Robust Speech Recognition Using a Cepstral Minimum-Mean-Square-Error-Motivated Noise Suppressor

Dong Yu, *Senior Member, IEEE*, Li Deng, *Fellow, IEEE*, Jasha Droppo, *Senior Member, IEEE*, Jian Wu, *Member, IEEE*, Yifan Gong, *Senior Member, IEEE*, and Alex Acero, *Fellow, IEEE*

***Abstract*—We present an efficient and effective nonlinear feature-domain noise suppression algorithm, motivated by the minimum-mean-square-error (MMSE) optimization criterion, for noise-robust speech recognition. Distinguishing from the log-MMSE spectral amplitude noise suppressor proposed by Ephraim and Malah (E&M), our new algorithm is aimed to minimize the error expressed explicitly for the Mel-frequency cepstra instead of discrete Fourier transform (DFT) spectra, and it operates on the Mel-frequency filter bank's output. As a consequence, the statistics used to estimate the suppression factor become vastly different from those used in the E&M log-MMSE suppressor. Our algorithm is significantly more efficient than the E&M's log-MMSE suppressor since the number of the channels in the Mel-frequency filter bank is much smaller (23 in our case) than the number of bins (256) in DFT. We have conducted extensive speech recognition experiments on the standard Aurora-3 task. The experimental results demonstrate a reduction of the recognition word error rate by 48% over the standard ICSLP02 baseline, 26% over the cepstral mean normalization baseline, and 13% over the popular E&M's log-MMSE noise suppressor. The experiments also show that our new algorithm performs slightly better than the ETSI advanced front end (AFE) on the well-matched and mid-mismatched settings, and has 8% and 10% fewer errors than our earlier SPLICE (stereo-based piecewise linear compensation for environments) system on these settings, respectively.**

***Index Terms*—Mel-frequency cepstral coefficient (MFCC), minimum-mean-square-error (MMSE) estimate, noise reduction, phase asynchrony, robust automatic speech recognition (ASR).**

I. Introduction

IT IS generally held that the desirable signal domain to which noise reduction or speech enhancement should be applied differs between human listening and automatic speech recognition (ASR). Conventional wisdom posits that the lower the distortion is between the enhanced speech and the clean speech in the domain closest to the "backend" (human perception or machine recognition), the better the enhancement performance will be. For subjective human listening, noise reduction has been traditionally applied in the spectral domain (e.g., spectral subtraction, Wiener filtering, and Ephraim/Malah spectral amplitude MMSE suppressor [7], [13]). Subjective human listening experiments [11] show that speech enhancement becomes more effective when it is applied to the logarithmic spectral amplitude domain [8]. This agrees with the observation that the periphery auditory system performs the kind of compression similar to logarithmic scaling [5].

In this paper, we apply the same line of thinking to speech feature enhancement for ASR applications, where Mel-frequency cepstral coefficients (MFCCs) have been proven to be effective and used pervasively as the direct input to the ASR backend [12]. Specifically, we propose a nonlinear feature-domain noise reduction algorithm motivated by the minimum-mean-square-error (MMSE) criterion on MFCCs, which are immediate to the ASR backend, for environment-robust speech recognition. We explain that the problem of seeking an MMSE estimator on MFCCs can be reduced to seeking a log-MMSE estimator on the Mel-frequency filter bank's output, which can be solved independently for each filter bank channel. We derive the algorithm by assigning uniformly distributed random phases to the real-valued filter bank's outputs and assuming that the artificially generated complex filter bank's outputs follow zero-mean complex normal distributions. We show that while the suppression rule derived in this way is similar in form to the log-MMSE spectral amplitude estimator proposed by Ephraim and Malah (E&M) [8], it has two important differences. First, the suppression rule in our algorithm is applied to the power spectral magnitude of the filter bank's output instead of the discrete Fourier transform (DFT) spectral amplitude. Second, the noise variance used in our algorithm has been derived to contain an additional term resulting from the fact that the clean speech and the mixing noise are not in phase with each other. We also demonstrate that operating on the MFCC, which is closer to the backend, does provide us better performance compared with approaches operating on the DFT domain.

Compared with our previous noise-robust technique of SPLICE (stereo-based piecewise linear compensation for environments) [2], [6], the new algorithm has three distinctive advantages. First, it does not require a codebook to be constructed using training data and thus is more robust to unseen environments and easier to deploy. Second, it introduces no additional look-ahead frame delay. Third, it is applied to the filter bank's output and hence can be easily plugged into the existing feature extraction pipeline. Speech recognition experiments on the Aurora-3 task, to be shown in Section V, demonstrate that our proposed algorithm can reduce word error rate (WER)

Manuscript received October 17, 2007; revised February 22, 2008. This is a significantly expanded version of the paper "a minimum-mean-square-error noise reduction algorithm on Mel-frequency cepstra for robust speech recognition" [18] presented at ICASSP 2008, Las Vegas, NV, in April, 2008. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Abeer Alwan.

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

Digital Object Identifier 10.1109/TASL.2008.921761

1558-7916/$25.00 © 2008 IEEE

by 48.33% relative to the ICSLP02 baseline, 25.59% over the cepstral mean normalization baseline, and 13.41% over the conventional E&M log-MMSE noise suppressor. Compared with the E&M log-MMSE noise suppressor, our new algorithm is also much more efficient since the number of the channels in the Mel-frequency filter bank is much smaller than the number of bins in the DFT (23 versus 256). The results also show that our algorithm performs slightly better than the ETSI advanced front end (AFE) [14] and SPLICE [2], [6] on the well-matched and mid-mismatched settings.

The rest of the paper is organized as follows. In Section II, we formulate the MMSE estimation problem in the MFCC domain and show how the problem can be reduced to log-MMSE estimation of the Mel-frequency filter bank's outputs. In Section III, we provide detailed derivation of the nonlinear noise reduction algorithm, with the emphasis on the special treatments employed and the difficulties introduced by the use of the Mel-frequency filter bank. In Section IV, we illustrate how the parameters used in the algorithm are estimated with the focus on noise tracking and the additional variance caused by the phase difference between the clean speech and the mixing noise. We describe the evaluation procedure on the Aurora-3 task and report the experimental results in Section V. In Section VI, we conclude the paper.

II. Problem Formulation

Without lack of generality, we denote $x(t)$ as the channel-convoluted clean speech waveform and refer to it as clean speech henceforth. We assume that $x(t)$ is corrupted with the independent additive noise waveform $n(t)$ to become the noisy speech waveform $y(t)$, i.e.,

$$y(t) = x(t) + n(t) \tag{1}$$

where t is the sampled time index. Given the additive noise assumption (1), we get the relationship in the DFT domain

$$Y(f) = X(f) + N(f) \tag{2}$$

where $Y(f)$, $X(f)$, and $N(f)$ are the DFT of the noisy speech waveform $y(t)$, the clean speech waveform $x(t)$, and the noise waveform $n(t)$, respectively.

The Mel-frequency filter bank's output power for noisy speech is

$$m_y(b) = \sum_f w_b(f)|Y(f)|^2 \tag{3}$$

where $w_b(f)$ is the bth Mel-frequency filter's weight for the frequency bin f. A similar relationship holds between the filter-bank output of clean speech $m_x(b)$ and its DFT $X(f)$, and between the filter bank output of the noise $m_n(b)$ and its DFT $N(f)$. The kth dimension of MFCC is calculated as

$$c_y(k) \cong \sum_b a_{k,b} m_y(b) \tag{4}$$

where

$$a_{k,b} = \cos \frac{\pi b}{B}(k - 0.5)$$

are the discrete cosine transform coefficients.

