Speech Enhancement Using a Mixture-Maximum Model

David Burshtein, *Senior Member, IEEE,* and Sharon Gannot, *Member, IEEE*

***Abstract*—We present a spectral domain, speech enhancement algorithm. The new algorithm is based on a mixture model for the short time spectrum of the clean speech signal, and on a maximum assumption in the production of the noisy speech spectrum. In the past this model was used in the context of noise robust speech recognition. In this paper we show that this model is also effective for improving the quality of speech signals corrupted by additive noise. The computational requirements of the algorithm can be significantly reduced, essentially without paying performance penalties, by incorporating a dual codebook scheme with tied variances. Experiments, using recorded speech signals and actual noise sources, show that in spite of its low computational requirements, the algorithm shows improved performance compared to alternative speech enhancement algorithms.**

***Index Terms*—Gaussian mixture model, MIXMAX model, speech enhancement.**

I. Introduction

Speech quality and intelligibility might significantly deteriorate in the presence of background noise, especially when the speech signal is subject to subsequent processing, such as speech coding or automatic speech recognition. Consequently, modern communications systems, such as cellular phones, employ some speech enhancement procedure at the preprocessing stage, prior to further processing (e.g., speech coding). Speech enhancement algorithms have therefore attracted a great deal of interest in the past two decades [1]–[14].

Speech enhancement algorithms may be broadly classified as belonging to one of the following two categories. The first is the class of time domain, parametric, model-based methods [6]–[12]. The second class of speech enhancement algorithms is the class of spectral domain algorithms. A subset of this class is the popular spectral subtraction-based algorithms, e.g., [1], [14]. Other spectral domain algorithms include the short time spectral amplitude (STSA) estimator and the log spectral amplitude estimator (LSAE), both proposed by Ephraim and Malah [2], [3], and the hidden Markov model (HMM)-based filtering algorithms proposed by Ephraim *et al.* [4], [5]. In general, the computational requirements of the spectral domain algorithms are lower than the computational requirements of the time domain algorithms. This property makes spectral domain algorithms attractive candidates, especially for low-cost and/or low-power (e.g., battery operated) applications, such as cellular telephony.

The purpose of the paper is to present a spectral domain algorithm, which produces high-quality enhanced speech on the one hand, and has low computational requirements on the other hand. The algorithm is similar to the HMM-based, minimum mean square error (MMSE) filtering algorithm proposed by Ephraim *et al.* [4], [5], in the sense that it also utilizes a Gaussian mixture to model the speech signal. However, while the previous set of algorithms utilize a mixture of auto-regressive models in the time domain, our algorithm models the log-spectrum by a mixture of diagonal covariance Gaussians. In this paper, we follow the MIXMAX approximation, which was originally suggested by Nádas *et al.* [15] in the context of speech recognition, and propose a new speech enhancement algorithm. For this purpose, various modifications, adaptations and improvements were made in the algorithm proposed in [15] in order to make it a high-quality, low-complexity speech enhancement algorithm. In [15], the MIXMAX model is used to design a noise adaptive, discrete density, HMM-based, speech recognition algorithm. In [16], we used the MIXMAX model to design various noise adaptive, continuous density, HMM-based speech recognition systems. In this paper, our approach is more similar to the adaptation algorithm presented in [16], when the feature vector comprises all the elements of the DFT of the frame (instead of the MEL spectrum used in [16]). We also discuss the computational complexity of the new speech enhancement algorithm and show how it can be reduced, essentially with no performance penalties. Our study is supported by extensive speech enhancement experiments using speech signals and various actual noise sources.

The organization of the paper is as follows. In Section II, we review the MIXMAX model that was originally suggested by Nádas *et al.* [15]. In Section III, we apply the MIXMAX model to the speech enhancement problem. In Section IV, we compare the MIXMAX speech enhancement algorithm to alternative enhancement algorithms. The comparison is supported by an experimental study. In Section V, we discuss the computational complexity of the algorithm and show how it can be reduced. Section VI concludes the paper.

II. MIXMAX Model

Let $x[l]$ $l = 0, 1, \ldots, L-1$ be the samples of some speech signal segment (frame), possibly weighted by some window function, and let $X(e^{j2\pi k/L})$ denote the corresponding short time Fourier transform

$$X(e^{j2\pi k/L}) = \sum_{l=0}^{L-1} x[l]e^{-j2\pi lk/L} \qquad k = 0, 1, \ldots, L-1. \tag{1}$$

Manuscript received December 5, 2000; revised April 21, 2002. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dirk van Compernolle.

D. Burshtein is with the Department of Electrical Engineering—Systems, Tel-Aviv University, Tel-Aviv, Israel (e-mail: burstyn@eng.tau.ac.il).

S. Gannot is with the Faculty of Electrical Engineering, Technion—Israel Institute of Technology, Haifa, Israel (e-mail: gannot@siglab.technion.ac.il).

Digital Object Identifier 10.1109/TSA.2002.803420.

1063-6676/02$17.00 © 2002 IEEE

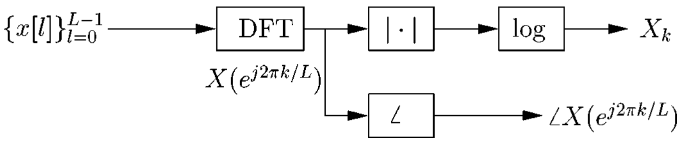

Fig. 1. Front-end signal processing.

Let $\mathbf{X}$ denote the $L/2+1$ dimensional, log-spectral vector with kth component, X_k, defined by

$$X_k = \log|X(e^{j2\pi k/L})| \qquad k = 0, 1, \ldots, K-1$$

where $K = L/2+1$ (X_k $k = L/2+1, \ldots, L-1$ may be obtained using symmetry, i.e., $X_k = X_{L-k}$). The relations between $x[l]$, $|X(e^{j2\pi k/L})|$, $\angle X(e^{j2\pi k/L})$ and X_k are shown in Fig. 1. The most common modeling approach of the log spectral vector, $\mathbf{X}$, is realized by an HMM with a state dependent mixture of diagonal covariance Gaussians. In this paper, a single state model is used. The corresponding probability density function, $f(\mathbf{x})$ [for simplicity, we avoid the more accurate notation, $f_{\mathbf{X}}(\mathbf{x})$], is given by

$$f(\mathbf{x}) = \sum_i c_i f_i(\mathbf{x}) = \sum_i c_i \prod_k f_{i,k}(x_k) \tag{2}$$

where

$$\begin{aligned} f_{i,k}(x) =& \mathcal{N}(x, \mu_{i,k}, \sigma_{i,k}) \\ =& \frac{1}{\sqrt{2\pi}\,\sigma_{i,k}} \exp\left\{-\frac{(x-\mu_{i,k})^2}{2\sigma_{i,k}^2}\right\}. \end{aligned} \tag{3}$$

In order to extend the Gaussian mixture model to the case where the speech signal is contaminated by (a possibly colored) additive noise, Nádas *et al.* [15] proposed the following model. Let $\mathbf{Y}$ and $\mathbf{Z}$ denote the log-spectral vectors of the noise and noisy speech signals, respectively, and let $g(\mathbf{y})$ denote the probability density function of $\mathbf{Y}$. We assume that the noise is statistically independent of the speech signal. In addition both signals have zero mean. For simplicity we also assume that $g(\mathbf{y})$ can be modeled by a single diagonal covariance Gaussian (the extension to a mixture of Gaussians noise density is straightforward), i.e.,

$$g(\mathbf{y}) = \prod_k g_k(y_k)$$

where

$$g_k(y) = \frac{1}{\sqrt{2\pi}\,\sigma_{Y,k}} \exp\left\{-\frac{(y-\mu_{Y,k})^2}{2\sigma_{Y,k}^2}\right\}. \tag{4}$$

Now, $z[l] = x[l] + y[l]$. Due to the statistical independence and zero mean assumptions we thus have

$$|Z(e^{j2\pi k/L})|^2 \approx |X(e^{j2\pi k/L})|^2 + |Y(e^{j2\pi k/L})|^2.$$

Hence

$$Z_k \approx \log(\exp(X_k) + \exp(Y_k)).$$

The assumption in the MIXMAX model, suggested by Nádas *et al.* [15], is that we can further approximate Z_k by $\max(X_k, Y_k)$, that is

$$\mathbf{Z} \approx \max(\mathbf{X}, \mathbf{Y})$$

where the maximum is carried out component-wise over the elements of the log-spectral vectors.

