Speech Enhancement Using a Multidimensional Mixture-Maximum Model

Yochay Yeminy, Sharon Gannot and Yosi Keller
School of Electrical and Computer Engineering
Bar-Ilan University, Ramat-Gan, Israel
yochay.ye@gmail.com; gannot@eng.biu.ac.il; yosi.keller@gmail.com

***Abstract*—We present a single-microphone speech enhancement algorithm that models the log-spectrum of the noise-free speech signal by a multidimensional Gaussian mixture. The proposed estimator is based on an earlier study which uses the single-dimensional mixture-maximum (MIXMAX) model for the speech signal. The experimental study shows that there is only a marginal difference between the proposed extension and the original algorithm in terms of both objective and subjective performance measures.**

I. Introduction

Speech quality is a major design aspect in a variety of applications, e.g. telephony, teleconferencing and automatic speech recognition (ASR) systems. Therefore, a significant effort has been made, and many algorithms have been developed aiming to reduce the noise level of corrupted speech signals. In this paper, we are only concerned with single channel applications in which a speech signal, contaminated by an additive stationary noise, is captured by a close-talk microphone. The latter expemts us from the problem of a possibly reverberated incoming speech signal. A variety of algorithms that addresses this problem is available. Some of them aim at enhancing the speech signal in the time-domain, and others in the spectral-domain. As the computational complexity of the time-domain algorithms is often higher than that of its frequency-domain counterparts, the latter is commonly regarded advantageous. Notable algorithms in the frequency domain are the log spectral amplitude (LSA) estimator proposed by Ephraim and Malah in [1] and its improved version, denoted optimally modified-LSA (OM-LSA), proposed by Cohen and Berdugo [2]. Other algorithms, that exploit the hidden Markov model (HMM) presumed structure of the speech production, were introduced by Ephraim et al. [3], [4].

The MIXMAX model was proposed by Nádas et al. [5] in the context of speech recognition problems, and was later applied by Burshtein and Gannot [6] to speech enhancement. The model is based on a Gaussian mixture modeling of the speech frequency bins, as well as on an approximation that replaces the observed log-amplitude of each frequency bin by the maximum of the log-amplitude of the corresponding speech and noise signals. The main attribute of this enhancement algorithm is the very low computational complexity, obtained without sacrificing the quality of the resulting speech signal. The experimental results in [6] show that the MIXMAX algorithm compares favorably to other speech enhancement algorithms, e.g. HMM-based algorithms [3] and the time-domain Kalman filter based algorithm [7]. The MIXMAX model was extended by several researchers in other contexts, too. Radfar and Dansereau [8] derived a single-microphone speech segregation technique by modeling the two speech sources as Gaussian mixtures. Other extensions include [9] and [10] (in the context of singer identification from music recordings).

All of the above mentioned algorithms are based on the underlying assumption that the frequency bins of the speech signal are statistically independent, although it is well-known that the speech power spectral density (PSD) is smooth. Moreover, many speech enhancement algorithms impose PSD smoothness as a post-processing stage, especially while estimating various speech attributes such as a priori signal-to-noise ratio (SNR) and speech probability [2]. The goal of the current contribution is therefore to test this assumption and to propose a speech enhancement algorithm that takes the dependence of the frequency bins into account.

The paper is organized as follows. In Section II, the problem is formulated, the new algorithm is introduced and some implementation issues are discussed. Section III is dedicated to the experimental study comparing the single- and multidimensional models. Section IV concludes the paper.

II. Speech Enhancement Using a Multidimensional MIXMAX Model

A. Signal Model

Let $X(e^{j2\pi q/L})$ denote the short-time Fourier transform (STFT) of a speech signal, where $q = 0, 1, \ldots, L-1$ is the frequency bin, L is the frame length, and the overlap between subsequent frames is $L/2$. Now, define $\mathbf{X}$ as a vector comprised of the first $L/2+1$ elements of $\log\left|X(e^{j2\pi q/L})\right|$. Note that the other frequency bins can be obtained by the symmetry of the discrete Fourier transform (DFT). Define $\mathbf{X}_k$ as the kth frequency subband, comprising a group of adjacent elements of $\mathbf{X}$. It is assumed that the elements within each subband are statistically correlated, while the subbands are mutually uncorrelated. The Gaussian mixture model (GMM) in [6] is subsequently extended to a multidimensional model to reflect the inter-bin correlation within each subband.