Our goal is to find the MMSE estimate $\hat{c}_x(k)$ against each separate and independent dimension k of the clean speech's MFCC vector $\boldsymbol{c}_x$, given the noisy MFCC $c_y(k)$. More specifically, we aim to find a mapping $\hat{f}$ from $c_y(k)$ to $\hat{c}_x(k)$ such that

$$\begin{aligned}\hat{c}_x(k) &= \hat{f}(c_y(k)) \\ &= \arg\min_f E\left\{(f(c_y(k)) - c_x(k))^2\right\} \\ &= \arg\min_f \int (f(c_y(k)) - c_x(k))^2 p(c_x(k)) dc_x(k).\end{aligned} \tag{5}$$

There are three reasons for choosing the dimension-wise instead of the full-vector MMSE criterion. First, each dimension of MFCC vector is known to be relatively independent with each other, and hence diagonal covariance matrices are usually used in modeling the MFCC space in ASR [12]. Second, the dynamic range of MFCC is vastly different across dimensions. If the MMSE criterion is applied to the MFCC vector, each dimension needs to be weighted differently to avoid the problem that the error is dominated by one or two dimensions. Choosing the appropriate weights not only is difficult but it also introduces unnecessary computational overhead. Third, the dimension-wise MMSE criterion decouples different dimensions, making the algorithm easier to develop and to implement.

Based on standard estimation theory, it can be shown that the solution to (5) is the conditional expectation

$$\begin{aligned}\hat{c}_x(k) &= E\{c_x(k)|\boldsymbol{m}_y\} = E\left\{\sum_b a_{k,b} \log m_x(b)|\boldsymbol{m}_y\right\} \\ &= \sum_b a_{k,b} E\left\{\log m_x(b)|\boldsymbol{m}_y\right\}.\end{aligned} \tag{6}$$

Note that according to (2) and (3), we can assume that $m_x(b)$ is independent of $m_y(b') \; \forall \, b' \neq b$ given $m_y(b)$ and thus can be reconstructed solely from $m_y(b)$; i.e., (6) can be further simplified to

$$\hat{c}_x(k) \cong \sum_b a_{k,b} E\{\log m_x(b)|m_y(b)\}. \tag{7}$$

The problem is thus reduced to finding the log-MMSE estimator of the Mel-frequency filter bank's output

$$\widehat{m}_x(b) = \exp(E\{\log m_x(b)|m_y(b)\}). \tag{8}$$

There can be many different solutions to (8) based on different assumptions on the noise and noisy speech models. In Sections III–V, we derive a solution by assigning uniformly distributed random phases to the real-valued filter bank's outputs and assuming that the artificially generated complex filter bank's outputs follow zero-mean complex normal distributions.

III. Noise Suppressor for MFCC

To motivate the solution, we set up a "straw man" by first rewriting (8) to

$$\begin{aligned}\widehat{m}_x(b) &= \exp(E\{\log m_x(b)|m_y(b)\}) \\ &= \exp(2E\{\log \sqrt{m_x(b)}|\sqrt{m_y(b)}\})\end{aligned} \tag{9}$$

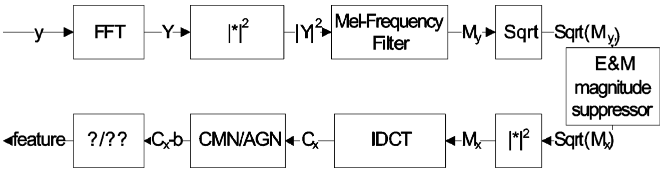


Fig. 1. Feature extraction pipeline where the E&M log-MMSE magnitude suppressor is directly applied to the magnitude spectrum of the filter bank output.

which has exactly the same form in the objective function as the E&M log-MMSE amplitude spectral suppressor [8], since $\sqrt{m_y(b)}$ is in the amplitude spectral domain. At the first glance of (9), it appears that we can apply the E&M log-MMSE noise suppressor directly to the filter bank output by first converting the power spectra to the amplitude spectra and then converting it back after the E&M suppression is applied. These steps, as part of the overall feature extraction process in our ASR system, are illustrated in Fig. 1. This naive approach, however, violates the original assumptions made in the E&M log-MMSE suppressor and has produced poor recognition results in our experiments (see details in Section V). Such failure motivates a more principled approach, which we have developed as will be described in this section.

Note that the filter banks' outputs $m_x(b)$, $m_n(b)$, and $m_y(b)$ take real values in the range of $[0, \infty)$ according to (3), and thus it is inappropriate to model them with real-valued normal distributions. To develop appropriate models, we construct three artificial complex variables $M_x(b)$, $M_n(b)$, and $M_y(b)$ such that

$$\begin{aligned}|M_x(b)| &= m_x(b) = \sum_f w_b(f)|X(f)|^2\\ |M_n(b)| &= m_n(b) = \sum_f w_b(f)|N(f)|^2\\ |M_y(b)| &= m_y(b) = \sum_f w_b(f)|Y(f)|^2.\end{aligned} \tag{10}$$

That is, we consider $m_x(b)$, $m_n(b)$, and $m_y(b)$ as modulus of the constructed complex variables $M_x(b)$, $M_n(b)$, and $M_y(b)$, respectively. Many $M_x(b)$, $M_n(b)$, and $M_y(b)$ would satisfy (10), among which we choose the ones with uniformly distributed random phases $\theta_x(b)$, $\theta_n(b)$, and $\theta_y(b)$ (which can be considered as the weighted sum of the phases over all the DFT bins). Selecting the uniformly distributed random phases permits us to make the assumption that complex variables $M_x(b)$ and $M_y(b) - M_x(b)$ both follow the zero-mean complex normal distributions. Note that mapping the real variables $m_x(b)$, $m_n(b)$, and $m_y(b)$ to the complex variables $M_x(b)$, $M_n(b)$ and $M_y(b)$ allows us to operate at a two-dimensional space with simpler statistical models.

Since $M_y(b)$ contains all information there is in $m_y(b)$, (8) can be rewritten as

$$\widehat{m}_x(b) = \exp(E\{\log m_x(b)|M_y(b)\}). \tag{11}$$

To solve (11), we follow the approach adopted in [8] by first evaluating the moment generating function

$$\begin{aligned}\Phi_b(\mu) &= E\{\exp(\mu \log m_x(b))|M_y(b)\}\\ &= E\left\{m_x^\mu(b)|M_y(b)\right\}\end{aligned} \tag{12}$$

and then find the solution to (11) as

$$\widehat{m}_x(b) = \exp\left(\frac{d}{d\mu}\Phi_b(\mu)|_{\mu=0}\right) \tag{13}$$

which can be easily verified by noting

$$\frac{d}{d\mu}m_x^\mu = m_x^\mu \log m_x.$$

Now we assume that $\theta_x(b)$, $\theta_n(b)$, and $\theta_y(b)$ are independent and uniformly distributed (from 0 to 2π) random variables, as shown by (14) at the bottom of the page.

