sum_m c_m p(y_\tau | m, n_{\tau-1})}$$

where likelihood $p(y_\tau | m, n_{\tau-1})$ is approximated by a Gaussian with the mean and variance of

$$\begin{aligned} \mu_m^y &\approx \mu_m^x + g + [1 - G](n_t - n_0) \\ \Sigma_m^y &\approx (1+G)^2 \Sigma_m^x + (1-G)^2 \Sigma^n. \end{aligned} \tag{21}$$

Here, Σ^n is the fixed variance (hyper-parameter) of the prior noise PDF $p(n_t)$, which is assumed to be Gaussian (with the fixed hyper-parameter mean of μ_n). g and G in (21) are computable quantities introduced in [8], [9] to linearly approximate the relationship among noisy speech y, clean speech x, and noise n (all in the form of log Mel power spectra). The expressions for these quantities are

$$g = \log[1 + \exp(n_0 - \hat{x})]$$

and

$$G = \frac{1}{1 + \exp(n_0 - \hat{x})}$$

respectively, where $\hat{x}$ is an estimate for clean speech, implemented in this work as the best matched Gaussian mean of the mixture-of-Gaussian clean speech model described in Section III, and n_0 is the Taylor series expansion point for noise, which is iteratively updated by the MAP estimate in the M-step described below.

B. M-Step

In the M-step, we estimate n_t by setting

$$\frac{\partial Q_{MAP}(n_t)}{\partial n_t} = 0.$$

Noting from (21) that μ_m^y is a linear function of n_t, we have

$$\sum_{\tau=1}^{t} \epsilon^{t-\tau} \sum_{m=1}^{M} \xi_\tau(m) \frac{(y_\tau - \mu_m^y)}{\Sigma_m^y}(1 - G_m) - \frac{\rho(n_t - \mu_n)}{\Sigma^n} = 0. \tag{22}$$

Substituting (21) into (22) and solving for n_t, we obtain the MAP estimate of noise

$$\hat{n}_t = \frac{s_t + \frac{\rho \mu_n}{\Sigma^n} + K_t n_0}{K_t + \frac{\rho}{\Sigma^n}} \tag{23}$$

where

$$s_t = \sum_{\tau=1}^{t} \epsilon^{t-\tau} \sum_{m=1}^{M} \xi_\tau(m)(y_\tau - \mu_m^x - g_m) \frac{(1 - G_m)}{\Sigma_m^y}$$

and

$$K_t = \sum_{\tau=1}^{t} \epsilon^{t-\tau} \sum_{m=1}^{M} \xi_\tau(m) \frac{(1 - G_m)^2}{\Sigma_m^y}.$$

[5] In one of the extremes when $\epsilon = 0$, the algorithm would, with low estimation reliability, closely track fast temporal changes of noise since only the current frame is used for noise estimation. When $\epsilon = 1$, on the other hand, all previous frames would be used for estimation (with an equal weight), increasing the estimation reliability and sacrificing the fast-tracking capability.

The s_t and K_t above can be efficiently computed by making use of previous computation for s_{t-1} and K_{t-1} via recursion, as in our earlier work for recursive ML noise estimation [8], [9], based on the original proposal from [23] and on recent work published in [3], [16]. For example, efficient recursive computation for K_t as we have implemented in this work is

$$K_t = \epsilon K_{t-1} + \sum_{m=1}^{M} \xi_t(m) \frac{(1-G_m)^2}{\Sigma_m^y}.$$

Note that the MAP estimate of (23) reverts to the ML estimate derived in [8], [9], as expected, when ρ is set to zero or when the variance of the noise prior distribution goes to infinity. In either of these extreme cases, the prior distribution of the noise would be expected to provide no information as far as noise estimation is concerned.

V. Noise-Robust Speech Recognition Experiments

The MMSE estimator for clean speech features and the sequential MAP noise estimate described so far in this paper have been evaluated on the Aurora2 database, using the standard recognition tasks designed for this database [15], [11]. 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 HMMs used in our evaluation experiments are specified by the Aurora2 task and trained using the clean-speech training set.