The probability density function (p.d.f.) $f(\mathbf{x})$ of $\mathbf{X}$ can be therefore written as

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

where

$$f_{i,k}(\mathbf{x}_k) = \frac{1}{\sqrt{(2\pi)^{d_k}\,|C_{i,k}|}} \times \exp\left\{ \left(\mathbf{x}_k - \boldsymbol{\mu}_{i,k}\right)^T C_{i,k}^{-1} \left(\mathbf{x}_k - \boldsymbol{\mu}_{i,k}\right) \right\}. \tag{2}$$

$f_i(\mathbf{x})$ is an abbreviation of $f(\mathbf{x}|I=i)$ and I is the mixture index. The product $\prod_k f_{i,k}(\mathbf{x}_k)$ is a result of the subband independence. c_i is defined as the probability of the ith mixture, $C_{i,k}$ and $\boldsymbol{\mu}_{i,k}$ are the covariance matrix and the mean of $\mathbf{X}_k$ in the ith mixture, respectively. d_k is the dimension of $\mathbf{X}_k$.

In a similar way, let $\mathbf{Y}$ be the log-spectral vector of the noise signal. Contrary to the log-spectral vector of the speech signal, it is assumed that the components of $\mathbf{Y}$ are statistically independent. Moreover, it is assumed that $\mathbf{Y}$ obeys a first-order GMM (i.e. Gaussian) distribution. Denote, $g(\mathbf{y})$ the p.d.f. of the log-spectral vector of the noise

$$g(\mathbf{y}) = \prod_k g_k(\mathbf{y}_k) \tag{3}$$

where

$$g_k(\mathbf{y}_k) = \frac{1}{\sqrt{(2\pi)^{d_k}\,|C_{Y,k}|}} \times \exp\left\{ \left(\mathbf{y}_k - \boldsymbol{\mu}_{Y,k}\right)^T C_{Y,k}^{-1} \left(\mathbf{y}_k - \boldsymbol{\mu}_{Y,k}\right) \right\}. \tag{4}$$

Note, that due to the frequency bin independence, the matrix $C_{Y,k}$ is diagonal. $g(\mathbf{y})$ is notated as a multidimensional distribution to keep the consistency with the definition of $f(\mathbf{x})$.

It is assumed that the speech signal is captured by a close-talk microphone. Under the additive noise model the received signal $z[n]$ is given by

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

Nádas et al. [5] showed that the received signal can be approximated in the log-spectral domain as

$$\mathbf{Z} \approx \max(\mathbf{X}, \mathbf{Y}) \tag{5}$$

where the maximization is component-wise over the elements of $\mathbf{X}$ and $\mathbf{Y}$. Based on this approximation, the p.d.f. of $\mathbf{Z}$ is given by

$$\begin{aligned} h(\mathbf{z}) &= \sum_i c_i h_i(\mathbf{z}) = \sum_i c_i \prod_k h_{i,k}(\mathbf{z}_k) \\ &= \sum_i c_i \prod_k \left[f_{i,k}(\mathbf{z}_k) G_k(\mathbf{z}_k) + F_{i,k}(\mathbf{z}_k) g_k(\mathbf{z}_k) \right] \end{aligned} \tag{6}$$

where $F_{i,k}(\mathbf{x}_k)$ and $G_k(\mathbf{y}_k)$ are the cumulative distribution function (c.d.f.) of $\mathbf{X}_{i,k}$ and $\mathbf{Y}_k$, respectively.