Since $M_x(b)$ is assumed to follow the zero-mean complex normal distribution [7], [17], we have

$$p(m_x(b), \theta_x(b)) = \frac{m_x(b)}{\pi\sigma_x^2(b)} \exp\left\{-\frac{m_x^2(b)}{\sigma_x^2(b)}\right\} \tag{15}$$

where $\sigma_x^2(b) \overset{\text{def}}{=} E\left\{|M_x(b)|^2\right\} = E\left\{m_x^2(b)\right\}$. Similarly, given that $M_y(b) - M_x(b)$ follows the zero-mean complex normal distribution, we obtain (16) at the bottom of the next page, where

$$\begin{aligned}\sigma_d^2(b) &\overset{\text{def}}{=} E\left\{|M_y(b) - M_x(b)|^2\right\}\\ &\geq E\left\{(m_y(b) - m_x(b))^2\right\}\end{aligned} \tag{17}$$

and triangular inequality is used above. Since

$$\begin{aligned}m_y(b) &= \sum_f w_b(f)|Y(f)|^2\\ &= \sum_f w_b(f)(|X(f)|^2 + |N(f)|^2\\ &\qquad + 2|X(f)||N(f)|\cos\varphi(f))\\ &= m_x(b) + m_n(b)\\ &\quad + \sum_f 2w_b(f)|X(f)||N(f)|\cos\varphi(f)\end{aligned} \tag{18}$$

$$\begin{aligned}\Phi_b(\mu) &= E\left\{m_x^\mu(b)|M_y(b)\right\}\\ &= \frac{\int_0^\infty \int_0^{2\pi} m_x^\mu(b)p(M_y(b), m_x(b), \theta_x(b))dm_x(b)d\theta_x(b)}{p(M_y(b))}\\ &= \frac{\int_0^\infty \int_o^{2\pi} m_x^\mu(b)p(M_y(b)|m_x(b), \theta_x(b))p(m_x(b), \theta_x(b))dm_x(b)d\theta_x(b)}{\int_0^\infty \int_0^{2\pi} p(M_y(b)|m_x(b), \theta_x(b))p(m_x(b), \theta_x(b))dm_x(b)d\theta_x(b)}\end{aligned} \tag{14}$$

where $\varphi(f)$ is the phase difference between $X(f)$ and $N(f)$, it follows that

$$\begin{aligned}\sigma_d^2(b) &\geq E\left\{\left(m_n(b)+\sum_f 2|X(f)||N(f)|\cos\varphi(f)w_b(f)\right)^2\right\}\\ &= E\left\{m_n^2(b)\right\}+E\left\{\left(\sum_f 2|X(f)||N(f)|\cos\varphi(f)w_b(f)\right)\right\}\end{aligned} \tag{19}$$

where we have used the result of

$$E\left\{2m_n(b)\left(\sum_f 2w_b(f)|X(f)||N(f)|\cos\varphi(f)\right)\right\} \cong 0$$

given that $X(f)$ and $N(f)$ are independent of each other. It was shown in [4] that $\sum_f 2w_b(f)|X(f)||N(f)|\cos\varphi(f)$ approximately follows a zero-mean normal distribution. If we denote its variance by $\sigma_\varphi^2(b)$, we then have

$$\sigma_d^2(b) \cong \sigma_n^2(b) + \sigma_\varphi^2(b) \tag{20}$$

where $\sigma_n^2(b) = E\left\{m_n^2(b)\right\}$. Note that (20) is one of the major differences between our approach and the original E&M log-MMSE algorithm. In the original E&M log-MMSE algorithm

$$N(f) = Y(f) - X(f)$$

and so

$$\sigma_d^2(f) \overset{\text{def}}{=} E\{|Y(f)-X(f)|^2\} = E\{|N(f)|^2\} = \sigma_n^2(f).$$

By substituting (15) and (16) into (14) and replacing variable $\theta_y(b) - \theta_x(b)$ by $\beta(b)$, we obtain

$$\Phi_b(\mu) = \frac{\int_0^\infty m_x^{\mu+1}(b)\exp\left\{-\frac{m_x^2(b)}{\sigma_x^2(b)} - \frac{m_x^2(b)}{\sigma_d^2(b)}\right\} g(m_x(b))dm_x(b)}{\int_0^\infty m_x(b)\exp\left\{-\frac{m_x^2(b)}{\sigma_x^2(b)} - \frac{m_x^2(b)}{\sigma_d^2(b)}\right\} g(m_x(b))dm_x(b)} \tag{21}$$

where

$$g(m_x(b)) = \int_0^{2\pi} \frac{1}{\pi\sigma_d^2(b)} \times \exp\left\{\frac{2m_x(b)m_y(b)\cos(\beta(b))}{\sigma_d^2(b)}\right\} d\beta(b). \tag{22}$$

This can be shown to be simplified to [8]

$$g(m_x(b)) = I_0\left(2m_x(b)\sqrt{\frac{v(b)}{\sigma^2(b)}}\right) \tag{23}$$

where

$$I_0(z) = \frac{1}{2\pi}\int_0^{2\pi} \exp\{z\cos\beta\}d\beta \tag{24}$$

is the integral representation of the modified zeroth-order Bessel function

$$\frac{1}{\sigma^2(b)} = \frac{1}{\sigma_x^2(b)} + \frac{1}{\sigma_d^2(b)} \tag{25}$$

and

$$v(b) = \frac{\xi(b)}{1+\xi(b)}\gamma(b) \tag{26}$$

is defined from the *a priori* signal-to-noise ratio (SNR) (modified and adjusted from that proposed in [7] and [8] by using the cepstra-domain representation and by accounting for phase asynchrony)

$$\xi(b) \overset{\text{def}}{=} \frac{\sigma_x^2(b)}{\sigma_d^2(b)} \cong \frac{\sigma_x^2(b)}{\sigma_n^2(b) + \sigma_\varphi^2(b)} \tag{27}$$

and the adjusted *a posteriori* SNR

$$\gamma(b) \overset{\text{def}}{=} \frac{m_y^2(b)}{\sigma_d^2(b)} \cong \frac{m_y^2(b)}{\sigma_n^2(b) + \sigma_\varphi^2(b)}. \tag{28}$$

Given the above definitions, (21) can be rewritten as

$$\Phi_b(\mu) = \frac{\int_0^\infty m_x^{\mu+1}(b)\exp\left\{-\frac{m_x^2(b)}{\sigma^2(b)}\right\} I_0(2m_x(b)\sqrt{\frac{v(b)}{\sigma^2(b)}})dm_x(b)}{\int_0^\infty m_x(b)\exp\left\{-\frac{m_x^2(b)}{\sigma^2(b)}\right\} I_0\left(2m_x(b)\sqrt{\frac{v(b)}{\sigma^2(b)}}\right) dm_x(b)}. \tag{29}$$

$$\begin{aligned} &p(M_y(b)|m_x(b),\theta_x(b))\\ &\quad= \frac{1}{\pi\sigma_d^2(b)}\exp\left\{-\frac{\left|M_y(b)-m_x(b)e^{j\theta_x(b)}\right|^2}{\sigma_d^2(b)}\right\}\\ &\quad= \frac{1}{\pi\sigma_d^2(b)}\exp\left\{-\frac{\left|m_y(b)e^{j\theta_y(b)}-m_x(b)e^{j\theta_x(b)}\right|^2}{\sigma_d^2(b)}\right\}\\ &\quad= \frac{1}{\pi\sigma_d^2(b)}\exp\left\{-\frac{\left|\begin{matrix} m_y(b)\cos\theta_y(b) - m_x(b)\cos\theta_x(b)\\ +j(m_y(b)\sin\theta_y(b) - m_x(b)\sin\theta_x(b))\end{matrix}\right|}{\sigma_d^2(b)}\right\}\\ &\quad= \frac{1}{\pi\sigma_d^2(b)}\exp\left\{-\frac{m_y^2(b)+m_x^2(b)-2m_x(b)m_y(b)\cos(\theta_y(b)-\theta_x(b))}{\sigma_d^2(b)}\right\}\end{aligned} \tag{16}$$