A. Results Using Phase-Removed Vectors of True Noise

In this set of experiments, we use the MFCC's and their inverse cosine transform computed from true noise (available in the Aurora2 database) as the deterministic noise $\bar{n}$ in (17) to evaluate the effects of various factors on the MMSE estimator's performance for noise-robust speech recognition. Other objectives of these experiments are to set the upper limit for the possible performance, and to demonstrate the effectiveness of incorporating the phase information in the speech distortion process.

Table I shows percent accuracy results on the full set of Aurora2 test data, when clean-speech HMMs are used, as a function of the number of iterations (R) for the MMSE estimator of (17). The initial clean-speech estimate, used as x_0 in (17) before any iteration in applying the MMSE estimator, is obtained from the algorithm published in [7] that has largely discarded the phase information in the speech corruption process. This forms the baseline, against which the phase-sensitive MMSE estimator is evaluated. The percent-accuracy performance of the baseline is 84.80% averaged over Sets A, B, and C.

When the MMSE estimator of (17) is applied iteratively to update the initial estimate, dramatic performance improvement is observed consistently across all three data sets. Performance convergence occurs at around seven iterations. In Table II, we list details of recognition rates (%) for each of the four noise conditions and for each of the SNR's in Set-A at the convergence. The same results for Set-B and Set-C are presented in Tables III and IV, respectively, with different noise types and distortion conditions.

In Fig. 1, we plot the convergence curve for example speech frames in the SetA test data when applying the MMSE estimator of (17) iteratively. The 13 curves correspond to the 13 MFCCs from the zeroth order to the 12th order. These MFCCs are computed from the estimated log Mel power spectra of a 23 dimension using the cosine transform. Convergence of the estimated quantities, at roughly the seventh iteration as illustrated in Fig. 1, has been observed for many utterances that we have examined.

TABLE I
Effects of the Total Number of Iterations (R) on the MMSE Estimator's Performance (Percent Accurate Digit Recognition Rate) for the Aurora2 Task. Phase-Removed MFCC Vectors of True Noise are Used for $\bar{n}$ in (17). The Baseline Performance is 84.80% Averaged Over the Three Sets

R	1	2	4	7	12
SetA	94.12	96.75	97.96	98.11	98.12
SetB	94.80	97.29	98.10	98.48	98.55
SetC	91.00	94.50	96.50	97.86	98.00
Ave.	93.77	96.52	97.72	98.21	98.27

TABLE II
Detailed Recognition Rates (Percent Accurate) Using the Phase-Sensitive MMSE Estimator After the Seventh Iteration. Four Noise Conditions: Subway, Babble, Car, Exhibition-Hall Noises; SNRs From 0 DB to 20 DB in 5-DB Increment; Set-A Results With Clean Speech Training

SNR	Subway	Babble	Car	Exhibition	Average
20 dB	98.74	98.97	98.87	99.11	98.92
15 dB	98.71	98.94	98.96	98.92	98.88
10 dB	98.28	98.85	98.69	98.43	98.56
5 dB	97.61	98.19	98.30	97.50	97.90
0 dB	95.55	97.31	97.08	95.28	96.31
Ave.	97.78	98.45	98.38	97.85	98.11

TABLE III
Detailed Digit Recognition Rates (Percent Accurate) Using the Phase-Sensitive MMSE Estimator After the Seventh Iteration. 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	98.99	98.94	98.96	99.14	99.01
15 dB	98.93	98.62	99.22	98.63	98.85
10 dB	98.68	98.52	98.78	98.86	98.71
5 dB	98.25	98.31	98.48	98.40	98.36
0 dB	97.70	96.43	98.39	97.38	97.48
Ave.	98.51	98.16	98.77	98.48	98.48

TABLE IV
DETAILED RECOGNITION RATES (PERCENT ACCURATE) USING THE PHASE-SENSITIVE MMSE ESTIMATOR AFTER THE SEVENTH ITERATION. 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	99.02	99.09	99.06
15 dB	98.89	98.79	98.84
10 dB	98.37	98.49	98.43
5 dB	97.33	97.91	97.62
0 dB	95.27	95.44	95.36
Ave.	97.78	97.94	97.86

This accounts for virtually no accuracy improvement after the seventh iteration in Table I.