Let $F_{i,k}(x)$, $G_k(y)$ denote the cumulative distribution functions of $f_{i,k}(x)$ and $g_k(y)$, respectively. Note that

$$\begin{aligned} G_k(y) &= \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}\,\sigma_{Y,k}} e^{-(1/2\sigma_{Y,k}^2)(u-\mu_{Y,k})^2}\, du \\ &= \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{y-\mu_{Y,k}}{\sqrt{2}\,\sigma_{Y,k}}\right) \end{aligned} \tag{5}$$

where

$$\operatorname{erf}(u) = \frac{2}{\sqrt{\pi}} \int_0^u e^{-t^2}\, dt$$

is the error function. Similarly

$$F_{i,k}(x) = \frac{1}{2} + \frac{1}{2}\operatorname{erf}\left(\frac{x-\mu_{i,k}}{\sqrt{2}\,\sigma_{i,k}}\right). \tag{6}$$

The cumulative distribution function of Z_k given the ith mixture, $H_{i,k}(z)$, is obtained by invoking the statistical independence of $\mathbf{X}$ and $\mathbf{Y}$ as follows:

$$\begin{aligned} H_{i,k}(z) &= \Pr\{Z_k < z \,|\, I = i\} \\ &= \Pr\{X_k < z, Y_k < z \,|\, I = i\} \\ &= F_{i,k}(z) G_k(z). \end{aligned} \tag{7}$$

Here I is the class (mixture) random variable. The density of Z_k given the ith mixture, $h_{i,k}(z)$, is obtained by differentiating (7), [15]

$$h_{i,k}(z) = f_{i,k}(z) G_k(z) + F_{i,k}(z) g_k(z).$$

The probability density of Z is hence given by

$$\begin{aligned} h(\mathbf{z}) &= \sum_i c_i h_i(\mathbf{z}) = \sum_i c_i \prod_k h_{i,k}(z_k) \\ &= \sum_i c_i \prod_k [f_{i,k}(z_k) G_k(z_k) + F_{i,k}(z_k) g_k(z_k)]. \end{aligned} \tag{8}$$

Nádas *et al.* used a probabilistic rule based on (8) to adapt a discrete density HMM-based speech recognition system in the presence of additive noise. In [16] the MIXMAX model is used in order to adapt other HMM-based speech recognition systems to noise, including systems that use continuous mixture of Gaussians and systems that utilize time derivative (delta) spectral features.

III. Application to Speech Enhancement

In this paper, we apply the MIXMAX model to the related problem of speech enhancement. In order to obtain an estimate,

$\hat{\mathbf{X}}$, to $\mathbf{X}$ given $\mathbf{Z}$, we use the following minimum mean square error (MMSE) estimator:

$$\hat{\mathbf{X}} = \mathrm{E}(\mathbf{X}\,|\,\mathbf{Z}) = \sum_i \hat{\mathbf{X}}_i q(i\,|\,\mathbf{Z}=\mathbf{z}) \tag{9}$$

where $q(i\,|\,\mathbf{Z}=\mathbf{z})$, the class conditioned probability is given by

$$q(i\,|\,\mathbf{Z}=\mathbf{z}) = \frac{c_i h_i(\mathbf{z})}{h(\mathbf{z})} = \frac{c_i h_i(\mathbf{z})}{\sum\limits_j c_j h_j(\mathbf{z})}. \tag{10}$$

$\hat{X}_{i,k}$, the kth component of $\hat{\mathbf{X}}_i$ is the expected value of X_k given the class i and the noisy observation z_k

$$\begin{aligned}\hat{X}_{i,k} &= \mathrm{E}\{X_k\,|\,Z_k=z_k,\, I=i\}\\ &= \int x_k \frac{f_{i,k}(x_k)h_{i,k}(z_k\,|\,X_k=x_k)}{h_{i,k}(z_k)}\,dx_k\end{aligned} \tag{11}$$

where $h_{i,k}(\cdot\,|\,X_k=x_k)$ is the conditional density of Z_k given $I=i$ and $X_k=x_k$. Note that

$$\Pr\{Z_k < z_k\,|\,X_k=x_k\} = \Pr\{Y_k < z_k\}u(z_k - x_k)$$

where $u(\;)$ is the unit step function. Differentiating the last expression with respect to z_k, $h_{i,k}(z_k\,|\,X_k=x_k)$ is obtained. Now, recalling the Gaussian assumption for $f_{i,k}$, and invoking the integration required by (11), we obtain

$$\hat{X}_{i,k} = z_k\rho_{i,k} + (\mu_{i,k} - \sigma_{i,k}^2 R_{i,k})(1-\rho_{i,k}) \tag{12}$$

where

$$\begin{aligned}R_{i,k} &= f_{i,k}(z_k)/F_{i,k}(z_k); \quad R_{Y,k} = g_k(z_k)/G_k(z_k);\\ \rho_{i,k} &= \frac{1}{1+R_{Y,k}/R_{i,k}}.\end{aligned} \tag{13}$$

Our estimate, $\hat{\mathbf{X}}$, is calculated using (9), (10), (12), and (13). In [16] we used $\hat{\mathbf{X}}$ in order to design a noise robust speech recognition system and compared it to alternative noise adaptation methods using the MIXMAX approach. For our present speech enhancement application the reconstructed speech signal, $\hat{x}[l]$, for the current frame is given by

$$\begin{aligned}\hat{x}[l] &= \frac{1}{L}\sum_{l=0}^{L-1} \hat{X}(e^{j2\pi k/L})e^{j2\pi lk/L}\\ \hat{X}(e^{j2\pi k/L}) &= \exp\{\hat{X}_k\}\angle Z(e^{j2\pi k/L}).\end{aligned}$$

Note that the reconstructed phase angle is the original phase angle of the noisy speech, as is usually the case when using spectral-domain enhancement methods [2].

We assume the availability of a voice activity detector (VAD). Based on the VAD indications of voice inactivity periods, we collect noise statistics, continuously and adaptively. Hence, we may assume that the (time varying) probability density of the noise, $\mathbf{Y}$, is known. For each frame we obtain an estimate $\hat{\mathbf{X}}$ to $\mathbf{X}$, based on $\mathbf{Z}$ and on the current density of the noise.

In order to apply the method a mixture model of the type of (2) needs to be trained. Let the training data consist of N log-spectrum frames, $x = (\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N)$. The objective is to set c_i, $\mu_{i,k}$, $\sigma_{i,k}$ so as to maximize the log-likelihood

$$\log f(\mathbf{x}) = \sum_n \log f(\mathbf{x}^n).$$

The maximization may be carried out by using the expectation–maximization (EM) algorithm [17].

