SPEECH ENHANCEMENT USING A MIXTURE-MAXIMUM MODEL

David Burshtein and Sharon Gannot
Dept. Electrical Engineering – Systems
Tel-Aviv University, Tel-Aviv 69978, Israel
burstyn,sharong@eng.tau.ac.il

10.21437/Eurospeech.1999-570

ABSTRACT

We present a new 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. The new algorithm is shown to be 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.

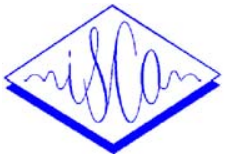
1 INTRODUCTION

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. 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, in which speech is enhanced by utilizing Wiener or Kalman filtering (e.g. [5]), based on the estimated parameters. 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], [7]. Other examples of spectral algorithms include the log spectral amplitude estimator (LSAE), proposed by Ephraim and Malah [2], and the hidden Markov model (HMM)-based filtering algorithm proposed by Ephraim [3]. In general, the computational requirements of the spectral domain algorithms is lower than those of the time domain algorithms.

The new proposed algorithm belongs to the class of spectral algorithms. It is similar to the HMM-based, minimum mean square error (MMSE) filtering algorithm proposed by Ephraim [3], in the sense that it also uses a mixture of Gaussians HMM to model the speech signal. We note, that in our experience the mixtured HMM model may be replaced by a mixture of Gaussians model which assumes independence from one frame to the other, without a noticeable performance degradation. Although the removal of the HMM assumption results in a significant simplification of the enhancement algorithm, it is still necessary to design and activate a series of Wiener filters (one for each mixture), whose outputs are properly combined to form the enhanced speech signal. Hence the simplified HMM MMSE algorithm might still be too complicated to implement in low-cost or low-power applications.

In the present paper we follow the simple MIXMAX model, which was originally suggested by Nadas *et al.* [6] to design a noise adaptive, speech recognition algorithm. The algorithm has been nicknamed the MIXMAX labeler, after the mixture and the maximum models. In [4] alternative noise robust speech recognition algorithms based on the MIXMAX model are proposed.

In this paper we present an effective speech enhancement algorithm based on the MIXMAX



6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

model, with some modifications and further simplifications. Our discussion is supported by an experimental study.

2 MIXMAX MODEL

Let $s[l] \quad l = 0, 1, \ldots, L-1$ denote the speech signal samples of the current frame (possibly weighted by some window function), and let $S(e^{j2\pi k/L})$ denote the Discrete Fourier Transform (DFT) of $s[l]$. Let $\mathbf{X}$ denote the $L/2+1$ dimensional, log-spectral vector of the speech signal, where its k-th component, $X_k = X(e^{j2\pi k/L})$, is defined by,

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

where $K = L/2 + 1$ (other values may be obtained using symmetry). We assume an additive colored noise model, which is statistically independent of the speech signal. Similarly, let $\mathbf{Y}$ and $\mathbf{Z}$ denote the log-spectral vectors of the noise and noisy speech signals, respectively. We assume the availability of a voice activity detector (VAD), thus noise statistics may be assumed 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. The reconstructed speech signal, $\hat{s}[l]$, is given by the inverse DFT of

$$\hat{S}(e^{j2\pi k/L}) = \exp\left\{\hat{X}_k\right\} \angle Z(e^{j2\pi k/L})$$

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]. The rest of the derivation is concerned with models and methods to obtain $\hat{\mathbf{X}}$.

Let $f(\mathbf{x})$ denote the probability density function of $\mathbf{X}$. We assume that $f(\mathbf{x})$ can be modeled by a mixture of diagonal covariance Gaussians, i.e.:

$$f(\mathbf{x}) = \sum_i c_i f_i(\mathbf{x}) = \sum_i c_i \prod_k f_{i,k}(x_k) \qquad (1)$$

where, $f_{i,k}(x) = \mathcal{N}(x, \mu_{i,k}, \sigma_{i,k})$. Let $g(\mathbf{y})$ denote the probability density function of the spectral noise, $\mathbf{Y}$. We assume that $g(\mathbf{y})$ can be modeled by a single diagonal covariance Gaussian, i.e., $g(\mathbf{y}) = \prod_k g_k(y_k)$, where $g_k(y) = \mathcal{N}(y, \mu_{Y,k}, \sigma_{Y,k})$. Let Z denote the log-energy frame vector of the noisy speech signal. Note that under the statistical independence, additive noise model assumptions

$$Z_k = \log\left(\exp(X_k) + \exp(Y_k)\right)$$

The assumption in the MIXMAX model, suggested by Nadas *et al.* [6], is that we can approximate Z_k by

$$Z_k \approx \max(X_k, Y_k)$$

The density of Z_k given the i-th mixture, $h_{i,k}(z)$, is obtained by differentiating the cumulative distribution function (CDF) [6],