B. Comparisons With Spectral Subtraction

To further demonstrate the benefits of the MMSE estimator of (17) in modeling the phase information, we use the same phase-removed true noise for (phase-insensitive) spectral subtraction (SS) and perform the identical Aurora2 evaluation. The SS algorithm is obtained by setting $\boldsymbol{\alpha} = 0$ in (3) (as well as $\mathbf{h} = 0$), which gives

$$\hat{\mathbf{x}} = \log(e^{\mathbf{y}} - e^{\mathbf{n}}) = \mathbf{y} + \log(1 - e^{\mathbf{n}-\mathbf{y}}).$$

To avoid the possibility of taking logarithm of negative values (when $\mathbf{n} > \mathbf{y}$ due to statistical variation arising from the random phase in mixing speech and noise), we introduce the floor parameter F according to

$$\hat{\mathbf{x}} = \mathbf{y} + \log\left[\max(1 - e^{\mathbf{n}-\mathbf{y}}, F)\right], \text{ or} \tag{24}$$

$$\hat{\mathbf{x}} = \mathbf{y} + \log\left[\max(|1 - e^{\mathbf{n}-\mathbf{y}}|, F)\right]. \tag{25}$$

These two ways of using the floor, in combination of applying the SS in the domains of direct Mel-scaled log spectra and of MFCC's as smoothed log spectra, result in four versions of the SS algorithm. Their respective digit recognition accuracies (%) as a function of the floor level are listed in Table V for Set A of the Aurora2 test data, where phase-removed, Mel-scaled log spectra (SS1 and SS2) or MFCCs (SS3 and SS4) are computed from true noise waveforms. SS1 and SS3 make use of (25). SS2 and SS4 make use of (24). Note that the best accuracy, 95.9%, still contains 54% more errors than that achieved by the converged MMSE estimator (98.1% accuracy), which models $\boldsymbol{\alpha}$ by a zero-mean Gaussian distribution rather than setting it to zero.

The results shown in Table V are somewhat surprising and against the conventional wisdom. The conventional wisdom holds that the main deficiency of spectral subtraction arises from inaccuracy in frame-specific spectral estimation of the noise. Now the surprising results of Table V demonstrate that even when the exact noise spectra are provided, spectral subtraction does not produce the exact spectra of clean speech. Rather, the spectra produced from such spectral subtraction are still quite inadequate for high-performance speech recognition. This suggests that the phase-sensitive term in (27), $2\alpha^{(l)}|\tilde{X}^{(l)}||\tilde{H}^{(l)}||\tilde{N}^{(l)}|$, which has been ignored in the conventional spectral subtraction accounts for its inadequacy. Effective exploitation of this term, as in the phase-sensitive model of the acoustic environment [(9)] and in the associated phase-sensitive MMSE estimator [(17)], has significantly reduced the recognition error rates shown in Table V to those in Table I, successfully overcoming the inadequacy of the conventional spectral subtraction.

C. Results Using Automatic Noise Estimates

In contrast to using the true noise log power spectral vector as $\bar{n}$ in (17) when applying the MMSE estimator to speech feature enhancement just described, in this section are presented the results using the estimated noise vectors. The best technique we have developed so far is the sequential MAP noise estimator described in Section IV, where the prior distribution of the noise is assumed to be diagonal Gaussian. In the current implementation and in the evaluation on the Aurora2 task, the mean and variance of the Gaussian change from utterance to utterance in the test data. They are fixed to be the sample mean and sample variance of the first 20 frames in each separate test utterance, which are assumed to be free of any speech material.

Applying the MAP noise estimator to the MMSE estimator (one iteration) for clean speech, we obtain the percent-accuracy performance results for all three sets of the Aurora2 test data. The results are shown in the last column of Table VI, using $\hat{x}$ in (17) (with MAP-tracked noise as $\bar{n}$) to score the pre-trained clean-speech HMMs.[6] This gives improvement over the baseline performance (established in the work of [7] using a phase-insensitive MMSE estimator and shown in Column 2 in Table VI), where the initial clean speech vector x_0 in (17) (i.e., without using the MMSE estimator) is used to score the HMMs. Compared with the performance shown in Column 3 of Table VI, the MAP-tracked noise (as described in Section IV) also provides slight improvement over the use of the sequential maximum likelihood (ML) noise estimator in the otherwise identical experimental setup (i.e., using $\hat{x}$ in (17) with the ML-tracked noise as $\bar{n}$). The algorithm for computing the ML-tracked noise estimator can be found in [8], [9], which gave the state-of-the-art performance in our earlier noise-robust recognition system [11].