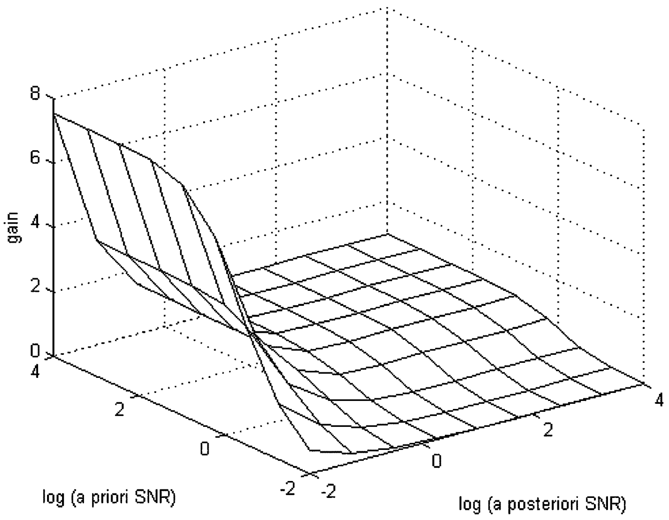


Fig. 2. Gain function against the *a priori* SNR and the *a posteriori* SNR.

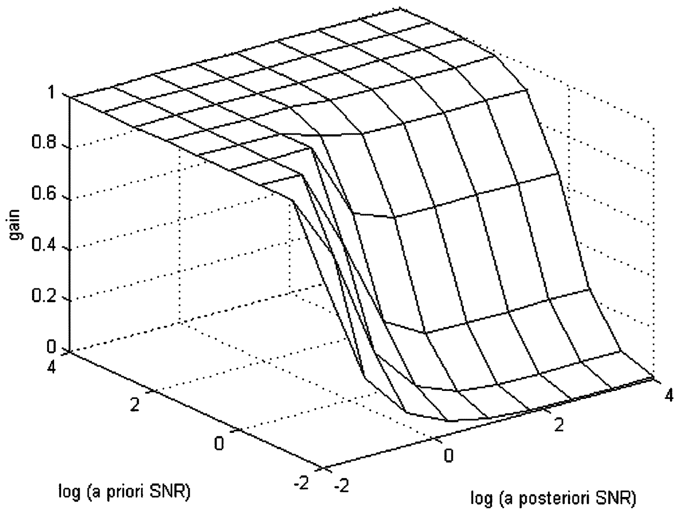


Fig. 3. Gain function restricted to be less than or equal to 1.

Following similar steps in [8], we obtain

$$\widehat{m}_x(b) = \exp(E\{\log m_x(b)|m_y(b)\}) \\ = G(\xi(b), v(b))m_y(b) \tag{30}$$

where

$$G(\xi(b), v(b)) = \frac{\xi(b)}{1+\xi(b)} \exp\left\{\frac{1}{2}\int_{v(b)}^{\infty} \frac{e^{-t}}{t} dt\right\} \tag{31}$$

can be calculated efficiently using the technique described in [8]. The MMSE estimate for the MFCC is thus

$$\hat{c}_x(k) \cong \sum_b a_{k,b} E\{\log m_x(b)|m_y(b)\} \\ = \sum_b a_{k,b} \log(G(\xi(b), v(b))m_y(b)). \tag{32}$$

Fig. 2 depicts the gain function (31) against the *a priori* SNR $\xi(b)$ and the *a posteriori* SNRs $\gamma(b)$. Note that normally the *a priori* SNR and *a posteriori* SNR do not differ too much since $E(\gamma(b) - 1) = E(\xi(b))$. In other words, some of the combinations of $\xi(b)$ and $\gamma(b)$ in the figure are rarely seen in the real situation. Also note that the gain function (31) allows for gains larger than 1 under some conditions. In our experiments; however, we have chosen to restrict the gains to be less than or equal to 1 and used the gain function illustrated in Fig. 3 instead. This restricted gain function performs slightly (not statistically significant) better than the original gain function in our experiments.

We would like to point out that while the noise suppression rule (30), (31) appears in the same form as that proposed in [8], the statistics used to estimate the parameters of the suppression rule is vastly different. First, as indicated by (30), the suppression rule is applied to the power spectral domain instead of the amplitude spectral domain as in [8]. In fact, applying (30) to $\sqrt{m_y(b)}$ (i.e., convert $m_y(b)$ to the amplitude spectral domain) would invalidate the derivations starting from (17) and leads to poor noise reduction performance as verified by our experiments. Second, the *a priori* and *a posteriori* SNRs defined in (27) and (28) are different from those defined in [7] and [8]. In our algorithm, they have to be adjusted to include not only the noise (in the power spectral domain) variance but also the additional variance $\sigma_\varphi^2(b)$ resulting from instantaneous phase differences between the clean speech and the mixing noise (the same quantity as treated in [4]) at the frame-wise DFT level. Note that the variances derived in this way are still underestimated as indicated in (19), and thus our noise suppressor is not a strict MMSE one. An underestimated variance $\sigma_d^2(b)$ would normally cause overestimation of *a-priori* SNR $\xi(b)$ and thus provide less suppression than is desired. Experiments in [7] have shown that using an overestimated $\xi(b)$ is more appropriate than using an underestimate, and experiments in [16] have shown that the optimal suppression rule tends to suppress less than the true log-MMSE suppressor [8] due to noise estimation errors.

IV. Estimation of Parameters

To apply the noise reduction algorithm (32), we need to estimate the noise variance $\sigma_n^2(b)$, the variance $\sigma_\varphi^2(b)$ introduced by the speech-noise phase differences, and the clean-speech variance $\sigma_x^2(b)$ (or equivalently the *a-priori* SNR $\xi(b)$). We discuss these estimates in this section.

A. Estimation of $\sigma_n^2(b)$

In our current implementation, the noise variance $\sigma_n^2(b)$ is estimated using a minimum-controlled recursive moving- average noise tracker similar to the one described in [1]. Briefly, a decision on whether a frame contains speech is made based on the energy ratio test

$$\frac{|\dddot{m}_y(b)|_t^2}{|\dddot{m}_n(b)|_{\min}^2} > \vartheta \tag{33}$$

where ϑ is the threshold, $|\dddot{m}_n(b)|_{\min}^2$ is the smoothed (across filter bank channels and time) minimum noise power within a sliding window which can be tracked efficiently, and $|\dddot{m}_y(b)|_t^2$ is the smoothed (using adjacent channels) power of the bth filter's output (which is in the power domain by itself) at the tth frame. If the energy ratio test is true the frame is assumed to contain speech and the new estimate of the noise variance becomes

$$\sigma_n^2(b)_t = \sigma_n^2(b)_{t-1}. \tag{34}$$

Otherwise, the noise variance is estimated as

$$\sigma_n^2(b)_t = \alpha\sigma_n^2(b)_{t-1} + (1-\alpha)|m_y(b)|_t^2 \tag{35}$$

using the smoothing factor α.