B. Algorithm Derivation

The c.d.f. is calculated by integrating the p.d.f.:

$$\begin{aligned} F_{i,k}(\mathbf{x}_k) &= \int \mathbf{x}_k f_{i,k}(\mathbf{x}_k) d\mathbf{x}_k \\ &\propto \int \mathbf{x}_k \exp\left\{ -\frac{1}{2} \left(\mathbf{x}_k - \boldsymbol{\mu}_{i,k}\right)^T C_{i,k}^{-1} \left(\mathbf{x}_k - \boldsymbol{\mu}_{i,k}\right) \right\} d\mathbf{x}_k. \end{aligned} \tag{7}$$

The calculation of $F_{i,k}(\mathbf{x}_k)$ can be performed by using the eigenvalue decomposition (EVD)

$$C_{i,k}^{-1} = V_{i,k} \Lambda_{i,k} V_{i,k}^T$$

where $V_{i,k}$ is an orthonormal matrix of the eigenvectors of $C_{i,k}^{-1}$, and $\Lambda_{i,k}$ is a diagonal matrix with the eigenvalues of $C_{i,k}^{-1}$ on its diagonal. Note that $C_{i,k}^{-1}$ is assumed to be full-rank matrix and therefore has exactly d_k eigenvalues, denoted $\lambda_{i,k,l}$ $l = 1, 2, \ldots, d_k$.

Now, it is convenient to exchange $V_{i,k}^T \left(\mathbf{x}_k - \boldsymbol{\mu}_{i,k}\right)$ by a new integration variable $\mathbf{r}_{i,k}$. Thus, the multidimensional integral simplifies to a product of one-dimensional integrals. Following several mathematical manipulations we finally obtain an expression for the speech c.d.f.

$$F_{i,k}(\mathbf{x}_k) = \prod_{l=1}^{d_k} \left(1 - \frac{1}{2} \mathrm{erfc}\left(\sqrt{\frac{\lambda_{i,k,l}}{2}} t_{i,k,l} \right) \right)$$

where $t_{i,k,l}$ is the lth component of $\mathbf{t}_{i,k} = V_{i,k}^T \left(\mathbf{x}_k - \boldsymbol{\mu}_{i,k}\right)$, and

$$\mathrm{erfc}\,(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-u^2} du$$

is the complementary error function (erfc).

The calculation of $G_k(\mathbf{y}_k)$ is straightforward since $C_{Y,k}^{-1}$ is diagonal. Again, for consistency, the presentation of $G_k(\mathbf{y}_k)$ is given in matrix form

$$G_k(\mathbf{y}_k) = \prod_{l=1}^{d_k} \left(1 - \frac{1}{2} \mathrm{erfc}\left(\sqrt{\frac{\lambda_{Y,k,l}}{2}} \left(\mathbf{y}_k - \boldsymbol{\mu}_{Y,k}\right)_l \right) \right)$$

where $\left(\mathbf{y}_k - \boldsymbol{\mu}_{Y,k}\right)_l$ is the lth component of $\left(\mathbf{y}_k - \boldsymbol{\mu}_{Y,k}\right)$ and $\lambda_{Y,k,l}$ is the lth eigenvalue of $C_{Y,k}^{-1}$.

The minimum mean square error (MMSE) estimate of the clean speech signal is given by the conditional expectation

$$\hat{\mathbf{X}} = \mathrm{E}(\mathbf{X} \mid \mathbf{Z}) = \sum_i \hat{\mathbf{X}}_i \Pr(i \mid \mathbf{Z} = \mathbf{z}) \tag{8}$$

where $\hat{\mathbf{X}}_i$ is the estimation of $\mathbf{X}$ given $\mathbf{Z}$ and the ith mixture. Due to the subband independence this estimation can be marginalized into separate estimates for each component, denoted $\hat{\mathbf{X}}_{i,k}$:

$$\begin{aligned} \hat{\mathbf{X}}_{i,k} &= \mathrm{E}(\mathbf{X}_k \mid \mathbf{Z}_k = \mathbf{z}_k,\, I = i) \\ &= \int \mathbf{x}_k \frac{f_{i,k}(\mathbf{x}_k) h_{i,k}(\mathbf{z}_k \mid \mathbf{X}_k = \mathbf{x}_k)}{h_{i,k}(\mathbf{z}_k)} \, d\mathbf{x}_k. \end{aligned} \tag{9}$$