The results shown in Fig. 1 and Table I indicate that iterations of the speech feature enhancement algorithm up to about six reached a level of convergence, where an optimal recognition performance was achieved when true noise features were used. This, however, did not happen when the estimated noise features were used. In our experiments, with one iteration, the performance was always improved as shown in Table VI. More iterations either degraded or improved the performance in an unpredictable manner. The cause of this behavior is currently under investigation. Preliminary analysis illustrates that the quality of the noise estimate appears to be responsible for our empirical observation. This can be understood from

[6]Based on all of our earlier noise-robust speech recognition work on the Aurora2 and other tasks [11], [5], any performance improvement achieved by new feature enhancement techniques carries over from the clean-speech HMM training case to the multi-style training case. In this paper, we will present the results on the clean-speech HMM training case only.

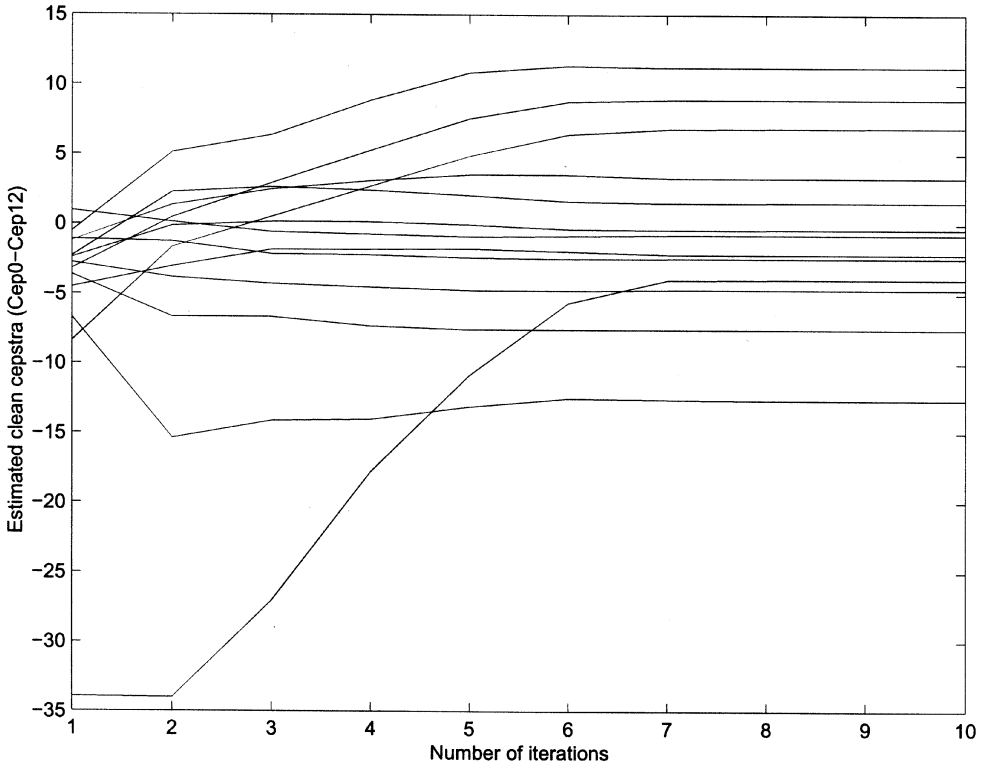

Fig. 1. Convergence curve showing the estimated clean speech MFCC's (cep0–12) using (17) as a function of the iteration number. The 13-dimensional MFCCs are computed from the estimated log Mel power spectra of 23 dimensions via cosine transformation.