Let $\gamma_{n,i}$, and $\alpha_{n,i}$ be defined by

$$\begin{aligned}\gamma_{n,i} &= f(\mathbf{x}^n,\, I_n=i) = c_i f(\mathbf{x}^n\,|\,I_n=i)\\ &= c_i \prod_{k=0}^{K-1} \frac{1}{\sqrt{2\pi}\,\sigma_{i,k}} \exp\left\{-\frac{(x_k^n - \mu_{i,k})^2}{2\sigma_{i,k}^2}\right\}\\ \alpha_{n,i} &= \Pr(I_n=i\,|\,\mathbf{x}^n) = \frac{\gamma_{n,i}}{\sum\limits_{i'=0}^{M-1}\gamma_{n,i'}}.\end{aligned} \tag{14}$$

M is the total number of mixtures. Note that $\alpha_{n,i}$ are the class-conditioned probabilities. Let c_i, $\mu_{i,k}$ and $\sigma_{i,k}^2$ denote the current values of the model parameters, and let $\hat{c}_i$, $\hat{\mu}_{i,k}$, and $\hat{\sigma}_{i,k}^2$ denote the values of the model parameters after the iteration. The EM iteration is given by

$$\hat{c}_i = \frac{\sum\limits_{n=0}^{N-1}\alpha_{n,i}}{N} \qquad i=0,\ldots,M-1 \tag{15}$$

$$\begin{aligned}\hat{\mu}_{i,k} &= \frac{\sum\limits_{n=0}^{N-1}\alpha_{n,i}x_k^n}{\sum\limits_{n=0}^{N-1}\alpha_{n,i}} \qquad i=0,\ldots,M-1\\ &k=0,\ldots,K-1\end{aligned} \tag{16}$$

$$\begin{aligned}\hat{\sigma}_{i,k}^2 &= \frac{\sum\limits_{n=0}^{N-1}\alpha_{n,i}(x_k^n - \hat{\mu}_{i,k})^2}{\sum\limits_{n=0}^{N-1}\alpha_{n,i}} = \frac{\sum\limits_{n=0}^{N-1}\alpha_{n,i}(x_k^n)^2}{\sum\limits_{n=0}^{N-1}\alpha_{n,i}} - \hat{\mu}_{i,k}^2\\ &i=0,\ldots,M-1 \quad k=0,\ldots,K-1\end{aligned} \tag{17}$$

where $\alpha_{n,i}$ are computed using the current values of the parameters, c_i, $\mu_{i,k}$ and $\sigma_{i,k}^2$.

To avoid numerical problems in the calculations, it is recommended to use logarithmic arithmetic [15]. Let $\{v_i\}$ be some given set of real numbers. Then, to evaluate $\log\sum_i e^{v_i}$, we use the following relation:

$$\log\sum_i e^{v_i} = v_{\max} + \log\sum_i e^{v_i - v_{\max}} \tag{18}$$

where $v_{\max} = \max_{1\le i\le N} v_i$. Equation (18) is then used in (8) and (10).

To further improve the subjective quality of the reconstructed speech, we found it useful to apply the nonlinear postprocessing method that was suggested in the past for spectral subtraction [1], [14]. Let $\gamma_k = \exp\{\hat{X}_k - Z_k\}$. γ_k is the spectral gain (in fact, suppression, since $\gamma_k < 1$) of the kth channel. The idea is to constrain γ_k to be above some frequency-dependent threshold, δ_k. That is, the reconstructed speech is now given by

$$\begin{aligned}\tilde{x}[l] &= \frac{1}{L}\sum_{l=0}^{L-1} \tilde{X}(e^{j2\pi k/L})e^{j2\pi lk/L}\\ \tilde{X}(e^{j2\pi k/L}) &= \exp\{\tilde{X}_k\}\angle Z(e^{j2\pi k/L})\\ \tilde{X}_k &= \max(Z_k + \log\delta_k,\, \hat{X}_k).\end{aligned}$$

IV. Comparison With Alternative Speech Enhancement Algorithms

The MIXMAX speech enhancement algorithm is closely related to the HMM-based minimum mean square error (MMSE) speech enhancement algorithm that was proposed by Ephraim*et al.* [4], [5]. Both the HMM MMSE and MIXMAX algorithms use the MMSE criterion and both utilize a Gaussian mixture model for the speech signal. In addition both need a clean speech database in order to train a speech model. However, while the HMM MMSE algorithm employs a mixture of auto-regressive models in the time domain, the MIXMAX enhancement algorithm models the log-spectrum by a mixture of diagonal covariance Gaussians. Both types of mixture models have been suggested for speech recognition systems. However, the time domain auto-regressive mixture yields a somewhat lower recognition rate, at least when the alternative spectral Gaussian mixture model is applied to the cepstrum representation [18]. The later model is thus much more popular in modern speech recognition systems. In fact when training our clean speech model using the auto-regressive spectrum, the quality of the enhanced speech degraded.

Since the HMM MMSE algorithm employs a mixture of auto-regressive models in the time domain, it results in a series of Wiener filters, such that the output signal is a mixture of the signals produced by these filters. Our estimator is based on a Gaussian mixture in the log-spectral domain. In this case the MMSE criterion results in a much more complicated solution. The MIXMAX assumption significantly simplifies the resulting MMSE estimator. As an alternative to the MIXMAX solution, one may use the MMSE estimator proposed in [19]. This estimator is based on a model for the log-spectrum, and is significantly more complicated than our MIXMAX estimator.

We compared the MIXMAX algorithm to the HMM MMSE algorithm using both objective and subjective listening tests. In our implementation of the HMM MMSE algorithm a single HMM state is used. However, in our experience this model is as effective as a multistate HMM, provided that sufficiently many mixtures are used. This is due to the fact that the information provided by temporal acoustic transitions is marginal compared to the spectral information. Consequently, it is sufficient to use a mixture of Gaussians model which assumes independence from one frame to the other. This simplifying assumption is also used by state-of-the-art speaker recognition systems [20]. In fact it is also straight-forward to extend our MIXMAX algorithm to a multistate HMM. In order to compare MIXMAX and HMM MMSE on equal terms, both were implemented using a single state HMM and with varying number of mixtures.

It has been noted in the past [13] that the performance of the simple nonlinear spectral subtraction algorithm proposed by Boll [1] is inferior to the HMM MMSE algorithm. Therefore we do not provide a detailed comparison with Boll's algorithm. For comparison with time-domain algorithms, we used the previously proposed KEM algorithm [6]. Essentially, this algorithm iterates between LPC parameters estimation and Kalman filtering.

To test the performance of the various algorithms we used 50 sentences from the TIMIT database (25 females, 25 males). All sentences were initially down-sampled from 16 KHz to 8 KHz. In order to apply the HMM MMSE and MIXMAX algorithms, it is first necessary to obtain a clean speech model. This was realized by using a set of additional 30 TIMIT sentences (15 females, 15 males). The performance of both algorithms essentially did not change when using a larger database with 50 sentences to train the clean speech model.

The postprocessing modification that was outlined in Section III was applied both for the HMM MMSE and MIXMAX algorithms using

$$\delta_k = \begin{cases} 0.35, & \text{if } 0 \le k \le 36 \\ 0.18, & \text{if } 37 \le k \le 128. \end{cases} \tag{19}$$

In our implementation the frame length is $L = 256$, which corresponds to $K = 129$. Hence δ_k is higher for frequencies lower than 1125 Hz ($k = 36$). As a result, the subjective quality of both algorithms improved significantly. Lower threshold values improved the objective criteria, and in particular the amount of noise reduction, but reduced the subjective quality.

In both algorithms frame overlapping of 50% was used, such that after synthesizing the reconstructed speech, we keep only the $L/2$ output samples that correspond to the center of the frame. The sentences were corrupted by additive noise, using various types of noise signals, including a synthetic white Gaussian noise source, and some noise signals from the NOISEX-92 database [21] resampled to 8 KHz. These include car noise, speech-like noise (synthetic noise with speech-like spectrum), operation room noise and a factory floor noise. The amplitude of the factory noise fluctuates in time periodically, with a period of about 0.5 s. The characteristics of the factory noise signal, as well as the other noise signals from the NOISEX-92 database used throughout this paper, are shown in Fig. 2.

