Iranian Journal of Science & Technology, Transaction B, Engineering, Vol. 31, No. B3, pp 361-375
Printed in The Islamic Republic of Iran, 2007

A GENERALIZED APPROACH FOR MODEL-BASED SPEAKER-DEPENDENT SINGLE CHANNEL SPEECH SEPARATION*

M. H. RADFAR[1], ** A. SAYADIYAN[1], AND R. M. DANSEREAU[2]

[1]Dept. of Electrical Engineering, Amirkabir University of Technology, Tehran, I. R. of Iran,15875-4413
Email: radfar@sce.carleton.ca
[2]Dept. of Systems and Computer Engineering, Carleton University, Ottawa, Ontario, Canada, K1S 5B6

Abstract– In this paper, we present a new technique for separating two speech signals received from one microphone or one communication channel. In this special case, the separation problem is too ill-conditioned to be handled with common blind source separation techniques. The proposed technique is a generalized approach to model-based speaker-dependent single channel speech separation techniques in which a priori knowledge of the underlying speakers is used to separate speech signals. The proposed technique not only preserves the advantages of model-based speaker dependent single channel speech separation algorithms (i.e. high separability), but also is able to separate the speech signals of an unlimited number of speakers given the speakers' models (i.e. generality). The whole algorithm consists of three stages: classification, identification, and separation. The identities of speakers speech signals form the mixed signal are first determined at the classification and identification stages. Identified speakers' model is then used to separate the underlying signals using a novel approach consisting of Gaussian mixture modeling, maximum likelihood estimation and Wiener filtering. Evaluation results conducted on a database consisting of 100 mixed speech signals with target-to-interference ratios (TIR) ranging from -9 dB to +9 dB show significant performance improvements over those techniques which use a single model for separation.

Keywords– Source separation, single channel speech separation, speaker identification, model-based single channel speech separation, Wiener filtering

1. INTRODUCTION

The human auditory system is able to pick one conversation out of dozens in a crowded room. This is a capability that no artificial system comes close to matching. Recently, many efforts have been carried out to mimic this fantastic human ability. Inspired by this, the separation of two speech signals received from one communication channel is a challenging topic in the speech processing context. Currently, blind source separation techniques [1]-[7] are commonly used in the speech separation problem. In fact, if the requirements of BSS methods are satisfied, these techniques separate out speech signals with higher accuracy in comparison with other state-of-the art techniques such as computational auditory scene analysis (CASA) [7]. One of these requirements is that the number of observations must be at least equal to the number of sources, a condition which is not held when we have just one microphone and two speakers. This drawback, which significantly confines the usefulness of the BSS techniques in the problem at hand, can be explained as follows. In the BSS context, the separation of I source speech signals when we have access to J observation signals can be formulated as

$$\boldsymbol{Y}^t = \boldsymbol{A}\boldsymbol{X}^t$$

*Received by the editors June 24, 2006; final revised form May 1, 2007.
**Corresponding author

where $\boldsymbol{Y}^t = [\boldsymbol{y}_1^t, \ldots, \boldsymbol{y}_j^t, \ldots, \boldsymbol{y}_J^t]^T$, $\boldsymbol{X}^t = [\boldsymbol{x}_1^t, \ldots, \boldsymbol{x}_i^t, \ldots, \boldsymbol{x}_I^t]^T$ and $\boldsymbol{A} = [a_{i,j}]_{I,J}$ is an $(I \times J)$ instantaneous mixing matrix which shows the relative position of the sources from the observations. Also, vectors $\boldsymbol{y}_j^t = \{y_j^t(n)\}_{n=1}^N$ and $\boldsymbol{x}_i^t = \{x_i^t(n)\}_{n=1}^N$ for $j = 1,2,\ldots,J$ and $i = 1,2,\ldots,I$ represent N-dimensional vectors of the j^{th} observation and i^{th} source signals, respectively. Additionally, $[\cdot]^T$ denotes the transpose operation and the superscript t indicates that the signals are in the time domain. When the number of observations is equal or greater than the number of sources ($J > I$), the solution to the separation problem is simply obtained by estimating the inverse of the mixing matrix, i.e. $\boldsymbol{W} = \boldsymbol{A}^{-1}$, and left multiplying both sides of the above equation by $\boldsymbol{W}$. Many solutions have, so far, been proposed for determining the mixing matrix and quite satisfactory results have been reported [1]-[6].