TABLE V
PERFORMANCE (PERCENT ACCURATE) FOR THE AURORA2 TASK (SET-A ONLY) USING FOUR VERSIONS OF SPECTRAL SUBTRACTION (SS)

Floor	e^{-20}	e^{-10}	e^{-5}	e^{-3}	e^{-2}
SS1	93.57	94.26	95.90	92.18	90.00
SS2	12.50	44.00	65.46	88.69	84.44
SS3	88.52	89.26	93.19	90.75	88.00
SS4	10.00	42.50	63.08	87.41	84.26

TABLE VI
MMSE ESTIMATOR'S PERFORMANCE (PERCENT ACCURATE) FOR THE AURORA2 TASK USING SEQUENTIAL ML AND MAP NOISE ESTIMATES. THE BASELINE RESULTS ARE FROM THE ALGORITHM PUBLISHED EARLIER AND WERE USED AS x_0 OF (17) TO INITIALIZE THE ITERATIVE MMSE ESTIMATOR. THE REMAINING RESULTS ARE FROM THE MMSE ESTIMATE OF $\hat{x}$ IN (17) USING DIFFERENT POINT ESTIMATES OF THE NOISE

	Baseline (x_0 in Eq. 17)	ML-tracked noise	MAP-tracked noise
SetA	85.66	86.34	86.39
SetB	86.15	86.24	86.30
SetC	80.40	82.50	83.35
Ave.	84.80	85.53	85.74

a careful examination of (2). It shows that the estimation error in the noise power $|N(k)|^2$ and the "phase" term of $2|X[k]||H[k]||N[k]|\cos\theta_k$ contribute equally to accounting for the observed noisy speech power $|Y(k)|^2$. When the noise estimation error exceeds the "phase" term, the speech feature algorithm designed to exploit the "phase" term will naturally lose its effectiveness and the converged estimate may be far away from the desired clean speech estimate.

The results of Table VI demonstrate that even with the noise being inaccurately estimated, the use of the phase information in the speech distortion process for noise reduction is beneficial for robust speech recognition under realistic conditions when the algorithm's iteration is appropriately controlled. The significantly lower recognition rates shown in Table VI than those in Table I highlight the importance of accurate estimation of the noise in enhancing the benefit of using the phase information.

D. Results on Sensitivity to Variances of the Phase Factor

The probabilistic, phase-sensitive model for the acoustic environment as presented in (9) and derived in Section II has only one parameter set — the covariance matrix $\mathbf{\Sigma}_\alpha$ of the phase factor $\boldsymbol{\alpha}$, or the variances of individual vector components $\Sigma_\alpha^{(l)}$, $l = 1, 2, \ldots, L$ under the diagonal covariance assumption.[7]

The parameter $\Sigma_\alpha^{(l)}$ is estimated from a set of the Aurora2 training data, disjoint from any set (A or B or C) of test data. The estimate is the sample variance computed from all the sample values of $\alpha^{(l)}$'s, which are computed using (28). (Data dependence of this estimate is extremely low, as is expected from its definition.) A linear regression line is fit to the computed $\Sigma_\alpha^{(l)}$ as a function of l. The estimated variances of the phase factor, $\hat{\Sigma}_\alpha^{(l)}$, used in the experiments presented so far, are taken from such a regression line fit and are shown in Table VII.

[7]The vector size is the total number of Mel-filter banks, and we use $L = 23$ for the experiments.

TABLE VII
ESTIMATED VARIANCES OF THE PHASE FACTOR AS A FUNCTION OF THE MEL-FILTER BANK (l). INDEX l INCREASES FROM LOW-FREQUENCY TO HIGH-FREQUENCY MEL-SCALED CHANNELS

Filter l	1	3	6	10	14	18	21	23
$\hat{\Sigma}_\alpha^{(l)}$	0.1943	0.1821	0.1577	0.1394	0.1150	0.0906	0.0723	0.0601

TABLE VIII
PERCENT ACCURATE DIGIT RECOGNITION RATE AS A FUNCTION OF THE VARIANCE OF THE PHASE FACTOR; SET A IN THE AURORA2 TEST DATA