It is necessary to obtain an expression for $h_{i,k}(\mathbf{z}_k \mid \mathbf{X}_k = \mathbf{x}_k)$ for solving this integral. It is straightforward to express

$$\Pr\left\{ \mathbf{Z}_k < \mathbf{z}_k \mid \mathbf{X}_k = \mathbf{x}_k \right\} = \Pr\left\{ \mathbf{Y}_k < \mathbf{z}_k \right\} u(\mathbf{z}_k - \mathbf{x}_k)$$

where the multidimensional step-function is defined as

$$u(\mathbf{z}_k - \mathbf{x}_k) = \begin{cases} 1 & z_{k,l} > x_{k,l},\ \forall l = 1, 2, \ldots, d_k \\ 0 & \text{otherwise} \end{cases}.$$

Differentiating $\Pr\{\mathbf{Z}_k < \mathbf{z}_k \mid \mathbf{X}_k = \mathbf{x}_k\}$ with respect to $\mathbf{z}_k$ we finally obtain

$$\begin{aligned} &h_{i,k}(\mathbf{z}_k \mid \mathbf{X}_k = \mathbf{x}_k) \\ &\quad = g_k(\mathbf{y}_k)\delta(\mathbf{z}_k - \mathbf{x}_k) + G_k(\mathbf{y}_k)u(\mathbf{z}_k - \mathbf{x}_k) \end{aligned} \tag{10}$$

where $\delta(\mathbf{z}_k - \mathbf{x}_k)$ is the multidimensional Dirac's Delta function. Applying again the EVD, the integral in (9) can be calculated:

$$\hat{\mathbf{X}}_{i,k} = \mathbf{z}_k\rho_{i,k} + (\boldsymbol{\mu}_{i,k} - \mathbf{m}_{i,k}/F_{i,k}(\mathbf{z}_k))(1 - \rho_{i,k}) \tag{11}$$

where

$$R_{i,k} = f_{i,k}(\mathbf{z}_k)/F_{i,k}(\mathbf{z}_k); \qquad R_{Y,k} = g_k(\mathbf{z}_k)/G_k(\mathbf{z}_k)$$

$$\rho_{i,k} = \frac{1}{1 + R_{Y,k}/R_{i,k}}; \qquad \mathbf{m}_{i,k} = V_{i,k}\mathbf{b}_{i,k}.$$

The components $b_{i,k,l}$, $l = 1, \ldots, d_k$ of $\mathbf{b}_{i,k}$ are given by

$$b_{i,k,l} = \frac{1}{\sqrt{2\pi\lambda_{i,k,l}}} \exp\left\{-\frac{\lambda_{i,k,l}r_{i,k,l}^2}{2}\right\} \tag{12}$$

$$\times \prod_{j=1,\, j\neq l}^{d_k} \left(1 - \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{\lambda_{i,k,j}}{2}}r_{i,k,j}\right)\right) \tag{13}$$

where $r_{i,k,j}$ is the jth component of $\mathbf{r}_{i,k} = V_{i,k}^T\left(\mathbf{z}_k - \boldsymbol{\mu}_{i,k}\right)$. The enhanced speech signal is obtained by applying (11) and (8) and by transforming $\hat{\mathbf{X}}$ back to the time-domain using the noisy signal phase.

It can be verified that by setting $d_k = 1$, $\forall k$ the multidimensional estimator (11) simplifies to

$$\hat{X}_{i,k} = z_k\rho_{i,k} + (\mu_{i,k} - \sigma_{i,k}^2 R_{i,k})(1 - \rho_{i,k}) \tag{14}$$

and identifies with the estimator obtained in [6].

C. Implementation Aspects

Several implementation aspects should be addressed before applying the algorithm in (11).