However, when the number of observations is less than the number of sources ($J < I$), (e.g. $J = 1$ and $I = 2$ for the case discussed in this paper) the mixing matrix is non-invertible such that the problem becomes too ill-conditioned to be solved using common BSS techniques. In this case, we need auxiliary information (e.g. a priori knowledge of sources) to solve the problem. This problem is commonly referred to as model-based single channel speech separation and has recently become a hot topic in the signal processing realm [8]. Although several solutions to this crux problem have been proposed by including the a priori knowledge of underlying speakers into the separation system [9]-[24], the problem has still remained a challenge such that current proposed algorithms deliver acceptable quality only in special cases. Generally, single channel model-based speech separation techniques are categorized into two classes: time domain and frequency domain.

In time domain techniques [9]-[13] each source is decomposed into independent basis functions in the training phase. The basis functions of each source are learnt from a training data set based on independent component analysis approaches. Then the trained basis functions along with the constraint imposed by the linearity of sources in the time domain are used to estimate the individual speech signals via a maximum likelihood optimization. While the techniques perform well when the speech signal is mixed with other sounds such as music, separability reduces significantly when the mixture consists of two speech signals since the learnt basis functions of two speakers overlap greatly. In frequency domain techniques [14]-[19], first a statistical model is fitted to the log spectral vectors of each speaker. Then, the two speaker models are combined to model the mixed signal. Finally, in the test phase, the states that best match the mixed signal are decoded based on some criteria (e.g., minimum mean square error, likelihood ratio).

In addition to the above approaches several works have been proposed from the audiology society who try to develop approaches based on human auditory mechanisms; the techniques are commonly referred to as computational auditory scene analysis (CASA) [25]-[31]. Though these methods are much faster and somehow simpler than model-based techniques, they suffer from two main problems. First, the current methods are unable to separate unvoiced speech and second, the separated speech signal suffers from crosstalk effects. Moreover, several techniques have been proposed that are categorized neither as BSS nor CASA methods [20]-[22]. In [20], a work has been presented based on neural networks and an extension of the Kalman filter. In [21] and [22], a generalized Wiener filter and an autoregressive model have been applied for general signal separation, respectively. The techniques have a mathematical depth that is worth further exploration, but no comprehensive results have been reported on the performance of these systems on speech signals. The previous model-based single channel separation techniques separate the sound signals with reasonable accuracy only for two special cases: the first when the mixture sound consists of a speech signal plus a non-speech signal, e.g. non-stationary noise or music, and the second when the system was trained for two known speakers. In the latter case, the separation system is speaker dependent such that the generality of the system is remarkably confined, though the separation results are impressive.

In this paper, we propose a new model-based single channel technique that not only takes on the advantages of speaker-dependent model-based approaches but also is able to separate the speech signals even if they come from unknown speakers. The system can be adapted to as many speakers as possible given a training data set of the speakers. The proposed technique consists of three stages: classification, identification, and separation. The algorithm first recognizes the underlying speakers, and then the trained models of the selected speakers are used in the separation process. We apply a new separation technique which employs Gaussian mixture modeling, maximum likelihood estimation and Wiener filtering to separate the speech signals. The classification stage is based on a new algorithm known as the harmonic matching classifier followed by the identification stage. We evaluate the performance of the whole system as well as the performance of each stage separately. Results show the proposed technique outperforms those techniques which apply a single trained model for all speakers.