$0.25\ \hat{\Sigma}_\alpha^{(l)}$	$0.8\ \hat{\Sigma}_\alpha^{(l)}$	$\hat{\Sigma}_\alpha^{(l)}$	$2\ \hat{\Sigma}_\alpha^{(l)}$	$5\ \hat{\Sigma}_\alpha^{(l)}$	$10\ \hat{\Sigma}_\alpha^{(l)}$	$100\ \hat{\Sigma}_\alpha^{(l)}$
85.52	86.32	86.34	86.25	86.15	86.07	81.00

To investigate the sensitivity of the recognition performance based on the MMSE estimator (17) to the variances of the phase factor, we artificially perturb the $\hat{\Sigma}_\alpha^{(l)}$ using a variable scaling. The resulting digit recognition performance on Set A of the Aurora2 test data is shown in Table VIII. The recognition results appear to be relatively robust against the variances over a rather wide range.

VI. SUMMARY AND CONCLUSION

In this paper we present an MMSE speech feature enhancement algorithm, capitalizing on a probabilistic and phase-sensitive environment model for acoustic distortion. The model effectively incorporates the phase relationship between the clean speech and the corrupting noise during the process of speech corruption via the use of the newly introduced phase factor as a random parameter. The MMSE estimator based on this phase-sensitive model is derived, which achieves high efficiency by exploiting single-point Taylor series expansion to approximate the joint probability of clean and noisy speech as a multivariate Gaussian. This forms sharp contrast to the use of the M-point Taylor series expansion (one for each mixture component in the clean speech model), which is computationally very expensive as implemented earlier [14] based on a phase-insensitive environment model.

As an integral component of the enhancement algorithm using the phase-sensitive MMSE estimator, a point estimator for (nonstationary) noise is derived and presented based on a new sequential MAP noise tracker. We show that under the special case where the variance of the noise prior distribution approaches infinity, the MAP noise estimator naturally reduces to the ML counterpart as published earlier.

Experimental results obtained on the Aurora2 task demonstrate the importance of exploiting the phase relationship in the speech corruption process captured by the MMSE estimator. The phase-sensitive MMSE estimator reported in this paper performs significantly better than phase-insensitive spectral subtraction (54% error rate reduction), and also noticeably better than a phase-insensitive MMSE estimator as our previous state-of-the-art technique reported in [7] (6% error rate reduction), both under carefully controlled experimental conditions for connected noisy digit recognition. In particular, we showed that spectral subtraction is a generated case of the new technique presented in this paper by setting the phase factor to zero. The Aurora2 recognition results demonstrate that such degeneration hurts the performance significantly, offering direct and quantitative evidence for the importance of including the phase factor in the model as the main novelty of this work. The experimental results also demonstrate superior performance of the MAP noise tracker over the ML counterpart when they are used in the otherwise identical MMSE estimator derived based on the phase-sensitive model of acoustic distortion.

The phase-sensitive environment model presented in this paper can be viewed as a generalization of the phase-insensitive model described in [14], [7] in that the SNR-dependent residual variances in the latter are automatically accommodated in the former. The sample residual variances computed from the phase-sensitive model shown in Table VII are invariant with respect to the (instantaneous) SNR. On the other hand, the sample variances derived from the phase-insensitive model are much higher using the data files of low SNRs than those with high SNRs. Since the (instantaneous) SNR is generally unknown, it is impractical to automatically represent the SNR-dependent variance in the phase-insensitive model of [14], [7]. The phase-sensitive model presented in this paper naturally overcomes this aspect of the weakness inherent in the phase-insensitive model.

While we have presented positive results in Table VI for the effectiveness of the phase-sensitive MMSE estimator for clean speech using realistic noise trackers, the accuracy obtained has been far below the results shown in Table I using true noise. This highlights the crucial role of accurate noise estimation in enhancing the use of the phase information. Accurate (point) noise estimation is extremely challenging, since noise is random and it is only possible to obtain a reliable estimate of noise statistics, not its instantaneous realization. In order to improve the phase-sensitive modeling technique for speech enhancement, we are currently working on sequential updating of noise statistics as the prior distribution, and on incorporating posterior noise distributions into a new version of the phase-sensitive MMSE estimator. Finally, based on the dramatic performance improvement of the phase-insensitive MMSE estimator by incorporating dynamic aspects of the speech prior distribution [7], another promising research direction is to strive for similar significant improvement on the current phase-sensitive MMSE estimator after incorporating the same type of dynamic information in a rigorous fashion.