First, the speech parameters c_i, $\boldsymbol{\mu}_{i,k}$, $C_{i,k}$ are estimated by applying the estimate-maximize (EM) procedure using clean speech signal as training data. The procedure is a straightforward extension of the procedure described in [6] to the multidimensional case. Noise features $\mu_{Y,k}$, $\sigma_{Y,k}^2$ were estimated in a noise-only segment of the noisy observation assuming the existence of a perfect voice activity detector (VAD) and noise stationarity.

It is well-known that integrating a threshold stage in speech enhancement algorithms is beneficial in maintaining low speech distortion. We use the same procedure applied in [6] to limit the amount of noise reduction using frequency-dependent threshold

$$\tilde{X}_k = \max(\hat{X}_k,\ Z_k + \log\delta_k) \tag{15}$$

where δ_k is maximum amount of signal suppression.

Finally, as observed in [6] it is possible to use in (8) the most probable mixture rather than the weighted sum of all mixtures:

$$\hat{i} = \arg\max\{\Pr(i \mid \mathbf{Z} = \mathbf{z})\}. \tag{16}$$

This choice reduces the computational complexity without sacrificing performance. In Section III we will validate this approximation, and compare the performance of the simplified algorithm with the performance of the algorithm using all mixtures.

III. Experimental Study

In this section we present an experimental study of the proposed algorithm and compare it with the original version of the MIXMAX algorithm [6].

A. Setup

The speech parameters were extracted using a training set comprising of 50 sentences, spoken by 25 males and 25 females. 40 Gaussians were used to model the GMM of the clean speech, and while applying the STFT, frames of 256 samples with 50% overlap were used. Hann window was used for both the analysis and synthesis. The dimension of the log-spectral vector of the clean signal, $\mathbf{X}$, is therefore 129.

The objective figures-of-merit used for the comparison between the speech enhancement algorithms were the weighted signal-to-noise ratio (W-SNR), with frequency weighting according to ANSI specifications [11], and the log spectral distance (LSD).

The subjective tests included unofficial listening test and assessment of sonograms (not presented here due to lack of space).

B. Correlation between Frequency Bins

Prior to the examination of the performance of the proposed estimators, we tried to justify the assumption that adjacent frequency bins are correlated. There are many ways to partition the frequency bins into subbands. In our test we used equal-dimension subbands, i.e. $d_k = d$, $\forall k$, with d ranging between 3 and 7. It was clearly demonstrated that the correlation matrices of the speech signal (obtained in the training stage) are not dominated by their elements on the diagonal. In the remainder of the experimental study we use 3 bins per subband (hence 43 subbands). It should be emphasized however that non-equal and even overlapping subbands might be reasonable choices as well.

C. Comparison of the One-dimensional and Multidimensional Estimators

Two variants of the proposed algorithm, namely using all mixtures or using only the most probable mixture, were tested and compared with the corresponding one-dimensional algorithm [6]. The test set consists of 50 sentences (25 males, 25 females), different from the training set. The noise signals were drawn from the NOISEX-92 database [12]. We used car, factory, room and speech-like noise signals as well as computer-generated white Gaussian noise. Results for the

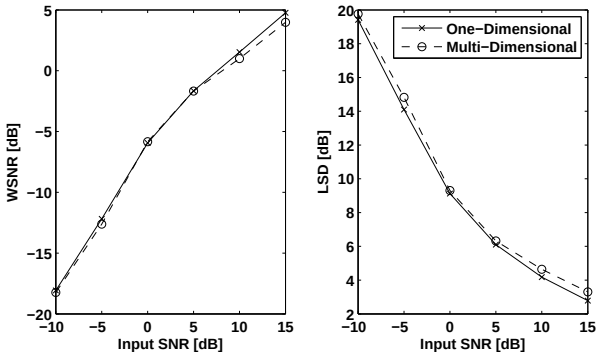

Figure 1. LSD and W-SNR comparison between the one-dimensional and the multidimensional estimators using all mixtures. Factory noise case.