The remainder of this paper is organized as follows. In Section 2, we present a brief overview of the whole system. In Section 3, we discuss the classification stage. The identification process is given in Section 4 followed by the separation system which is explored in Section 4. Experimental results are reported in Section 5 and, finally, conclusions are discussed in Section 6.

2. MODEL OVERVIEW

In this section, we present a brief overview of the proposed technique and in the subsequent sections we elaborate on the details of the algorithm. Fig. 1 shows the system's block diagram which consists of three stages: classification, identification, and separation.

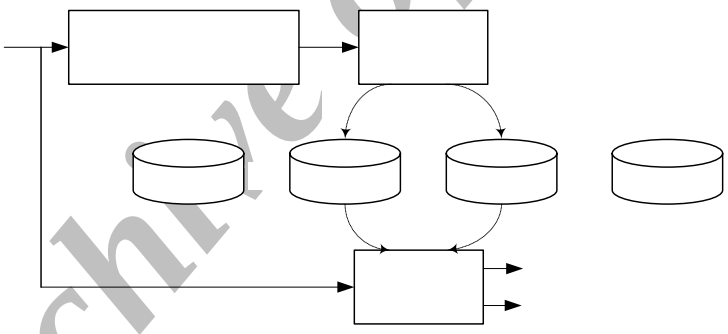

Fig. 1. Schematic diagram of the proposed system. The system consists of classification, identification, and separation stages

The task of the classification stage is to extract the segments by which we can identify the speakers' identity. From the human speech production mechanism, we know that the speech signal is generally categorized into voiced (V) and unvoiced (U) segments. Consequently, the mixed speech contains U-U, U-V, and V-V segments among which the U-V segments are extracted and passed to the speaker identification stage. We use the U-V frames for speaker identification because, in this case, the unvoiced frames are nearly masked by the voiced frames whose energy contents are generally greater than unvoiced frames. Hence, a U-V frame contains information related to one speaker which is appropriate for identification. The classification stage consists of two parts; the first part recognizes the U-U frames from the U-V and V-V frames and the second part distinguishes the U-V frames from V-V frames. For the first part of the classification stage, we use the technique introduced in [32]. While this technique has essentially been designed to classify the voiced and unvoiced frames in the single-talker scenario, we found through simulations that the technique accurately recognizes the U-U frames from the U-V and V-V frames. For the second part of the classification stage, we introduce a new technique which we call the harmonic matching classifier.

The extracted features (i.e U-V frames) are then transformed to the mel frequency cepstral coefficients (MFCC) and passed to the identification stage. The task of the identification stage is to identify the two speakers among the N speakers. We apply a speaker identification algorithm based on the techniques known as VQ-based speaker identification [33]-[35]. These techniques, however, are designed to identify one speaker among the N speakers. Therefore, we modify the VQ-based speaker identification algorithms to be able to recognize two speakers among the N speakers.

Finally, the last stage of the proposed algorithm is intended to separate the underlying speech signals of the identified speakers. In this stage, we introduce a new technique which applies Gaussian mixture modeling, maximum likelihood estimation, and Wiener filtering to separate speech signals. A block diagram of the proposed separation algorithm is shown in Fig. 2. In this stage, the power spectrum density (PSD) of the underlying speech signals is estimated in a maximum likelihood estimation process at the frame level. Then the estimated PSDs are fed into the Wiener filter so as to estimate the speech signals. In the following sections we present the details of these algorithms.