Various SNRs were used in the experiments. We assumed the existence of a reliable VAD. Later we note on this assumption. Hence, prior to speech enhancement we estimated the noise parameters using some independent segment from the noise source. The duration of this segment was set to 250 ms. When using the MIXMAX algorithm, the noise parameters, $\mu_{Y,k}$ and $\sigma_{Y,k}$ are estimated using the standard empirical mean and variance equations. When using the HMM MMSE algorithm, we employed the Blackman–Tukey method for spectrum estimation.

Our objective set of criteria comprises total output SNR, segmental SNR and Itakura–Saito distance measure. These distortion measures are known to be correlated with the subjective perception of speech quality [22].

The total output SNR is defined by

$$\text{SNR} = \frac{\sum_t x^2[t]}{\sum_t (x[t] - \hat{x}[t])^2} \tag{20}$$

where $x[t]$ and $\hat{x}[t]$ are the reference (e.g., clean) and test (e.g., enhanced) speech signals, and where the time summations are over the entire duration of the signals. Prior to the application of

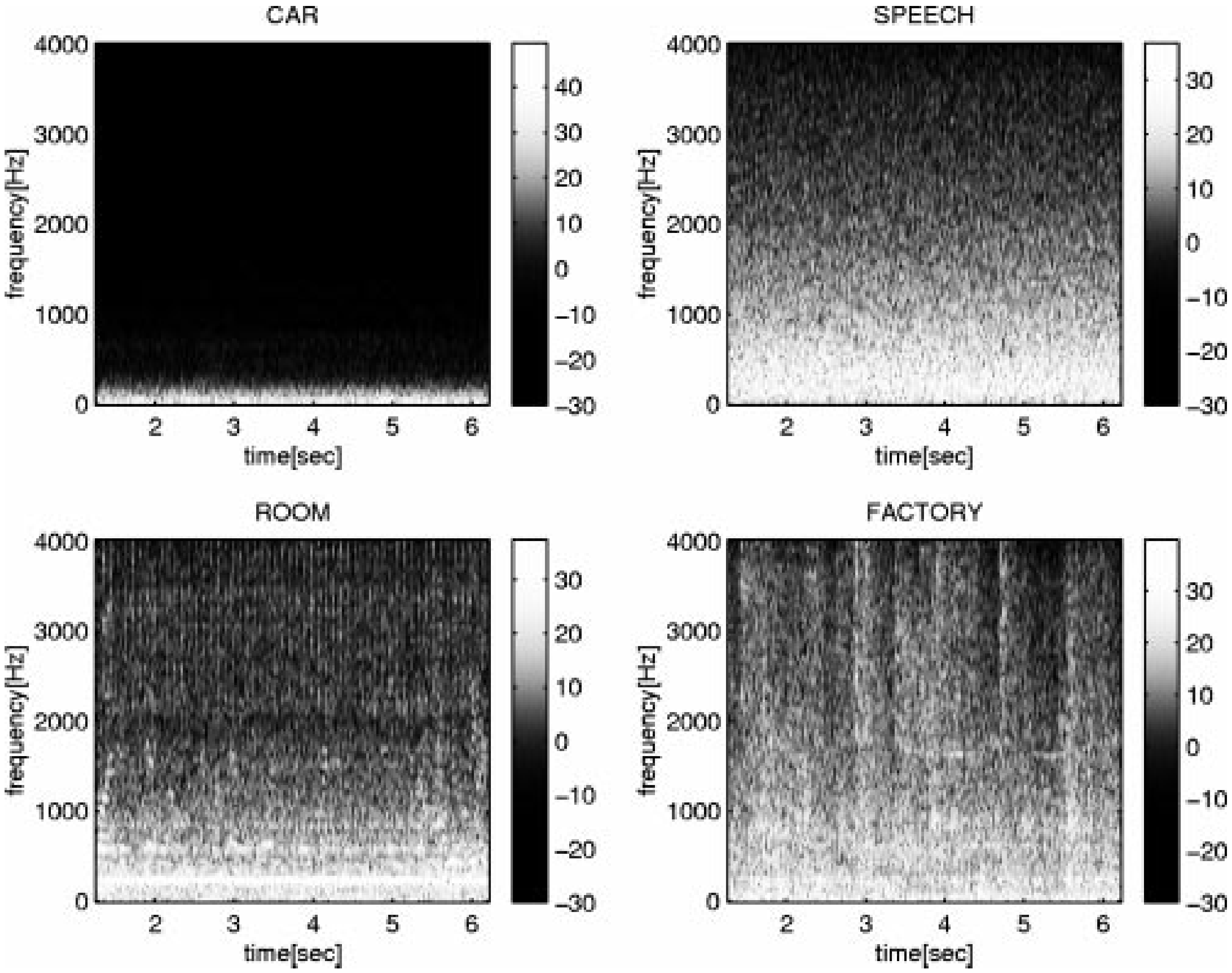

Fig. 2. Sonograms of the car, speech-like, operation room, and factory noise signals.

(20), $x[t]$ and $\hat{x}[t]$ are scaled to have unit energy over the entire sentence.

Segmental SNR is usually defined by the mean value of the individual SNR measurements [using (20)] over the frames of the sentence. Segmental SNR is known to be more strongly correlated with subjective quality, and is similar in that sense to the performance of the Itakura–Saito distance measure [22]. However, total output SNR is more robust to the presence of low energy regions (frames), or to frames for which the energy of $x[t]-\hat{x}[t]$ is small. To increase the robustness of the segmental SNR measure and to eliminate outliers (which are due to the reasons outlined above) we used the median value of the individual SNR measurements instead of using their mean. Likewise, we have modified the standard definition of the Itakura–Saito distance measure by replacing the mean value with median averaging.

Figs. 3 and 4 show the total SNR, segmental SNR and Itakura–Saito (IS) distance measure of the HMM MMSE, MIXMAX, and KEM algorithms, for the case where 20 Gaussian mixtures are used, for a factory noise source and white Gaussian noise, respectively. All three distance measures consistently show an advantage to the MIXMAX algorithm. Similar trend was observed for other noise sources from the NOISEX-92 database [21], including car noise, operation room noise and the speech-like noise. In Figs. 3 and 4, we provide results for the case where postprocessing [(19)] was applied at the output of both the HMM MMSE and MIXMAX algorithms. When postprocessing is not applied the objective criteria tend to improve for both algorithms. However the improvement is usually more significant for the MIXMAX algorithm such that the gap between these algorithms slightly increases. For example, for a factory noise signal and input SNR of 12.5 dB, the output SNR of HMM MMSE is 14.5 dB (same as with postprocessing). The output SNR of MIXMAX is 16.1 dB (15.8 dB when postprocessing is used). When the input SNR is 0.5 dB, the output SNR of HMM MMSE is 5.7 dB (2.4 dB when postprocessing is used), while the output SNR of MIXMAX is 6 dB (2.7 dB when postprocessing is used).