B. Estimation of $\sigma_x^2(b)$

In our current implementation $\sigma_x^2(b)$ is estimated using the same decision-directed approach as that described in [7]. That is, $\sigma_x^2(b)$ for the current frame is estimated using the estimated clean speech from the previous frame and smoothed over the past frames. The reason to use the decision-directed approach instead of other approaches such as maximum-likelihood (ML) estimation is that decision-directed approach has been proven to perform better than the ML-based approaches when combined with MMSE or Wiener filter noise suppressor [7].

C. Estimation of $\sigma_\varphi^2(b)$

According to the definition $\sigma_\varphi^2(b)$ can be computed by

$$\begin{aligned}\sigma_\varphi^2(b) &= E\left\{\left(\sum_f 2|X(f)||N(f)|\cos\varphi(f)w_b(f)\right)^2\right\}\\ &= 4\sum_f E\left\{(|X(f)||N(f)|\cos\varphi(f)w_b(f))^2\right\}\\ &= 4\sum_f E\{|X(f)||N(f)|w_b(f)\}^2\\ &\quad\times\int_0^{2\pi}\frac{1}{2\pi}\cos^2\varphi(f)d\varphi(f)\\ &= 2\sum_f E\{|X(f)||N(f)|w_b(f)\}^2\\ &= 2\sum_f w_b^2(f)E\{|X(f)|\}^2E\{|N(f)|\}^2\end{aligned}$$

where we have used the assumption that the clean speech and noise are independent and the phase difference $\varphi(f)$ between clean speech and mixing noise at the DFT level is uniformly distributed. This approach, however, requires the availability of each DFT bin's estimated speech and noise statistics, which needs to be tracked with a high computational cost. Since we only estimate and keep track of the statistics at the real-valued filter bank's output, we approximate $\sigma_\varphi^2(b)$ as

$$\begin{aligned}\sigma_\varphi^2(b) &= 2\sum_f w_b^2(f)E\{|X(f)|\}^2E\{|N(f)|\}^2\\ &\cong 2E\{m_x(b)\}E\{m_n(b)\}\frac{\sum_f w_b^2(f)}{\sum_{f1}\sum_{f2}w_b(f1)w_bf(2)}\\ &\cong 2\frac{E\{m_x(b)\}}{E\{m_n(b)\}}E\left\{m_n^2(b)\right\}\frac{\sum_f w_b^2(f)}{(\sum_f w_b(f))^2}\\ &\cong 2\frac{\sum_f w_b^2(f)}{(\sum_f w_b(f))^2}\sqrt{\sigma_x^2(b)\sigma_n^2(b)}\end{aligned} \tag{36}$$

in our current implementation. Note that (36) depends on $\sigma_n^2(b)$ and $\sigma_x^2(b)$. Therefore, $\sigma_\varphi^2(b)$ needs to be estimated after estimating $\sigma_n^2(b)$ and $\sigma_x^2(b)$ as described earlier in this section.

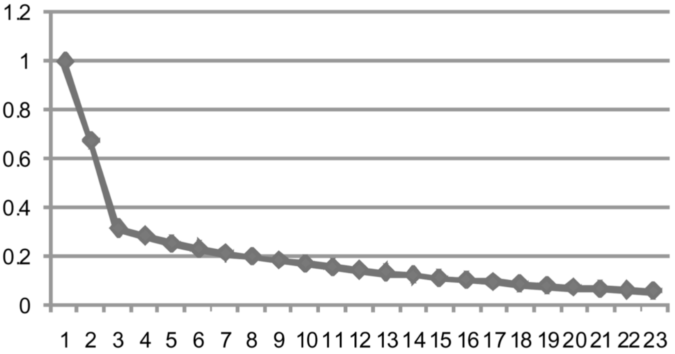


Fig. 4. Precalculated scale weights (see text for detail) used to speed up the estimation of $\sigma_\varphi^2(b)$.

In (36), $\sum_f w_b^2(f)/(\sum_f w_b(f))^2$ is a fixed value for the same Mel-frequency filter bank, and hence it can be precalculated and stored for saving the computational cost. Fig. 4 illustrates the values used for the Mel-frequency filter bank in our experiments. In this figure, the x-axis indicates the Mel-frequency filter bank channel ID and the y-axis indicates the precalculated $\sum_f w_b^2(f)/(\sum_f w_b(f))^2$

V. Performance Evaluation

We have conducted extensive speech recognition experiments on the standard Aurora-3 task [6], [10], [14] to evaluate the performance of the nonlinear MMSE noise reduction algorithm on MFCC described in Sections II–IV.

A. Experimental Setup

The Aurora-3 task consists of noisy digit recognition subtasks under realistic automobile environments [10], [15]. In the Aurora-3 corpus, each utterance is labeled as coming from either a high, low, or quiet noise environment, and as being recorded using a close-talk microphone or a hands-free, far-field microphone.

Based on the languages, the task can be classified into Finnish, Spanish, German, and Danish digit recognition subtasks. For each language, three standard experimental settings are defined for the evaluation.

Well-matched—Both the training and the testing set contain all combinations of noise environments and microphones.

Mid-mismatch—The training set contains quiet and low noise data recorded using the far-field microphone, and the testing set contains the high noisy data recorded using the far-field microphone. The mismatch is mainly caused by the noise.

High-mismatch—The training set contains close-talk data from all noise classes, and the testing set contains high-noise and low-noise far-field data for testing. The mismatch is mainly caused by the channel distortion.

To better understand the performance of our algorithm when evaluated using clean speech data, we have also constructed a test set that consists of only the utterances recorded under the quiet conditions. This test set is evaluated against all three settings mentioned above.

All speech recognition results reported in this section use the HMMs trained in the manner prescribed by the scripts included

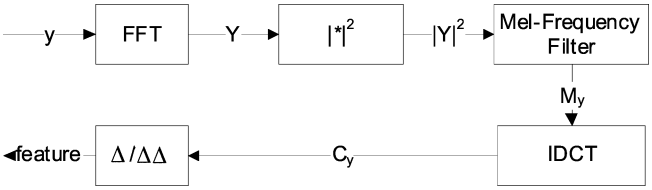


Fig. 5. Feature extraction pipeline for the ICSLP02 baseline system.

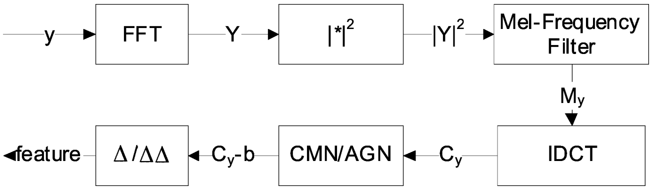


Fig. 6. Feature extraction pipeline for the CMN baseline system.

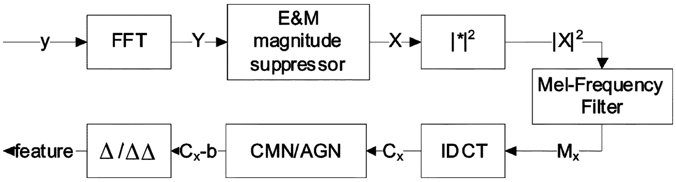


Fig. 7. Feature extraction pipeline for the E&M log-MMSE system [8], where the suppressor is applied to the DFT bins.