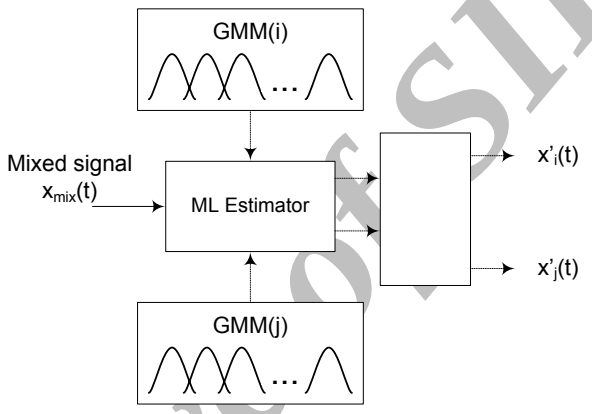

Fig. 2. Block diagram of the separation stage which separates the underlying speech

3. CLASSIFICATION STAGE

As mentioned earlier the task of the classification stage is to recognize the U-V frames which are appropriate for speaker identification. Fig. 3 shows a schematic of the proposed classifier which consists of two stages. At the first stage, we distinguish the U U frames from V-V and U-V frames, and at the second stage the U-V frames are recognized from the V-V frames.

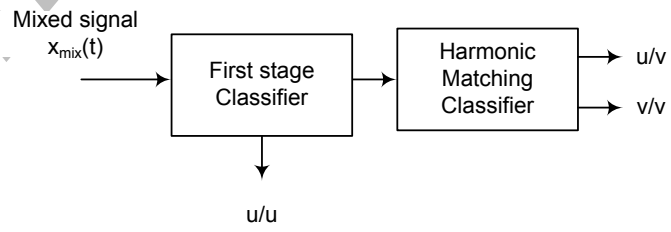

Fig. 3. Block diagram of the classification stage consisting of two parts: the U-U classifier and the harmonic match classifier to recognize U-V from V-V

The first classification stage is not a tough task such that common single speech classification techniques can be effectively used for the co-channel case as well. Therefore, for the first stage, we use the algorithm proposed by Talkin [32] which not only performs a pitch detection task, but also accomplishes voiced and unvoiced classification. However, the difficult part of the classification process is to

distinguish the V-V frames from the U-V frames. Recently, this topic has been dubbed in speech separation literature as usable speech detection [36]. The main application of usable speech detection is in the co-channel speaker identification problem where the aim is to recognize the U-V frames whereby speaker identities are recognized. Several approaches have been proposed for distinguishing U-V frames from V-V frames, namely using the spectral autocorrelation peak valley ratio (SAPVR) criterion [37], nonlinear speech processing [38], wavelet analysis [39], Bayesian classifiers [40], or pitch information [41]. In this paper, we introduce a new technique which we call the harmonic matching classifier (HMC). The algorithm is explained in the following paragraph.

As mentioned earlier, the main characteristic of voiced frames, i.e. periodic nature, are preserved when a voiced frame interacts with an unvoiced frame in the mixed speech signal. In this case, we can fit a harmonic model to a U-V analysis frame with a modeling error which is considerably less than that of fitting a harmonic model to a V-V frame. In the latter case, harmonic modeling just covers the spectral peaks belonging to one speaker and thus leads to a high modeling error. The detailed algorithm for extracting the U-V frames is described in Table I. In this algorithm, $|X_{mix}^{t}(\omega)|^2$ denotes the spectrum of the t^{th} mixed signal frame. Moreover, the harmonic model is represented by $\sum_{l=1}^{L(\omega_i)} A_{l\omega_i}^2 W^2(\omega - l\omega_i)$, in which the applied spectrum window, $W(\omega)$, is repeated at integer multiples of the fundamental frequency ω_i with an amplitude proportional to $A_{l\omega_i}$. Also $L(\omega_i)$ represents the number of harmonics in the speech bandwidth. For each frame we find the best harmonic match and compute the model error. If the corresponding error is less than the threshold σ, the frame is classified as a U-V frame, otherwise it is a V-V frame. Using a training data set we obtained the best value for $\sigma = mean(\{e^t\}_{t=1}^{T})$, where e^t is the model error for the frame *t* (see Table 1 for more details). We report the performance of the harmonic matching classifier along with a comparison with a state-of-the-art technique in Section 5. Finally, it should be noted that although the silence segment in a classification process is desirable, but in this paper we consider the silence segments as a special case of unvoiced segments.