In Fig. 5, we present the sound sonograms of the clean, noisy, HMM MMSE enhanced and MIXMAX enhanced speech, when using an operation room noise source at an SNR level of 9 dB. The reconstructed speech produced by both algorithms is characterized by an almost equal noise reduction. However, the MIXMAX output is less distorted compared to the HMM MMSE output. These results were verified by

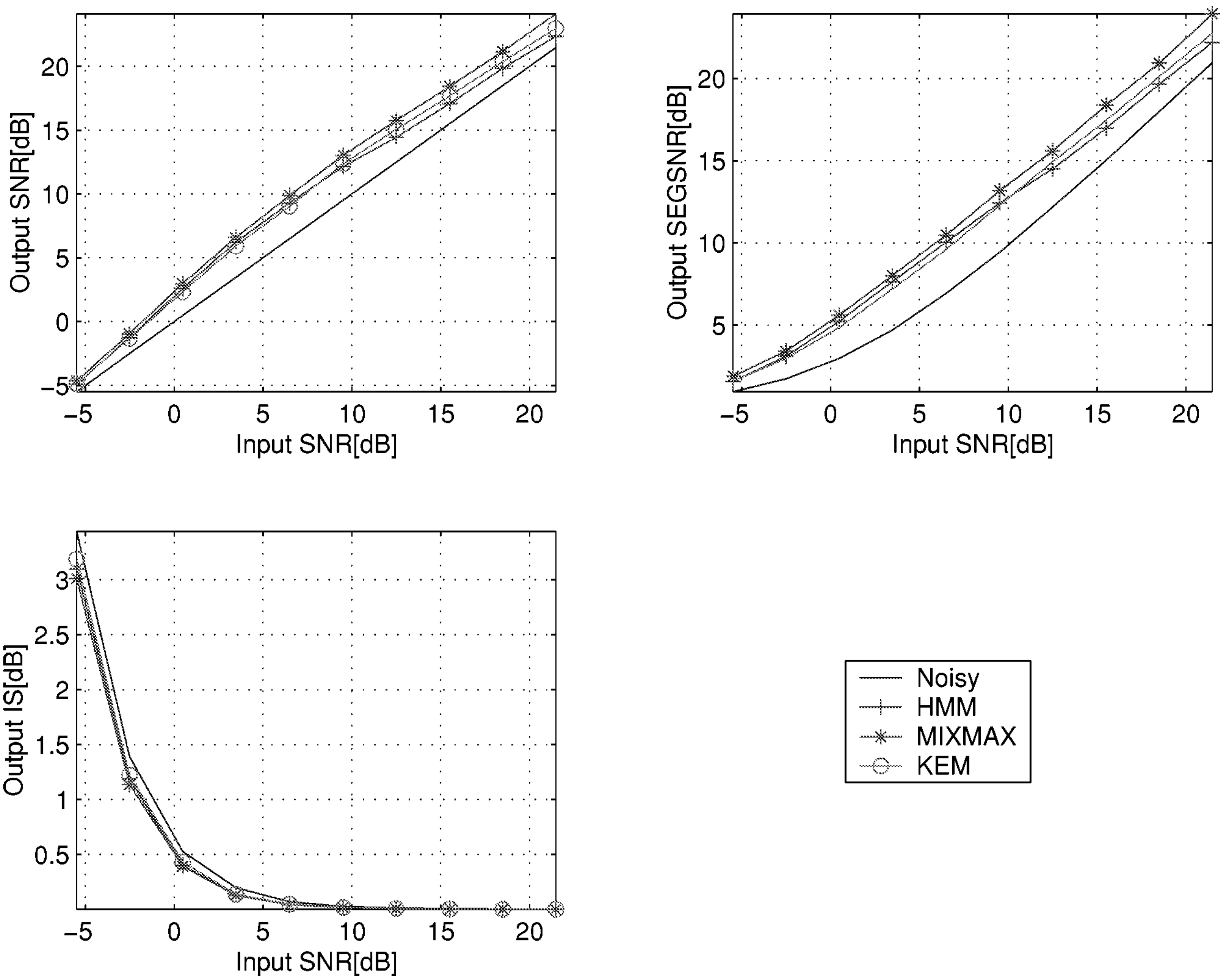

Fig. 3. Comparison between MIXMAX, HMM MMSE, and KEM algorithms (factory noise, 20 mixtures).

informal listening tests using several listeners. Although the noise reduction of MIXMAX and HMM MMSE is about the same, the quality of the enhanced MIXMAX signal is superior to that of HMM MMSE over the entire SNR range examined. In particular, it seems that at low SNRs the MIXMAX output respects the unvoiced part. The distortion of the speech produced by the KEM algorithm is low, but its noise reduction is inferior. Speech samples can be found in [23].

So far we assumed an ideal VAD. In order to test the significance of this assumption we repeated the experiments with a simple energy based VAD. While tested with the factory noise source, the application of the VAD did not impose any significant degradation in performance, both in objective and subjective measures. Note, that while in high SNR levels the simple VAD performance is very good, it might collapse in the low SNR region. However, we found that in this SNR range, any corrupted speech segment might be used by the enhancement algorithm, since the noisy signal is dominated by the noise.

To assess the sensitivity of the various algorithms to channel mismatch, we repeated the experiments for the factory noise summarized in Fig. 3 with the NTIMIT database, which is the same database as TIMIT except that a telephone channel is used (training was performed with the standard TIMIT database). The results of this experiment were essentially the same as those provided in Fig. 3. This shows that in spite of the fact that non of these algorithms considers the effect of the channel, they all seem to be insensitive to channel mismatch.

Our algorithm needs to be trained using some clean speech database. To assess the sensitivity of the algorithm to the language of this database, we tested the enhancement algorithm on Dutch sentences (both male and female) taken from the *Amsterdam Free University* database. First we used the TIMIT database (English) for the training stage (thus, there was a language mismatch between the training and the enhancement stages). In the second experiment, we used Dutch sentences for both the training and enhancement stages. For example, for a background speech noise signal at input SNR of 9.8 dB, the output SNR of the MIXMAX algorithm trained with English database and tested on Dutch sentences was 9.2 dB (degradation) and while trained with Dutch database the output SNR was 11.9 dB. For input SNR of 0.8 dB the output SNR for English training was 1.2 dB and for Dutch training it increased to 2 dB. The HMM MMSE algorithm is more sensitive to language mismatch in terms of the objective criteria. Subjective listening shows that although some degradation due to language mismatch probably exists, it is certainly not significant.