TABLE I
SUMMARY OF ABSOLUTE WER ON THE STANDARD TEST SETS IN THE AURORA-3 TASK UNDER DIFFERENT EXPERIMENTAL SETTINGS

Summary of Aurora 3 Absolute Word Error Rate (Standard Set)				
	Well	**Mid**	**High**	**Average**
ICSLP02 Baseline	8.96%	21.96%	48.85%	**23.48%**
CMN	6.87%	16.52%	31.11%	**16.31%**
FB Output Magnitude	6.87%	15.21%	31.29%	**15.89%**
E&M log-MMSE	5.57%	12.79%	29.23%	**14.01%**
MFCC-MMSE	5.08%	12.26%	23.26%	**12.13%**

TABLE II
SUMMARY OF RELATIVE WER REDUCTION ON THE STANDARD TEST SETS IN THE AURORA-3 TASK UNDER DIFFERENT EXPERIMENTAL SETTINGS

Summary of Aurora 3 Relative Improvement (Standard Set)			
Relative to →	**ICSLP02 Baseline**	**CMN**	**E&M log-MMSE**
CMN	30.55%	--	--
E&M log-MMSE	40.33%	14.08%	--
MFCC-MMSE	48.33%	25.59%	13.41%

TABLE III
DETAILED AURORA-3 ABSOLUTE WER RESULTS ON THE STANDARD TEST SETS UNDER THE MFCC-MMSE EXPERIMENTAL SETTING

Aurora-3 Word Error Rate with MFCC-MMSE (Standard Set)					
	Finnish	**Spanish**	**German**	**Danish**	**Average**
Well (x40%)	3.54%	5.90%	5.20%	5.66%	**5.08%**
Mid (x35%)	15.12%	5.39%	10.67%	17.84%	**12.26%**
High (x25%)	17.99%	34.77%	10.78%	29.49%	**23.26%**
Overall	**11.21%**	**12.94%**	**8.51%**	**15.88%**	**12.13%**

TABLE IV
DETAILED AURORA-3 WER REDUCTION RESULTS ON THE STANDARD TEST SETS AGAINST THE ICSLP02 BASELINE UNDER THE MFCC-MMSE

Aurora-3 Relative Improvement with MFCC-MMSE (Standard Set)					
	Finnish	**Spanish**	**German**	**Danish**	**Average**
Well (x40%)	51.24%	16.43%	40.91%	55.50%	**43.36%**
Mid (x35%)	22.42%	67.71%	43.72%	45.41%	**44.18%**
High (x25%)	69.75%	28.24%	59.82%	51.36%	**52.39%**
Overall	**54.44%**	**37.73%**	**49.54%**	**49.88%**	**48.32%**

with the Aurora-3 task. The HMMs used consist of 16-state whole-word models for each digit in addition to the "sil" and "sp" models. The 39-dimension features used in our experiments contain the 13-dimension (with energy and without C0) static MFCC features and their delta and delta-delta features. The parameters (such as smoothing factors and the size of the minimum tracking windows) used for noise tracking are similar to those described in [1] and have not been tuned in the experiments reported in this section. More specifically, the threshold ϑ was set to 5, the noise tracking window was set to 100 frames (or 1 s), and the noise power smooth parameter α was set to 0.9. The smooth parameter used in the decision-directed approach to estimate the clean-speech variance $\sigma_x^2(b)$ and the *a priori* SNR was set to 0.8.

B. Experimental Results

The purpose of our experiments is to examine to what extent our new algorithm is effective for its designed purpose: noise robustness under the additive noise environment. With this goal in mind, we have conducted a series of experiments to compare our algorithm with other noise-robust algorithms such as the conventional E&M log-MMSE magnitude spectral suppressor (operates on the DFT domain), the ETSI AFE, and the SPLICE.

In all the five sets of the results reported in this section, the ICSLP02 baseline refers to the baseline system using the standard WI007 front-end [10], [15], whose feature extraction information flow is shown in Fig. 5. The CMN baseline is the system with the WI007 front-end and a standard active gain normalization [12] and cepstral mean normalization algorithm (Fig. 6). In the FB Output Magnitude technique, we show the results of applying the E&M log-MMSE noise suppressor directly to the magnitude spectrum of the Mel-frequency filter bank output (the information flow was shown in Fig. 1). In both the MFCC-MMSE and the E&M log-MMSE systems, we apply the noise suppression algorithms on top of the CMN baseline system (Figs. 7 and 8).

Tables I and II summarize the average absolute recognition word error rate (WER) results and the relative improvements on the standard test sets for five different experimental system settings, respectively. We observe that our proposed approach has achieved 48.33% WER reduction relative to the ICSLP02 baseline system, 25.59% WER reduction over the CMN baseline system, and 13.41% WER reduction over the conventional E&M log-MMSE algorithm yet with significantly less computation (23 versus 256 vector-component estimates). We can also see that directly applying the E&M log-MMSE noise suppressor to the amplitude spectrum of the Mel-frequency filter bank output (the FB Output Magnitude setting) gives us only slight gain over the CMN baseline. Detailed results on each subtasks of our MFCC-MMSE noise suppressor on the standard test sets are reported in Tables III and IV.

TABLE V
SUMMARY OF ABSOLUTE WER ON THE QUIET TEST SET IN THE AURORA-3 TASK UNDER DIFFERENT EXPERIMENTAL SETTINGS

Summary of Aurora 3 Absolute Word Error Rate (Quiet Set)				
	Well	Mid	High	Average
CMN	3.84%	5.57%	4.75%	**4.67%**
FB Output Magnitude	3.85%	6.42%	5.73%	**5.22%**
E&M log-MMSE	2.97%	4.25%	4.19%	**3.72%**
MFCC-MMSE	3.20%	4.38%	3.38%	**3.66%**

TABLE VI
SUMMARY OF RELATIVE WER REDUCTION ON THE QUIET TEST SET IN THE AURORA-3 TASK UNDER DIFFERENT EXPERIMENTAL SETTINGS

Summary of Aurora 3 Relative Improvement (Quiet Set)		
Relative to ->	CMN	E&M log-MMSE
E&M log-MMSE	20.33%	--
MFCC-MMSE	21.72%	1.75%

TABLE VII
COMPARISON BETWEEN THE MFCC-MMSE SYSTEM AND THE ETSI'S AFE ON THE AURORA-3 TASK

Compare with AFE on Aurora 3 (Standard Set)			
	Well	Mid	High
ETSI AFE	4.70%	13.21%	12.75%
MFCC-MMSE	5.08%	12.26%	23.26%

Tables V and VI summarize the average absolute recognition word error rate (WER) results and the relative improvements on the "quiet" portion of the full test sets. It can be seen that our proposed approach has achieved 21.72% WER reduction over the CMN baseline system, and is slightly better than the E&M log-MMSE algorithm with significantly less computation.

To further understand the effectiveness of the new algorithm, we have evaluated its performance against ETSI AFE [14] and SPLICE [2], [6].