Table 1. Harmonic matching classifier

- Find error introduced by fitting a harmonic model to the t[th] mixed analysis frame

$$e^t = \min_{\omega_i^*} \left| |X_{mix}^t(\omega)|^2 - \sum_{l=1}^{L(\omega_i)} A_{l\omega_i}^2 W^2(\omega - l\omega_i) \right|$$

- V-V and U-V classification

if e^t ¡Ü σ
  frame ¡ú U-V
else
  frame ¡ú V-V
end

- Repeat the algorithm for all frames

t=1,2,…,*T*

Let $\Theta = \{1,2,\ldots,S\}$ be a group of speakers among whom we wish to identify the two speaker identities given a mixed utterance. Having the training data set for each speaker, we first partition the feature space (MFCCs) of each speaker into *K* partitions using the Linde-Buzo-Gray (LBG) vector quantization algorithm [42]. Then, partition centers $\boldsymbol{c}_k^i$ (known as codewords) are extracted and form the speaker *i* codebook $\Psi^i = \{\boldsymbol{c}_1^i, \boldsymbol{c}_2^i, \ldots, \boldsymbol{c}_K^i\}$. Each codeword, in turn, contains the first *M* MFCCs (excluding the first one) that is, $\boldsymbol{c}^i = [c^i(1), c^i(2), \ldots, c^i(M)]^T$, where $[\cdot]^T$ denotes the transpose operation.

Accordingly, performing quantization on the training data set of all speakers we obtain the set $\Psi = \{C^1, C^2, \ldots C^S\}$ consisting of all speakers' codebooks. Now the objective is to find two speakers by minimizing the following criteria

$$\arg\min_{i^* \in \Theta} \sum_{t \in U-V} \min_k D(\boldsymbol{c}_{mix}^t, \boldsymbol{c}_k^i) \tag{1}$$

$$\arg\min_{j^* \in \Theta - \{i^*\}} \sum_{t \in U-V} \min_k D(\boldsymbol{c}_{mix}^t, \boldsymbol{c}_k^j) \tag{2}$$

where i^* and j^* are selected speakers and $\boldsymbol{c}_{mix}^t$ mix is the MFCC vector extracted from the t^{th} U-V mixed analysis frame. Additionally, $D(\boldsymbol{c}_{mix}^t, \boldsymbol{c}_k^i)$ represents the Euclidean distance between the vectors $\boldsymbol{c}_{mix}^t$ and $\boldsymbol{c}_k^i$ as defined by

$$D(\mathbf{c}_{mix}^t, \mathbf{c}_k^i) = \sum_{m=1}^{M} (c_{mix}^t(m) - c_k^i(m))^2 \tag{3}$$

Equations (1) and (2) can be interpreted as follows. First, for those frames recognized as the U-V frames in the mixed speech signal, a search is done through all codewords (k=1:K) of all speakers' codebooks (i=1:S). The speaker who obtains the minimum distortion measure for all U-V frames is selected as the first underlying speaker. A similar process is again performed to select the second underlying speaker but the selected speaker from the first search process is excluded from the searching process.

4. SEPARATION STAGE

a) Relation between the log spectra vectors

In this subsection, we assume that the two speakers whose utterances form the mixed speech signal were specified from the identification stage. Let $x_1(t)$ and $x_2(t)$ be the speech signal of speaker one and two, respectively. An N-dimensional vector of samples of $x_1(t)$ and $x_2(t)$ at time m are denoted by

$$\boldsymbol{x}_1^t(m) = [x_1^t(m), x_1^t(m+1), \ldots, x_1^t(m-N+1)]^T \tag{4}$$

$$\boldsymbol{x}_2^t(m) = [x_2^t(m), x_2^t(m+1), \ldots, x_2^t(m-N+1)]^T \tag{5}$$

where $[\cdot]^T$ denotes the transpose operation and the superscript t denotes the time domain notation. We assume that the observed signal $\boldsymbol{y}^t(m)$ is the sum of the speech signals of the two speakers as follows