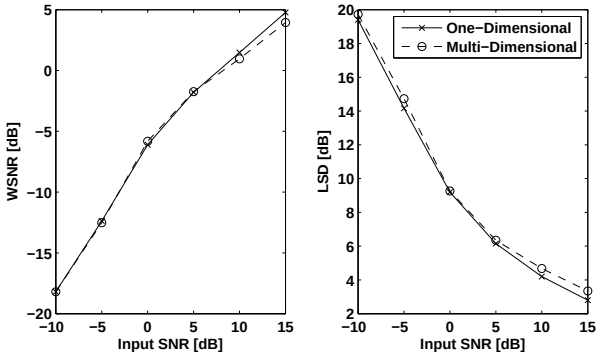

Figure 2. LSD and W-SNR comparison between the one-dimensional and the multidimensional estimators using the most probable mixture. Factory noise case.

algorithm variant using all mixtures and factory noise are depicted in Fig. 1. Only insignificant difference between the one-dimensional and the multidimensional estimators were encountered (in favor of the former). This result is also verified by our informal listening test for all noise types. A comparison of the original one-dimensional MIXMAX algorithm be found in [6].

Applying the one-dimensional and the multidimensional estimators to the same data, while using only the most probable mixture, exhibits only minor degradation, keeping both estimators comparable. The results for this case are shown in Fig. 2.

D. Discussion

It has been shown in the previous sections that the multidimensional estimator does not exhibit any advantage over its one-dimensional counterpart. Here we try to analyze this somewhat disappointing result. For this purpose we applied the algorithm to an artificial scenario. In this scenario we simulated log spectral segments of a desired signal obeying the multidimensional Gaussian mixture regime and log spectral segments of a noise signal using Gaussian densities. The noisy observations were simulated according to (5). This simulation showed that the estimator is trying to fit the observed noisy segment $\mathbf{Z}$ to the closest Gaussian mixture of the desired signal model. We therefore concluded that using a one-dimensional model increases the flexibility of the clustering procedure, and hence its robustness. This increased robustness may compensate for the inferior smoothness of the one-dimensional estimator.

IV. Summary

In this paper, we proposed an extension of the MIXMAX speech enhancement algorithm to a multidimensional GMM. In the extended model, the assumed correlation between frequency bins is manifested by multidimensional GMMs. The performance of the proposed estimator was compared to the original estimator that does not assume any correlation between log spectral bins. Only insignificant differences between estimators were encountered. We also showed that for both estimators, using only the most probable mixture can reduce the computational burden, without any notable performance degradation. We hypothesize that the multidimensional model fails to improve the enhancement performance over the single-dimensional model due to its decreased frequency-domain resolution.

References

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443 – 445, apr 1985.

[2] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403 – 2418, 2001.

[3] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden markov models for enhancing noisy speech," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 12, pp. 1846 –1856, dec 1989.

[4] Y. Ephraim, "A bayesian estimation approach for speech enhancement using hidden markov models," *Signal Processing, IEEE Transactions on*, vol. 40, no. 4, pp. 725 –735, apr 1992.

[5] A. Nádas, D. Nahamoo, and M. Picheny, "Speech recognition using noise-adaptive prototypes," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 37, no. 10, pp. 1495 –1503, oct 1989.

[6] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *Speech and Audio Processing, IEEE Transactions on*, vol. 10, no. 6, pp. 341 – 351, sep 2002.

[7] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential kalman filter-based speech enhancement algorithms," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 4, pp. 373 –385, jul 1998.

[8] M. Radfar and R. Dansereau, "Single-channel speech separation using soft mask filtering," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2299 –2310, nov. 2007.

[9] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "Monaural speech segregation based on fusion of source-driven with model-driven techniques," *Speech Communication*, vol. 49, no. 6, pp. 464 – 476, 2007.

[10] W. Tsai, H. Wang, and D. Rodgers, "Automatic singer identification of popular music recordings via estimation and modeling of solo vocal signal," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[11] S. Davis, "Octave and fractional octave band digital filtering based on the proposed ansi standard," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '86.*, vol. 11, apr 1986, pp. 945 – 948.

[12] A. Varga and H. J. 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 Communication*, vol. 12, no. 3, pp. 247 – 251, 1993.