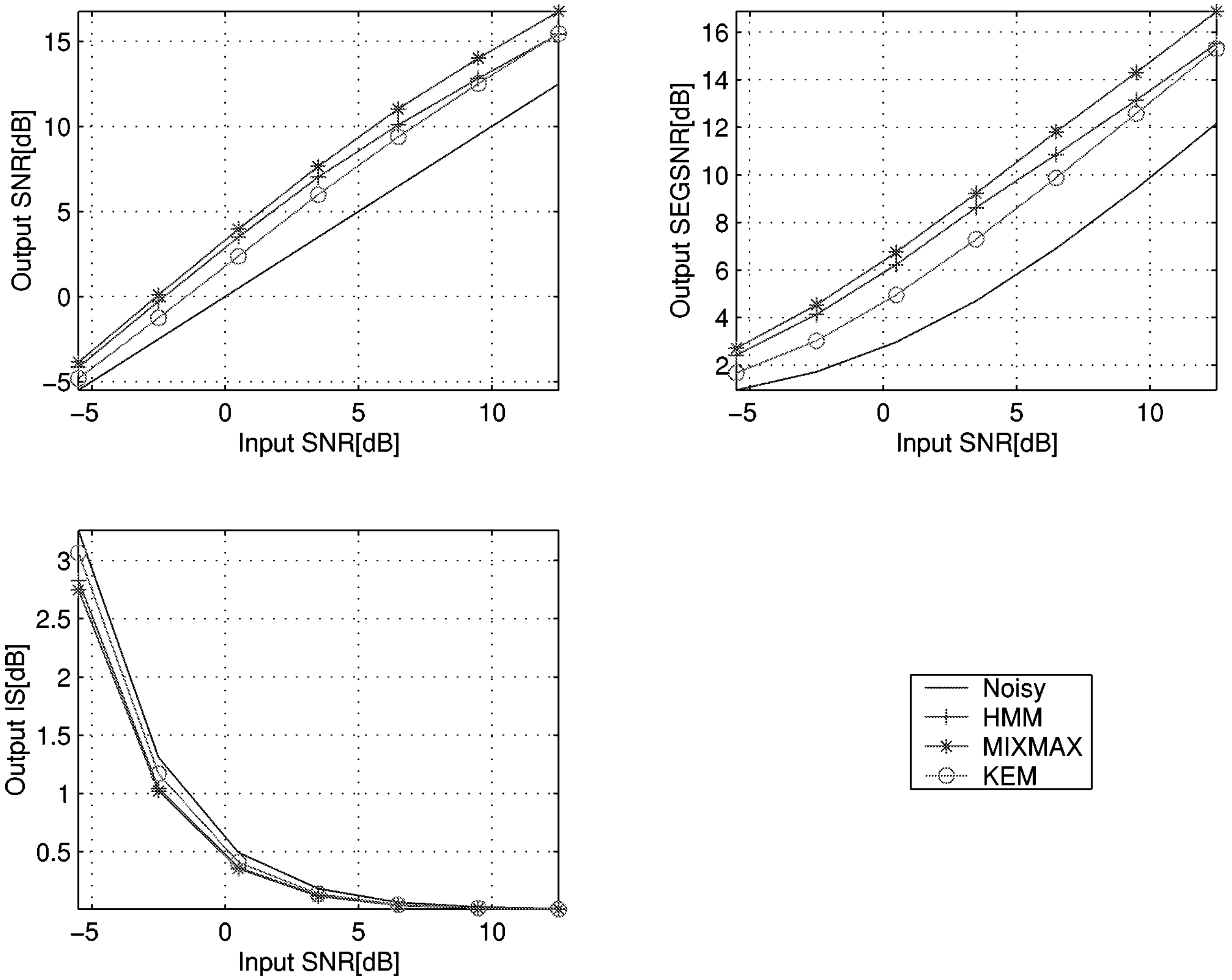

Fig. 4. Comparison between MIXMAX, HMM MMSE, and KEM algorithms (white Gaussian noise, 20 mixtures).

V. Reduced-Complexity MIXMAX Enhancement

In this section, we discuss the complexity of the algorithm and its memory requirements. We then suggest some improvements and simplifications that were found useful.

The algorithm processes the data block-wise, $L/2$ new samples are produced from each input block of size L. The algorithm comprises the following computational stages: spectral analysis, class-conditioned probability calculation, filtering, and synthesis. Under the assumption of K and M sufficiently large, the computational complexity of these stages is as follows.

Spectral Analysis and Synthesis: In the spectral analysis stage, we compute the log-spectrum and phase. The computational complexity is dominated by a DFT of a block of L real numbers. The corresponding number of real multiplications is $2K \log_2 K$, the number of real additions is $3K \log_2 K$. In the spectral synthesis stage, we convert the log-spectrum and phase back to the time domain. The computational complexity is the same as that for the spectral analysis stage.

Class Conditioned Probability Calculation: To compute $q(i \,|\, \mathbf{Z} = \mathbf{z})$, the class conditioned probabilities for $i = 1, \ldots, M$ we use (10), (8), (3), (4), (6), and (5). Recall that we are using logarithmic arithmetic. By (18) we have for $v_1 > v_2$

$$\log(e^{v_1} + e^{v_2}) = v_1 + \psi(v_2 - v_1) \tag{21}$$

where $\psi(u) = \log(1+e^u)$. Assuming that $\psi(u)$ is realized by a table, (21) is implemented by two additions and one table lookup (TLU). We also assume that (6) and (5) are calculated using a table for the function $\phi(u) = \log(1/2 + 1/2\,\mathrm{erf}(u/\sqrt{2}\,))$. The total number of operations to implement this stage is dominated by $7KM$ additions, $2KM$ multiplications and $2KM$ TLUs.

Filtering: To compute $\hat{X}_{i,k}$ we use (12) and (13). To calculate $\rho_{i,k}$ we use a table form of the function $\xi(u) = 1/(1+e^u)$. The number of operations is dominated by $5KM$ additions, $3KM$ multiplications and KM TLUs. Finally, we use (9) to construct $\hat{\mathbf{X}}$ in KM additions and KM multiplications.

The total number of operations required by the MIXMAX algorithm is summarized in Table I (recall that the computational complexity in Table I is per output sample, while previously we listed the complexity per frame, i.e., per $L/2 \approx K$ output samples). We note that the computational burden imposed by the HMM MMSE is also a sum of two terms, where the first is proportional to $\log_2 K$ and the second is proportional to the number of mixtures, M.

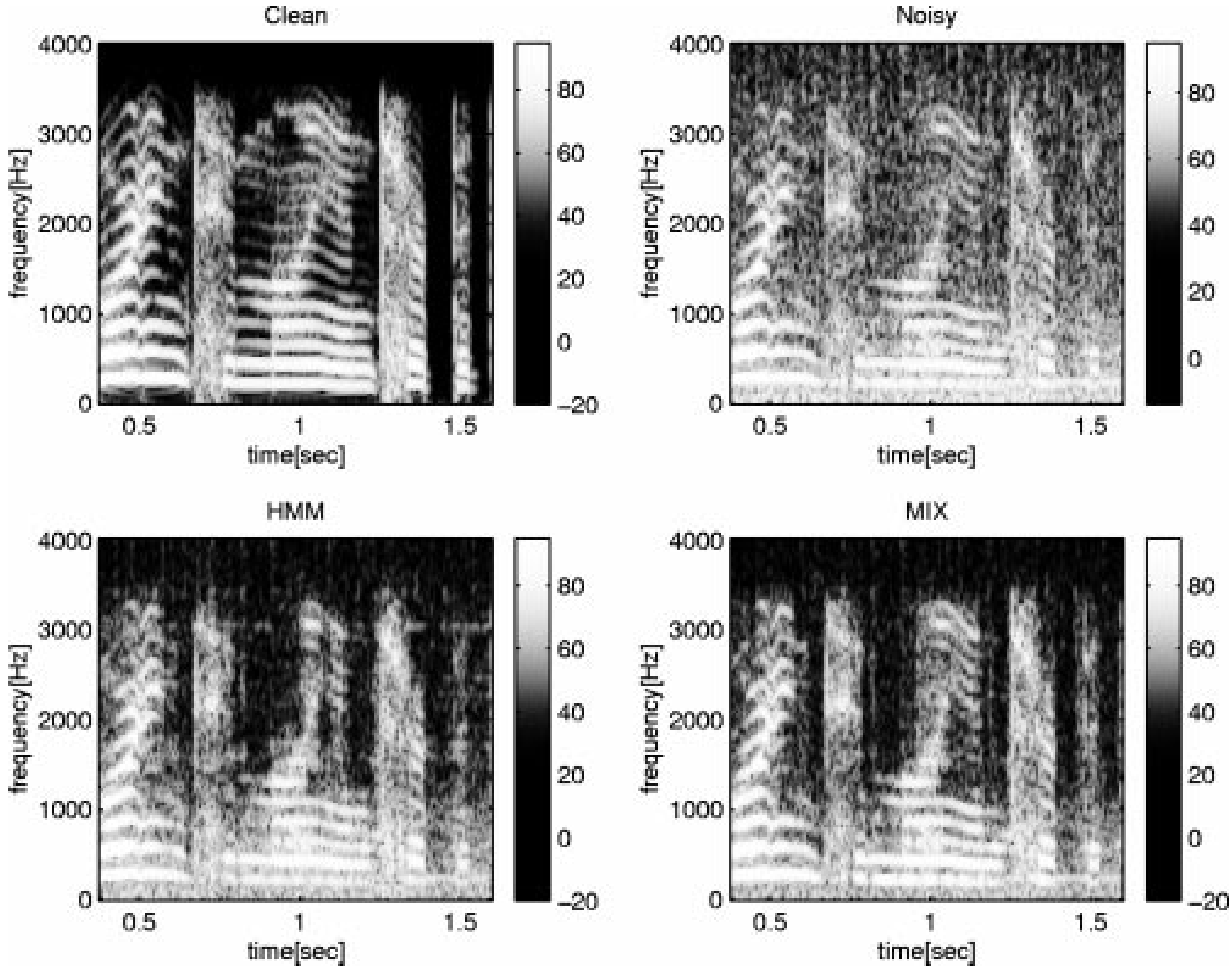

Fig. 5. Sonograms of the clean, noisy, HMM MMSE enhanced and MIXMAX enhanced speech in operation room environment at SNR level of 9 dB.