$$\boldsymbol{y}^t(m) = \boldsymbol{x}_1^t(m) + \boldsymbol{x}_2^t(m). \tag{6}$$

We next form the following vectors

$$\boldsymbol{x}_1 = \log_{10}(|F_D(\boldsymbol{x}_1^t(m))|) = [x_1(1), \ldots, x_1(2), \ldots, x_1(D)]^T \tag{7}$$

$$\boldsymbol{x}_2 = \log_{10}(|F_D(\boldsymbol{x}_2^t(m))|) = [x_2(1), \ldots, x_2(2), \ldots, x_2(D)]^T \tag{8}$$

$$\boldsymbol{y} = \log_{10}(|F_D(\boldsymbol{y}^t(m))|) = [y(1), \ldots, y(2), \ldots, y(D)]^T \tag{9}$$

where $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{y}$ denote the D-dimensional log spectral vectors of speaker one and speaker two, the mixed signal, $F_D(\cdot)$, denotes the D-point discrete Fourier transform, and $|\cdot|$ denotes the magnitude operator. The relation between the log spectral vectors of the mixed signal and those of the individual signals can be expressed by the Log Max approximation. This approximation was first used in the context of robust speech recognition by Nadas *et al*. [43]. In [44], we have shown that this approximation is, in fact, a non-linear minimum mean square error estimator for phase information and implies that the log

spectrum of the mixed signal is nearly the element wise maximum of the log spectrum of the two underlying signals. Mathematically, the approximation can be formulated as follows

$$\hat{\boldsymbol{y}} = Max(\boldsymbol{x}_1, \boldsymbol{x}_2) = [\max(x_1(1), x_2(1)), \ldots, \max(x_1(d), x_2(d)), \ldots, \max(x_1(D), x_2(D))]^T \quad (10)$$

Hence, $\hat{\boldsymbol{y}}$ is an approximation to $\boldsymbol{y}$ with reasonable accuracy. It should be noted that MFFC coefficients used in the identification stage cannot be used in the separation stage for two reasons. First, the MFCC is a non linear feature such that we cannot re-synthesize the speech signal from the MFCC coefficients. For this reason, although MFFC is widely used in classification based techniques such as speech or speaker recognition, we cannot use it for re-synthesizing a speech signal. Second, there is no straightforward relationship between the MMFSs of the mixture and those of the underlying signals.

b) Maximum likelihood estimator

We next model the probability density function of the i^{th} speaker's log spectral vectors by a mixture of K_i Gaussian densities in the following form

$$f_{\boldsymbol{x}_i}(\boldsymbol{x}_i) = \sum_{k=1}^{K_i} c_{x_i,k} N(\boldsymbol{x}_i, \mu_{x_i,k}, \boldsymbol{U}_{x_i,k}) \quad i \in \{1,2\} \quad (11)$$

where $c_{x_i,k}$ represents the a priori probability for the k^{th} Gaussian in the mixture and satisfies $\sum_k c_{x_i,k} = 1$, and

$$N(\boldsymbol{x}_i, \mu_{x_i,k}, \boldsymbol{U}_{x_i}) = \frac{\exp(-\frac{1}{2}(\boldsymbol{x}_i - \mu_{x_i,k})^T \boldsymbol{U}_{x_i,k}^{-1}(\boldsymbol{x}_i - \mu_{x_i,k}))}{\sqrt{(2\pi)^D \,|\boldsymbol{U}_{x_i,k}|}} \quad (12)$$

represents a D-dimensional normal density function with the mean vector $\mu_{x_i,k}$ and covariance matrix $\boldsymbol{U}_{x_i,k}$. The D-variant Gaussians are assumed to be diagonal covariant to reduce the order of computation. This assumption enables us to represent the multivariate Gaussian as the product of D uni-variant Gaussians given by