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

(Lower case letters denote densities. Capital letters denote CDFs). The density of Z is hence given by,

$$h(\mathbf{z}) = \sum_i c_i h_i(\mathbf{z}) = \sum_i c_i \prod_k h_{i,k}(z_k) \qquad (2)$$

Eq. (2) was used by Nadas *et al.* and by Erell and Burshtein [4] to design noise adaptive speech recognition systems. In the present paper we apply the MIXMAX model to the related problem of speech enhancement.

The estimated speech vector $\hat{\mathbf{X}}$ can be calculated from the expected value of X_k given the class i and the noisy observation z_k,

$$\hat{X}_{i,k} = \mathrm{E}\left\{X_k \mid Z_k = z_k, I = i\right\} \qquad (3)$$

Now, performing the required calculations, we obtain

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

where,
$R_{i,k} = f_{i,k}(z_k)/F_{i,k}(z_k)$; $R_{Y,k} = g_k(z_k)/G_k(z_k)$, and $\rho_{i,k} = \frac{1}{1+R_{Y,k}/R_{i,k}}$.

3 IMPLEMENTATION

To apply the method of Section 2, a mixture model of the type of equation (1) needs to be trained, using the maximum likelihood approach outlined in [6]. The objective is to set $c_i, \mu_{i,k}, \sigma_{i,k}$ so as to maximize the log-likelihood of the observed data. The maximization may be carried out by using the expectation-maximization (EM) algorithm, as in [6].

The following improvements and simplifications were found useful.

3.1 Tied Variances

We use the same mixture model (1), except that the variance of the k-th spectral component is now independent of the mixture. That is, the variances, $\{\sigma_{i,k}\}_{i=0}^{M-1} = \sigma_k \ \forall \ k = 0, \ldots, K-1$, are tied together, which enables a more compact representation. Our experiments indicate that this saving is usually possible without a significant loss in the performance.

3.2 Dual Codebook Scheme

Given the speech signal samples of the current frame $s[l] \quad l = 0, \ldots, L-1$ (possibly weighted by some window function), we define, , $=$ $\log \sqrt{\sum_{l=0}^{L-1} s^2[l]}$, hence, $X_k = \tilde{X}_k +$, . , and $\tilde{\mathbf{X}}$ are the (logarithmic) gain and gain normalized spectrum of the frame. We assume separate mixture models to $\tilde{X}_k$ and , . Let i denote the mixture index that corresponds to $\tilde{\mathbf{X}}$, and let j denote the mixture index that corresponds to , . 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 k-th component of the i-th mixture of $\tilde{\mathbf{X}}$. Similarly, μ_j^g is the mean value that corresponds to the j-th mixture of , . Note that we assume a tied variances model. The parameters are estimated by the K-means procedure.

3.3 Most probable mixture

A simplification that was found useful, is to construct the enhanced speech based only on the most probable mixture, $l = \arg\max_i c_i h_i(\mathbf{z})$, instead of weighting all the possible mixtures.

3.4 Non-linear Post-processing

Non-linear post-processing was applied in the past in spectral subtraction methods [1], [7]. We found non-linear postprocessing to be very effective in improving the quality of the enhanced speech. Let $\gamma_k = \exp\left\{\hat{X}_k - Z_k\right\}$. $\gamma_k < 1$ is the spectral suppression of the k-th channel. A simple and useful post-processing is obtained by constraining γ_k to be above some frequency dependent threshold, δ_k. That is, the estimated speech vector is replaced by, $\tilde{X}_k = \max(Z_k + \log \delta_k, \hat{X}_k)$.