TABLE I
TOTAL NUMBER OF OPERATIONS PER OUTPUT SAMPLE FOR THE MIXMAX ALGORITHM

Multiplications	$4\log_2 K + 6M$
Additions	$6\log_2 K + 13M$
Table lookups	$3M$

The memory requirement is dominated by the $2KM$ cells required to store $\mu_{i,k}$ and $\sigma_{i,k}$.

Our algorithm can be easily implemented using a low cost DSP chip (e.g., for $K = 128$, $M = 20$ and a sampling rate of 8 kHz, Table I shows that the total number of operations per second is less than 4 million). However, in some applications, such as cellular communications, the DSP chip is responsible for a variety of tasks including speech coding and the receive–transmit modem. In such applications the speech enhancement task should consume only a small fraction of the total computational resources. By reducing the number of operations per second we also reduce the power consumption of the DSP, which may be limited in some applications, such as cellular telephony. In some applications, the speech enhancement should be performed on several channels at the same time (e.g., in a communication center). In this case it is also important to reduce the number of operations as much as possible in order to reduce the size and cost of the required hardware. Thus, we are motivated to reduce the computational requirements of the algorithm and make it closer to the complexity of spectral subtraction algorithms. In the rest of this section, we show how this goal can be achieved.

A. Tied Variances

In this case, the same mixture model (2) is used, except that the variance of the kth spectral component is now independent of the mixture

$$\sigma_{i,k} = \sigma_k \qquad \forall i = 0, \ldots, M-1 \quad k = 0, \ldots, K-1.$$

That is, the variances, $\{\sigma_{i,k}\}_{i=0}^{M-1}$ are tied together. The EM iteration is now described by (15), (16), and by the following equation that replaces (17):

$$\hat{\sigma}_k^2 = \frac{1}{N} \sum_{i=0}^{M-1} \sum_{n=0}^{N-1} \alpha_{n,i} (x_k^n - \hat{\mu}_{i,k})^2 \qquad k = 0, \ldots, K-1.$$

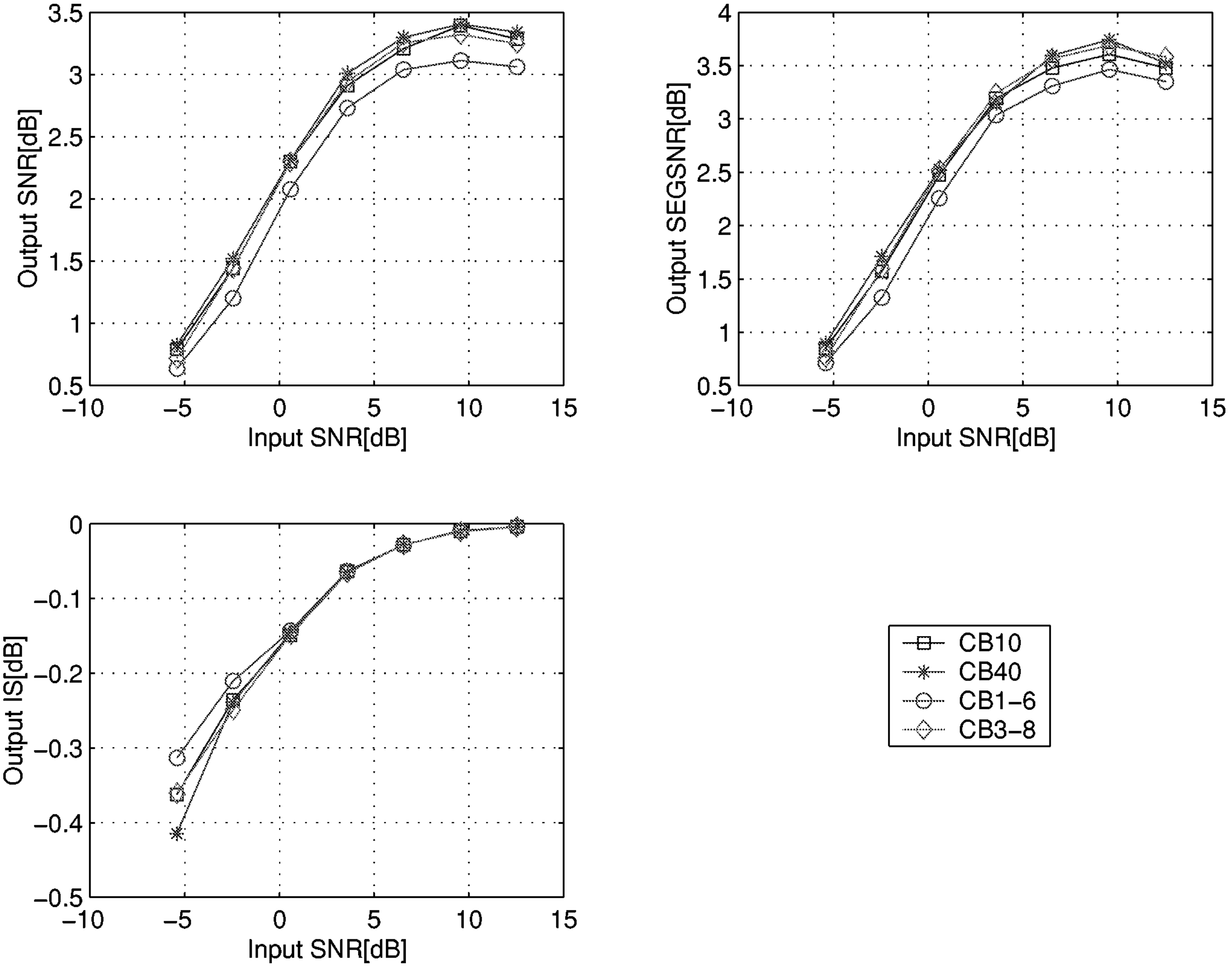

Fig. 6. Comparison between the performances of several codebook configurations in factory noise.

Tied variances enable a more compact representation, that is, when tying is applied, only K variance parameters are required (instead of KM), thus lowering memory requirements.

B. Dual Codebook Scheme

Given the speech signal samples of the current frame $x[l]$ $l = 0, \ldots, L-1$ (possibly weighted by some window function), we define

$$\Gamma = \log \sqrt{\sum_{l=0}^{L-1} x^2[l]}$$

$$\tilde{X}_k = \log |X(e^{j2\pi k/L})| - \Gamma = X_k - \Gamma$$

$$k = 0, \ldots, K-1$$

$$\tilde{\mathbf{X}} = [\tilde{X}_0, \tilde{X}_1, \ldots, \tilde{X}_{K-1}]^T$$

where $X(e^{j2\pi k/L})$ is defined by (1). Hence

$$X_k = \tilde{X}_k + \Gamma.$$

Γ and $\tilde{\mathbf{X}}$ are the (logarithmic) gain and gain normalized spectrum of the frame, respectively. We assume separate mixture models to $\tilde{X}_k$ and Γ. Let i denote the mixture index that corresponds to $\tilde{\mathbf{X}}$, and let j denote the mixture index that corresponds to Γ. The class conditioned density of X_k is

$$f_{i,j,k}(x_k) = \mathcal{N}(x_k, \mu_{i,k} + \mu_j^g, \sigma_k).$$

$\mu_{i,k}$ is the mean value that corresponds to the kth component of the ith mixture of $\tilde{\mathbf{X}}$. Similarly, μ_j^g is the mean value that corresponds to the jth mixture of Γ. Note that we assume a tied variances model. Denote by M_1, the total number of mixtures that correspond to $\tilde{\mathbf{X}}$. Similarly, denote by M_2, the total number of mixtures that correspond to Γ. The density of $\mathbf{X}$ is

$$f(\mathbf{x}) = \sum_{i=1}^{M_1} \sum_{j=1}^{M_2} c_i c_j^g f_{i,j}(\mathbf{x}) = \sum_i \sum_j c_i c_j^g \prod_k f_{i,j,k}(x_k)$$

where c_i, c_j^g are the mixture components that correspond to $\tilde{\mathbf{X}}$ and Γ respectively.