$$f_{x_i}(\boldsymbol{x}_i) = \sum_{k=1}^{K_i} c_{x_i,k} \prod_{d=1}^{D} \frac{\exp\left(-\frac{1}{2}\left(\frac{x_i(d) - \mu_{x_i,k}(d)}{\sigma_{x_i,k}(d)}\right)^2\right)}{\sqrt{2\pi}\sigma_{x_i,k}(d)} \quad (13)$$

where, $x_i(d)$, $\mu_{x_i,k}(d)$ and $\sigma^2_{x_i,k}(d)$ are the d^{th} component of $\boldsymbol{x}_i$, d^{th} component of the mean vector, and the d^{th} element on the diagonal of the covariance matrix, respectively.

As mentioned earlier, the log spectral vectors of the mixed signal are almost exactly the maximum element-wise components of the log spectral vectors of the underlying signal, that is

$$\mathbf{y} \approx \max(\mathbf{x}_1, \mathbf{x}_2). \quad (14)$$

The cumulative distribution function (CDF) of the mixed log spectral vectors $F_y(\boldsymbol{y})$ is given by

$$F_y(\boldsymbol{y}) = F_{x_1,x_2}(\boldsymbol{y}, \boldsymbol{y}) \quad (15)$$

where $F_{x_1,x_2}(\boldsymbol{y}, \boldsymbol{y})$ is the joint CDF of the random vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Since the speech signals of the two speakers are independent, then

$$F_y(\boldsymbol{y}) = F_{x_1}(\boldsymbol{y}) \times F_{x_2}(\boldsymbol{y}). \quad (16)$$

Thus $f_y(\boldsymbol{y})$ is obtained by differentiating both sides of Eq. (16) to give

$$f_y(\boldsymbol{y}) = f_{x_1}(\boldsymbol{y})F_{x_2}(\boldsymbol{y}) + f_{x_2}(\boldsymbol{y})F_{x_1}(\boldsymbol{y}). \tag{17}$$

The CDF to express $F_{x_i}(\boldsymbol{y})$ is obtained by

$$F_{x_i}(\mathbf{y}) = \int_{-\infty}^{y} f_{x_i}(\xi)d\xi = \int_{-\infty}^{y(d)} \sum_{k=1}^{K_i} c_{x_i,k} \prod_{d=1}^{D} \left[\frac{1}{\sigma_{x_i,k}(d)\sqrt{2\pi}} \times \exp\left(-\frac{1}{2}\left(\frac{\xi_d - \mu_{x_i,k}(d)}{\sigma_{x_i,k}(d)}\right)^2\right) \right] d\xi_d \tag{18}$$

Since the integration of the sum of the exponential functions is identical to the sum of the integral of exponentials as well as assuming a diagonal covariance matrix for the distributions, we conclude that

$$F_{x_i}(\boldsymbol{y}) = \sum_{k=1}^{K_i} c_{x_i,k} \prod_{d=1}^{D} \left[\frac{1}{\sigma_{x_i,k}(d)\sqrt{2\pi}} \times \int_{-\infty}^{y(d)} \exp\left(-\frac{1}{2}\left(\frac{\xi_d - \mu_{x_i,k}(d)}{\sigma_{x_i,k}(d)}\right)^2\right) d\xi_d \right] \tag{19}$$

The term in the bracket in Eq. 19 is often expressed in terms of the error function

$$erf(\alpha) = \frac{1}{\sqrt{2\pi}} \int_0^{\alpha} \exp\left(-\frac{1}{2}\nu^2\right) d\nu \tag{20}$$

Thus, we conclude that

$$F_{x_i}(\boldsymbol{y}) = \sum_{k=1}^{K_i} c_{x_i,k} \prod_{d=1}^{D} \left[erf(z_{x_i,k}(d)) + \frac{1}{2} \right] \tag{21}$$

where

$$z_{x_i,k}(d) = \frac{y(d) - \mu_{x_i,k}(d)}{\sigma_{x_i,k}(d)} \quad i \in \{1,2\} \tag{22}$$