The ETSI AFE is the standard advanced front end adopted by the ETSI. It is a package that includes signal preprocessing, noise estimation, two-pass Wiener filter-based noise suppression, and blind feature equalization. Table VII shows that our proposed approach has comparable performance compared to the ETSI AFE on the well-matched and mid-mismatched settings where noise distortion is the dominant cause of the mismatch. In fact, if we only count errors under these two conditions, the MFCC-MMSE achieved 8.43% WER, which is slightly better than that achieved by the ETSI AFE (which is 8.67% WER). Our approach, however, performs worse than the ETSI AFE system under the high-mismatched setting. This is attributed mainly to the fact that the distortion in the high-mismatched setting is largely caused by channel distortion, which is only handled in our system by the simple CMN method but which was much more carefully handled by the ETSI AFE [14].

SPLICE is a general framework used to model and remove the effect of any consistent degradation of speech cepstra. SPLICE learns a joint probability distribution of noisy and clean cepstra, and uses this distribution to infer clean speech estimates from noisy inputs. SPLICE does not include any assumptions about how noisy cepstra are produced from clean cepstra, and can model any combination of noise and channel distortions. Prior to the work presented in this paper, SPLICE was a standard noise-robust technique in our ASR system [2], [6].

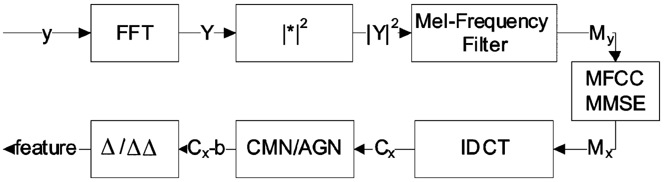


Fig. 8. Feature extraction pipeline for the MFCC-MMSE system.

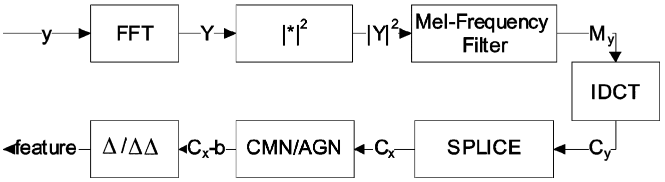


Fig. 9. Feature extraction pipeline for the SPLICE systems.

TABLE VIII
COMPARISON BETWEEN THE MFCC-MMSE SYSTEM AND THE SPLICE ON AURORA-3 WHERE THE SPLICE CODE BOOK WAS TRAINED USING ADDITIONAL INFORMATION TO MAKE A MATCHING CONDITION

Comparisons with SPLICE on Aurora-3 (Standard Set)			
	Well	Mid	High
SPLICE	5.49%	13.55%	11.42%
MFCC-MMSE	5.08%	12.26%	23.26%

The results in Table VIII compare the performance of our new approach with SPLICE on the Aurora-3 task. To make a fair comparison, the SPLICE noise reduction algorithm is applied upon the same CMN baseline system as depicted in Figs. 6 and 9. In this way, the only difference between the MFCC MMSE feature extraction pipeline and the SPLICE feature extraction pipeline lies in the noise robustness techniques used. Note that in this experiment the SPLICE codebook is trained using all combinations of the noise environments and microphones as reported in [6]. That is, the codebook was trained using additional information so that the matched condition is established. This gives an upper bound performance of the SPLICE system.

Table VIII shows that our proposed approach outperforms the SPLICE system by 7.56% and 9.57% on the well-matched and mid-mismatched settings, respectively, where noise distortion is the dominant cause of the mismatch, even though additional information has been used by SPLICE to train the code book in the mid-mismatched setting. Again, our approach performs worse than the SPLICE system under the high-mismatched setting. This is attributed mainly to the fact that the distortion in the high-mismatched setting is largely caused by channel distortion, which is handled by the SPLICE automatically by design if the matched training data are available (which is the case in our setting since SPLICE has seen the far-field microphone data when building the codebook).

VI. SUMMARY AND CONCLUSION

In this paper, we have described a new nonlinear noise reduction algorithm motivated by the MMSE criterion in the MFCC domain for environment-robustness ASR. We have described the algorithm and the parameter estimation methods, showed the differences between our algorithm and the conventional E&M

log-MMSE noise suppressor, and demonstrated its effectiveness in the standard Aurora-3 task.

Our new approach has several key attributes. First, it does not require a codebook to be constructed using training data; hence, it is highly robust to general unseen acoustic environments and it is easy to deploy. Second, it is computationally efficient compared with the conventional E&M log-MMSE noise suppressor since the number of the channels in the Mel-frequency filter bank is usually much smaller (23 in our case) than the number of bins in the FFT domain (256). It introduces no additional look-ahead frame delay. Third, it is designed to apply to filter bank's outputs and hence can be easily plugged into the feature extraction pipeline of many commonly used ASR systems. The proposed approach as developed so far, however, only deals with additive noises and has not been developed to handle channel distortions. Our current work involves expanding on this capability. We are also investigating the combination of the current algorithm, which does not rely on any data, with the data-driven approach (as exploited in SPLICE) to take advantage of the mutual strengths.

ACKNOWLEDGMENT

The authors would like to thank Dr. J. Stokes at Microsoft Research for valuable discussions.

REFERENCES

[1] G. I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[2] L. Deng, J. Droppo, and A. Acero, "Recursive estimation of non-stationary noise using iterative stochastic approximation for robust speech recognition," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 6, pp. 568–580, Nov. 2003.

[3] L. Deng, J. Droppo, and A. Acero, "Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 3, pp. 218–233, May 2004.

[4] L. Deng, J. Droppo, and A. Acero, "Enhancement of log-spectra of speech using a phase-sensitive model of the acoustic environment," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 2, pp. 133–143, Mar. 2004.

[5] L. Deng, "Processing of acoustic signals in a cochlear model incorporating laterally coupled suppressive elements," *Neural Netw.*, vol. 5, pp. 19–34, 1992.

[6] J. Droppo, L. Deng, and A. Acero, "Evaluation of SPLICE on the Aurora 2 and 3 tasks," in *Proc. Int. Conf. Spoken Lang. Process.*, Denver, CO, Sep. 2002, pp. 29–32.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[8] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[9] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 7th ed. New York: Academic, 2007.

[10] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR*, 2000.

[11] Y. Hu and P. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. ICASSP*, 2006, vol. I, pp. 153–156.

[12] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing*. Upper Saddle River, NJ: Prentice-Hall, 2001.

[13] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007.

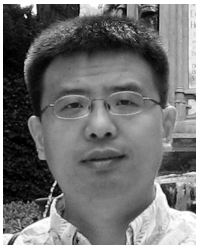
[14] D. Machola, L. Mauuary, B. Noé, Y. -M. Cheng, D. Ealey, D. Jouve, H. Kelleher, D. Pearce, and F. Saadoun, "Evaluation of a noise-robust DSR front-end on Aurora databases," in *Proc. Interspeech*, 2002, pp. 17–20.

[15] A. Peinado and J. Segura, *Speech Recognition Over Digital Channels—Robustness and Standards*. West Sussex, U.K.: Wiley, 2006.

[16] I. Tashev, J. Droppo, and A. Acero, "Suppression rule for speech recognition friendly noise suppressors," in *Proc. ICDSPA*, 2006.

[17] A. van den Bos, "The multivariate complex normal distribution—A generalization," *IEEE Trans. Inf. Theory*, vol. 41, no. 2, pp. 537–538, Mar. 1995.