We estimate $\boldsymbol{\mu}_i = (\mu_{i,0}, \mu_{i,1}, \ldots, \mu_{i,K-1})^T$ $i = 0, \ldots, M_1 - 1$ by clustering the gain normalized spectrum $\tilde{\mathbf{X}}$, using a K-means algorithm. We then estimate μ_j^g $j = 0, \ldots, M_2 - 1$ by clustering the gains, Γ. c_i is obtained as a by-product of the K-means algorithm, by calculating the relative frequency of gain normalized spectrum vectors, classified as belonging to the ith mixture. c_j^g is obtained similarly, by calculating the relative frequency of gains classified as belonging to the jth mixture. Finally, the variances, σ_k are obtained using

$$\hat{\sigma}_k^2 = \frac{1}{N}\sum_{n=0}^{N-1}\left(x_k^n - \mu_{i_n,k} - \mu_{j_n}^g\right)^2$$

where i_n is the index of the mixture mean which is closest to $\tilde{x}^n$, i.e.,

$$i_n = \arg\min_i |\boldsymbol{\mu}_i - \tilde{\mathbf{x}}^n|^2.$$

Similarly

$$j_n = \arg\min_j |\mu_j^g - \Gamma^n|^2.$$

i_n and j_n are obtained as a byproduct of the K-means procedure.

In Fig. 6, we compare the performance of a standard (nontied) mixture (one with ten mixtures and one with 40 mixtures) with that of two dual codebook configuration. The first dual codebook configuration has $M_1 = 1$ and $M_2 = 6$. The second configuration has $M_1 = 3$ and $M_2 = 8$. In Fig. 6, we present the results for factory noise. Similar trend was observed for other noise sources from the NOISEX-92 database [21], including car noise and speech-like noise. As can be seen, even a very compact dual codebook configuration with $M_1 = 1$ and $M_2 = 6$ yields only a small degradation in the objective criteria examined. Subjective listening tests support these findings by showing no difference in the quality of the reconstructed speech produced by each one of these codebook configurations. Thus, a dual codebook scheme with M_1M_2 relatively small can be as effective as a standard (nontied) mixture with a larger value of M (i.e., $M_1M_2 < M$). In this way both the computational and memory requirements of the algorithm may be reduced.

C. Replacing Weighted Mixtures by the Most Probable Mixture Element

In this case we construct the enhanced speech based only on the most probable mixture. That is, (9) is now replaced by

$$\hat{\mathbf{X}} = \hat{\mathbf{X}}_l$$

where

$$l = \arg\max_i q(i \,|\, \mathbf{Z} = \mathbf{z}) = \arg\max_i c_i h_i(\mathbf{z}).$$

This simplification saves a fraction of $(M-1)/M$ of the filtering stage in the enhancement algorithm (approximately $6M$ additions, $4M$ multiplications and M TLUs per output sample), essentially with no noticeable reduction in the performance.

VI. CONCLUSIONS

We presented a new speech enhancement algorithm which was shown to be effective for improving the quality of the reconstructed speech. The derivation is based on the MIXMAX model which was originally proposed for designing noise adaptive speech recognition algorithms. Several modifications and simplifications were found useful. In particular, by using a dual codebook scheme that also incorporates tied variances, it is possible to significantly reduce the amount of model parameters (thus minimizing the memory and computational requirements of the algorithm), essentially without paying performance penalties.

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," in *Speech Enhancement*, J. S. Lim and A. V. Oppenheim, Eds. Englewood Cliffs, NJ: Prentice-Hall, 1983, pp. 61–68.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 1109–1121, 1984.

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

[4] Y. Ephraim, D. Malah, and B. H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1846–1856, Dec. 1989.

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

[6] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 373–385, July 1998.

[7] J. D. Gibson, B. Koo, and S. D. Gray, "Filtering of colored noise for speech enhancement and coding," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 39, pp. 1732–1742, Aug. 1991.

[8] B. G. Lee, K. Y. Lee, and S. Ann, "An EM-based approach for parameter enhancement with an application to speech signals," *Signal Process.*, vol. 46, pp. 1–14, 1995.

[9] K. Y. Lee and K. Shirai, "Efficient recursive estimation for speech enhancement in colored noise," *IEEE Signal Processing Lett.*, vol. 3, pp. 196–199, July 1996.

[10] J. B. Kim, K. Y. Lee, and C. W. Lee, "On the applications of the interacting multiple model algorithm for enhancing noisy speech," *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 349–352, May 2000.

[11] J. S. Lim, *Speech Enhancement*. Englewood Cliffs, NJ: Prentice-Hall, 1983.

[12] K. K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1987, pp. 177–180.

[13] H. Sameti, H. Sheikhzadeh, L. Deng, and R. Brennan, "Comparative performance of spectral subtraction and HMM-based speech enhancement strategies with application to hearing aid design," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, vol. 1, Adelaide, Australia, Apr. 1994, pp. 13–16.

[14] R. J. Vilmur, J. J. Barlo, I. A. Gerson, and B. L. Lindsley, "Noise suppression system," U.S. patent 4 811 404, 1989.

[15] A. Nádas, D. Nahamoo, and M. A. Picheny, "Speech recognition using noise-adaptive prototype," *IEEE Trans. Speech Audio Processing*, vol. 37, pp. 1495–1505, Oct. 1989.

[16] A. Erell and D. Burshtein, "Noise adaptation of HMM speech recognition systems using tied-mixtures in spectral domain," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 72–74, Jan. 1997.

[17] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. Ser. 3g, pp. 1–38, 1977.

[18] B. H. Juang and L. R. Rabiner, "Mixture autoregressive hidden Markov models for speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-33, pp. 1404–1413, 1985.

[19] A. Erell and M. Weintraub, "Filterbank-energy estimation using mixture and Markov models for recognition of noisy speech," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 68–76, Jan. 1993.

[20] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 72–73, Jan. 1995.

[21] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, 1993.

[22] A. H. Gray, R. M. Gray, A. Buzo, and Y. Matsuyama, "Distortion measures for speech processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 28, pp. 367–376, 1980.

[23] S. Gannot and D. Burshtein. (2001, Aug.) Audio sample files. [Online]. Available: http://www-sipl.technion.ac.il/~gannot/examples1.html.

David Burshtein (M'92–SM'99) received the B.Sc. and Ph.D. degrees in electrical engineering from Tel-Aviv University, Tel-Aviv, Israel, in 1982 and 1987, respectively.

During 1988–1989, he was a Research Staff Member in the Speech Recognition Group of IBM, T. J. Watson Research Center, Yorktown Heights, NY. In 1989, he joined the Department of Electrical Engineering—Systems, Tel-Aviv University. His research interests include information theory, speech, and signal processing.

Sharon Gannot (S'95–M'01) received the B.Sc. degree (summa cum laude) from the Technion—Israel Institute of Technology, Haifa, in 1986 and the M.Sc. (cum laude) and Ph.D. degrees from Tel-Aviv University, Tel-Aviv, Israel, in 1995 and 2000, respectively, all in electrical engineering.

From 1986 to 1993, he was Head of a Research and Development Section, in the R&D Center of the Israeli Defense Forces. In 2001, he held a postdoctoral position with the Department of Electrical Engineering (SISTA), K.U.Leuven, Belgium. Currently he holds a research fellowship position with the Technion-Israeli Institute of Technology. His research interests include parameter estimation, statistical signal processing, and speech processing using either single- or multimicrophone arrays.