APPENDIX
DERIVATION OF (3)

Starting from (2). After applying a set of Mel-scale filters (L in total) to the spectrum $|Y[k]|^2$ in the frequency domain, where the l^{th} filter is characterized by the transfer function $W_k^{(l)} \geq 0$

(where $\sum_k W_k^{(l)} = 1$), we obtain a total of L channel (Mel-filter) energies of

$$\sum_k W_k^{(l)}|Y[k]|^2 = \sum_k W_k^{(l)}|X[k]|^2|H[k]|^2 + \sum_k W_k^{(l)}|N[k]|^2 + 2\sum_k W_k^{(l)}|X[k]||H[k]||N[k]|\cos\theta_k \quad (26)$$

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

Denoting the various channel energies in (26) 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 (26) to

$$|\tilde{Y}^{(l)}|^2 = |\tilde{X}^{(l)}|^2|\tilde{H}^{(l)}|^2 + |\tilde{N}^{(l)}|^2 + 2\alpha^{(l)}|\tilde{X}^{(l)}||\tilde{H}^{(l)}||\tilde{N}^{(l)}| \quad (27)$$

where we define the "phase factor" as

$$\alpha^{(l)} \equiv \frac{\sum_k W_k^{(l)}|X[k]||H[k]||N[k]|\cos\theta_k}{|\tilde{X}^{(l)}||\tilde{H}^{(l)}||\tilde{N}^{(l)}|}. \quad (28)$$

Since $\cos\theta_k \leq 1$, we have

$$|\alpha^{(l)}| \leq \frac{\sum_k W_k^{(l)}|X[k]||H[k]||N[k]|}{|\tilde{X}^{(l)}||\tilde{H}^{(l)}||\tilde{N}^{(l)}|}.$$

The right-hand side above is the normalized inner product of two vectors $\bar{N}$ and $\bar{X}^H$, whose elements are $\bar{N}_k \equiv \sqrt{W_k^{(l)}}|\tilde{N}^{(l)}(k)|$ and $\bar{X}_k^H \equiv \sqrt{W_k^{(l)}}|\tilde{X}^{(l)}(k)||\tilde{H}^{(l)}(k)|$. Hence

$$|\alpha^{(l)}| \leq \frac{<\bar{N}, \bar{X}^H>}{|\bar{N}||\bar{X}^H|} \leq 1.$$

Further, we define the log channel energy (log-spectrum) vectors

$$\mathbf{y} = \begin{bmatrix} \log|\tilde{Y}^{(1)}|^2 \\ \log|\tilde{Y}^{(2)}|^2 \\ \vdots \\ \log|\tilde{Y}^{(l)}|^2 \\ \vdots \\ \log|\tilde{Y}^{(L)}|^2 \end{bmatrix},\ \mathbf{x} = \begin{bmatrix} \log|\tilde{X}^{(1)}|^2 \\ \log|\tilde{X}^{(2)}|^2 \\ \vdots \\ \log|\tilde{X}^{(l)}|^2 \\ \vdots \\ \log|\tilde{X}^{(L)}|^2 \end{bmatrix},$$
$$\mathbf{n} = \begin{bmatrix} \log|\tilde{N}^{(1)}|^2 \\ \log|\tilde{N}^{(2)}|^2 \\ \vdots \\ \log|\tilde{N}^{(l)}|^2 \\ \vdots \\ \log|\tilde{N}^{(L)}|^2 \end{bmatrix},\ \mathbf{h} = \begin{bmatrix} \log|\tilde{H}^{(1)}|^2 \\ \log|\tilde{H}^{(2)}|^2 \\ \vdots \\ \log|\tilde{H}^{(l)}|^2 \\ \vdots \\ \log|\tilde{H}^{(L)}|^2 \end{bmatrix} \quad (29)$$

and define the vector of phase factors

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

Then, we rewrite (27) as

$$e^{\mathbf{y}} = e^{\mathbf{x}} \bullet e^{\mathbf{h}} + e^{\mathbf{n}} + 2\,\boldsymbol{\alpha} \bullet e^{\mathbf{x}/2} \bullet e^{\mathbf{h}/2} \bullet e^{\mathbf{n}/2} = e^{\mathbf{x}+\mathbf{h}} + e^{\mathbf{n}} + 2\,\boldsymbol{\alpha} \bullet e^{(\mathbf{x}+\mathbf{h}+\mathbf{n})/2} \quad (30)$$