[18] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "A minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition," in *ICASSP'08*, Las Vegas, NV, pp. 4041–4044.



Dong Yu (M'97–SM'06) received the B.S. degree (with honors) in electrical engineering from Zhejiang University, Hangzhou, China, the M.S. degree (with a presidential award) in electrical engineering from the Chinese Academy of Sciences, Beijing, the M.S. degree in computer science from Indiana University, Bloomington, and the Ph.D. degree in computer science from University of Idaho, Moscow.

He joined Microsoft Corporation, Redmond, WA, in 1998 and the Microsoft Speech Research Group in 2002, where he is currently a Researcher. His current research interests include speech and audio signal processing, robust speech recognition, discriminative training for speech recognition, speech-centric multimodal dialog systems, voice search, machine learning, and pattern recognition. He has published over 50 book chapters, journals, and conference papers in these areas, and is the inventor/co-inventor of dozens of awarded and pending patents.

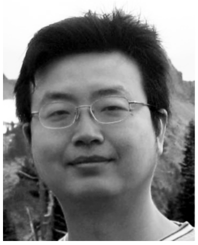
Li Deng (M'86–SM'91–F'05) received the Ph.D. degree in electrical engineering from the University of Wisconsin-Madison.

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 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 the ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan. In 1999, he joined Microsoft Research, Redmond, WA, as a Senior Researcher, where he is currently Principal Researcher. He is also an Affiliate 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 250 refereed papers in leading international conferences and journals, 12 book chapters, and has given keynotes, tutorials, and lectures worldwide. He has been granted over 25 U.S. or International patents in acoustics, speech, and language technology, and signal processing. He coauthored the book *Speech Processing—A Dynamic and Optimization-Oriented Approach* (Marcel Dekker, 2003), and authored the book *Dynamic Speech Models—Theory, Algorithms, and Applications* (Morgan & Claypool 2006).

Dr. Deng served on the Education Committee and Speech Processing Technical Committee of the IEEE Signal Processing Society from 1996 to 2000 and was an Associate Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING from 2002 to 2005. He currently serves on the Society's Multimedia Signal Processing Technical Committee, on the Editorial Board of IEEE SIGNAL PROCESSING LETTERS, as Area Editor for IEEE *Signal Processing Magazine* and as Editor-In-Chief Elect for the same magazine. He was a Technical Chair of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04), and was the General Chair of the IEEE Workshop on Multimedia Signal Processing in 2006. He is Fellow of the Acoustical Society of America.

Jasha Droppo (M'03–SM'07) 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 the summer of 2000. His core interest is speech enhancement for automatic speech recognition, including the SPLICE algorithm, several techniques for model-based speech feature enhancement, and algorithms for learning nonparametric feature space warpings. His other interests include techniques for acoustic modeling, pitch tracking, multiple stream ASR, novel speech recognition features, multimodal interfaces, and cepstral compression and transport.



Jian Wu (M'05) received the B.S. and M.S. degrees in computer science from Tsinghua University, Beijing, China, in 1998 and 2000, respectively, and the Ph.D. degree in computer science from the University of Hong Kong, Hong Kong, China, in 2004.

From 1997 to 2000, he was with the Center of Speech Technology, Tsinghua University. From 2000 to 2004, he was a member of the Human–Machine Communication Laboratory, University of Hong Kong. In 2004, he joined the Speech Component Group, Microsoft Corporation, Redmond, WA, working on speech recognition for desktop, telephony server, and other devices. His current research interests include speech recognition in adverse acoustical environments, acoustic modeling, front-end processing, and machine learning for speech recognition.

Yifan Gong (SM'93) received the B.Sc. degree from the Department of Communication Engineering, Southeast University, Nanjing, China, the M.Sc. degree in electrical engineering and instrumentation from the Department of Electronics, University of Paris, Paris, France, and the Ph.D. degree (with highest honors) in computer science from the Department of Mathematics and Computer Science, University of Henri Poincaré, Nancy, France.

He served the National Scientific Research Center (CNRS), Paris, and INRIA-Lorraine,Villerslès-Nancy, France, as Research Engineer and then joined CNRS as Senior Research Scientist. As Associate Lecturer, he taught computer programming and digital signal processing at the Department of Computer Science, University of Henri Poincaré. He also worked as a Visiting Research Fellow at the Communications Research Center of Canada. He worked for Texas Instruments as a Senior Member of Technical Staff at the Speech Technologies Laboratory. He developed speech modeling technologies robust against noisy environments, designed systems, algorithms, and software for speech and speaker recognition, and delivered memory- and CPU-efficient speech recognizers for mobile devices.

His research interests included mathematical models, software tools, and systems for signal processing, speech and speaker recognition, speech recognition in noisy conditions, and pattern recognition. He is currently a Senior Development Lead at Microsoft Corporation, Redmond, WA, leading a technology and software development team on speech technology and modeling. His current interests include developing technologies and speech recognition models for improved speech recognition performance across multiple languages for desktop, telephony, and mobile devices. He has authored over 100 publications in journals, IEEE Transactions, books, and conferences. His inventions have been awarded 18 U.S. patents. He is an Associate Editor of the *Pattern Recognition Journal*. He has been selected to give tutorials and other invited presentations in international conferences. He has served as a member of technical committees and session chairs for many international conferences.

Dr. Gong served on the IEEE Signal Processing Society Speech Technical Committee from 1998 to 2002.

Alex Acero (S'85–M'90–SM'00–F'04) received the M.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 worked in Apple Computer's Advanced Technology Group from 1990 to 1991. In 1992, he joined Telefonica $I + D$, Madrid, as a Manager of the Speech Technology Group. In 1994, he joined Microsoft Research, Redmond, WA, where he became a Senior Researcher in 1996 and Manager of the Speech Research Group in 2000. Since 2005, he has been a Research Area Manager directing an organization with over 60 engineers conducting research in speech technology, natural language, computer vision, communication, and multimedia collaboration. He is currently an Affiliate Professor of Electrical Engineering at the University of Washington, Seattle. He is author of the books *Acoustical and Environmental Robustness in Automatic Speech Recognition* (Kluwer, 1993) and *Spoken Language Processing* (Prentice-Hall, 2001), has written invited chapters in four edited books and over 150 technical papers. He holds 35 U.S. patents. His research interests include speech and audio processing, natural language processing, image understanding, multimedia signal processing, and multimodal human–computer interaction.

Dr. Acero has served the IEEE Signal Processing Society as Vice President Technical Directions (2007–2009), 2006 Distinguished Lecturer, member of the Board of Governors (2003–2005), Associate Editor for the IEEE SIGNAL PROCESSING LETTERS (2003–2005) and the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (2005–2007), and member of the editorial board of the IEEE JOURNAL ON SELECTED TOPICS IN SIGNAL PROCESSING (2006–2008) and the IEEE SIGNAL PROCESSING MAGAZINE (2008–2010). He also served as member (1996–2002) and Chair (2000–2002) of the Speech Technical Committee of the IEEE Signal Processing Society. He was Publications Chair of ICASSP'98, Sponsorship Chair of the 1999 IEEE Workshop on Automatic Speech Recognition and Understanding, and General Co-Chair of the 2001 IEEE Workshop on Automatic Speech Recognition and Understanding. He is member of the editorial board of *Computer Speech and Language*.