4 EXPERIMENTS

To test the performance of the new MIXMAX algorithm we used 8 sentences from the TIMIT and Resource Management databases (3 females, 5 males). Clean speech model training was performed using 10 other TIMIT sentences. The sentences were corrupted by additive noise, using two types of noise signals. The first was a synthetic white Gaussian noise source. The second was a computer fan signal, that was recorded in our laboratory. Various signal to noise ratios (SNRs) were used in the experiments. In all our experiments noise parameters were estimated from noise only segments. Both objective and subjective listening tests were employed. The performance was compared to the performance of the HMM-based minimum mean square error (MMSE) speech enhancement algorithm [3]. Both algorithms used the postprocessing modification that was outlined in subsection 3.4. The performance of the spectral subtraction algorithm in [1] was inferior both to the HMM MMSE algorithm and to the new proposed MIXMAX algorithm. Figure 1 illustrates the performances of the HMM MMSE and MIXMAX algorithms, for the case where 20 Gaussian mixtures are used, both for white Gaussian noise (left) and for computer fan noise (right). Figure 1 presents the enhancement of the total output SNR (TOTSNR), and the median of the segmental SNR (MEDSEG) taken over all speech frames, in dB scale. Both distance measures show clear advantage to the simpler MIXMAX algorithm. The same trend was observed when 5 Gaussian mixtures were used. These results were verified by informal listening tests, using several listeners. The quality of the enhanced MIXMAX speech signal was clearly improved compared to the HMM MMSE enhanced speech, over the entire SNR range examined.

It should be noted that postprocessing reduces objective quality measures such as SNR enhancement. However, at the same time it results in a significant improvement in the quality of the enhanced speech, especially for low SNR inputs.

Figure 2 compares the performance of the MIXMAX algorithm when 20 mixtures are used, as in Figure 1 (denoted by 1CB) to the per-

formance of a reduced parameters, simplified variant of the algorithm (denoted by 2CB), in which we used the two codebook scheme (with $M_2 = 6$ mixtures for the gain and only $M_1 = 1$ mixture for the gain-normalized spectrum), the most probable mixture variant of the algorithm, and a uniform mixture components ($c_i = 1/M_1$ and $c_j^g = 1/M_2$). In spite of these simplifications, and the significant reduction in the number of parameters used, there is almost no loss in terms of objective SNR. There is virtually no loss in the performance in terms of subjective speech quality.

5 CONCLUSIONS

We presented a new, computationally efficient 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. 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. Postprocessing was essential to produce high quality speech.

6 REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 27, no. 2, pp. 113-120, 1979.

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

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

[4] A. Erell and D. Burshtein, "Noise adaptation of HMM speech recognition systems using tied-mixtures in the spectral domain", *IEEE Transactions on Speech and Audio Processing*, volume 5, pp. 72-74, January 1997.

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

[6] A. Nadas, D. Nahamoo and M. Picheny, "Speech recognition using noise adaptive prototypes", *IEEE Trans. on ASSP*, vol. 37, pp. 1495-1503, 1989.

[7] R. J. Vilmur, J. J. Barlo, I. A. Gerson and B. L. Lindsley, "Noise Suppression System", U.S. patent no. 4,811,404, 1989.

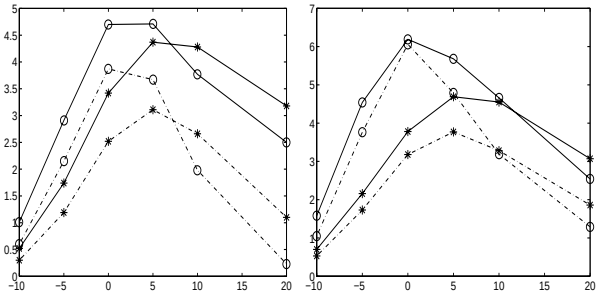


Figure 1: HMM (dash-dot line) vs. MIXMAX (full line). $*$ denotes TOTSNR, $\circ$ denotes MEDSEG. Left: white noise. Right: Computer fan.

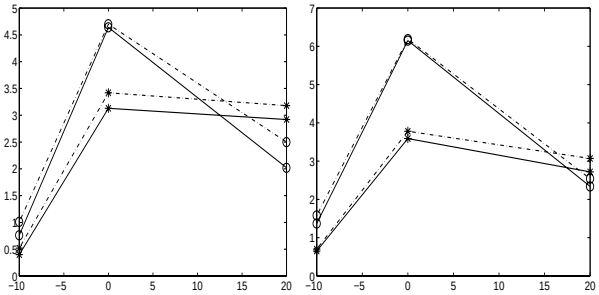


Figure 2: One codebook (1CB, dash-dot line) vs. two codebooks (2CB, full line). $*$ denotes TOTSNR, $\circ$ denotes MEDSEG.