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 (30)

$$\mathbf{y} = \log\left[e^{\mathbf{x}+\mathbf{h}} \bullet (\mathbf{1} + e^{\mathbf{n}-\mathbf{x}-\mathbf{h}} + 2\boldsymbol{\alpha} \bullet e^{(\mathbf{x}+\mathbf{h}+\mathbf{n})/2-\mathbf{x}-\mathbf{h}})\right] = \mathbf{x} + \mathbf{h} + \log[\mathbf{1} + e^{\mathbf{n}-\mathbf{x}-\mathbf{h}} + 2\boldsymbol{\alpha} \bullet e^{(\mathbf{n}-\mathbf{x}-\mathbf{h})/2}].$$

ACKNOWLEDGMENT

The authors wish to thank B. Frey and T. Kristjansson for the earlier joint work of [14] that motivated the iterative approach presented in this paper. They also thank the reviewers who provided constructive comments that improve the presentation of this paper.

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. ICSLP*, vol. 3, 2000, pp. 869–872.

[3] M. Afify and O. Siohan, "Sequential noise estimation with optimal forgetting for robust speech recognition," in *Proc. ICASSP*, vol. 1, 2001, pp. 229–232.

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

[5] 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.

[6] 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.

[7] L. Deng, J. Droppo, and A. Acero, "A Bayesian approach to speech feature enhancement using the dynamic cepstral prior," in *Proc. ICASSP*, vol. 1, May 2002, pp. 829–832.

[8] ——, "Recursive noise estimation using iterative stochastic approximation for stereo-based robust speech recognition," in *Proc. Automatic Speech Recognition and Understanding*, Trento, Italy, Dec. 2001.

[9] ——, "Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 11, pp. 568–580, Nov. 2003.

[10] 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.

[11] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora2 database," in *Proc. Eurospeech*, vol. 1, Sept. 2001, pp. 217–220.

[12] 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.

[13] ——, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, pp. 1526–1555, Oct. 1992.
[14] 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*, vol. 2, Sept. 2001, pp. 901–904.
[15] 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.
[16] N. S. Kim, "Nonstationary environment compensation based on sequential estimation," *IEEE Signal Processing Lett.*, vol. 5, pp. 57–60, 1998.
[17] *Speech Enhancement*, J. S. Lim, Ed., Prentice-Hall, London, U.K., 1983.
[18] P. Lockwood and J. Boudy, "Experiments with a nonlinear spectral subtraction (NSS), hidden Markov models and the projection for robust speech recognition in cars," *Speech Commun.*, vol. 11, pp. 215–228, 1992.
[19] J. Mendel, *Lessons in Estimation Theory for Signal Processing, Communications and Control*. Englewood Cliffs, NJ: Prentice-Hall, 1995.
[20] P. Moreno, B. Raj, and R. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proc. ICASSP*, vol. 2, 1996, pp. 733–736.
[21] *ESE2 Special Sessions on Noise Robust Recognition, Proc. Eurospeech*, D. Pearce, Ed., Aalborg, Denmark, Sept. 2001.
[22] 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. Eurospeech*, Aalborg, Denmark, Sept. 2001.
[23] D. M. Titterington, "Recursive parameter estimation using incomplete data," *J. R. Statist. Soc. B*, vol. 46, pp. 257–267, 1984.
[24] *Speech Commun.*, vol. 34, O. Viikki, Ed., 2001.
[25] Q. Zhu and A. Alwan, "The effect of additive noise on speech amplitude spectra: A quantitative analysis," *IEEE Signal Processing Lett.*, vol. 9, pp. 275–277, Sept. 2002.

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.