Finally, we obtain the PDF of the log spectral vectors of the mixed signal by substituting Eq. (13) and Eq. (21) into Eq. (17) to give

$$\begin{aligned} f_{\mathbf{y}}(\mathbf{y}) = \sum_{k=1}^{K_1}\sum_{l=1}^{K_2} c_{x_1,k} c_{x_2,l} \times \Bigg(& \prod_{d=1}^{D} \left[(2\pi\sigma_{x_1,k}^2(d))^{\frac{1}{2}} \times \left(erf(z_{x_2,l}(d)) + \frac{1}{2}\right) \times \exp\left(\frac{1}{2} z_{x_1,k}^2(d)\right) \right] \\ & + \prod_{d=1}^{D} \left[(2\pi\sigma_{x_2,l}^2(d))^{\frac{1}{2}} \times \left(erf(z_{x_1,k}(d)) + \frac{1}{2}\right) \times \exp\left(\frac{1}{2} z_{x_2,l}^2(d)\right) \right] \Bigg) \end{aligned} \tag{23}$$

Equation (23) gives the PDF of log spectral vectors for the mixed signal in terms of the mean and variance of the log spectral vectors of the underlying signals.

Now we apply $f_y(\boldsymbol{y})$ in a maximum likelihood framework to estimate the parameters of the underlying signals. The main objective of the Maximum Likelihood estimator is to find the k^{th} Gaussian in $f_{x_1}(\boldsymbol{x}_1;\lambda_{x_1})$ and the l^{th} Gaussian in $f_{x_2}(\boldsymbol{x}_2;\lambda_{x_2})$ such that $f_y(\boldsymbol{y})$ is maximized. The estimator is given by

$$\{\hat{k},\hat{l}\}_{ML} = \arg\max_{\theta_{k,l}} f_y(\boldsymbol{y} \mid \theta_{k,l}) \tag{24}$$

where

$$\theta_{k,l} = \{\mu_{x_1,k}, \mu_{x_2,l}, \sigma_{x_1,k}, \sigma_{x_2,l}\} \tag{25}$$

The estimated mean vectors are then passed to the Wiener filtering stage to estimate the underlying speech signals.

c) Wiener filtering

From the Wiener filtering theory [45], we know that the optimal filter for stationary processes that can estimate a signal corrupted by noise (for our case the term noise means the other speaker's signal) is given by

$$\left|F_D(\mathbf{x}_1^t(m))\right|^2 \approx \frac{S_{x_1}(\omega)}{S_{x_1}(\omega)+S_{x_2}(\omega)} S_y(\omega) \tag{26}$$

where $S_{x_1}(\omega)$, $S_{x_2}(\omega)$, and $S_y(\omega)$ are the power spectral densities associated with speaker one, speaker two, and the mixed signal, respectively. Approximation to $\left|F_D(\boldsymbol{x}_2^t(m))\right|^2$ is also obtained in a similar way. In Eq. (26), however, we have no access to the speakers' PSDs, so we replace them by the estimated log spectral vectors, i.e. $\mu_{x_1,k}$ and $\mu_{x_2,l}$, obtained from the previous subsection (Eq. (25)). Thus, we have

$$\left|F_D(\mathbf{x}_1^t(m))\right|^2 \approx \frac{10^{2\mu_{x_1,l}(\omega)}}{10^{2\mu_{x_1,l}(\omega)}+10^{2\mu_{x_2,k}(\omega)}} S_y(\omega) \tag{27}$$

Finally, the estimated signals are obtained in the time domain by

$$\hat{x}_i^t(m) = F_D^{-1}\left(\left|F_D(x_i^t(m))\right| \angle F_D(y^t(m))\right) \quad i \in \{1,2\} \tag{28}$$

where F_D^{-1} denotes the inverse Fourier transform and $\angle F_D(\mathbf{y}^t(m))$ is the phase of the Fourier transform of the mixed signal. In this way, we obtain an estimate of $\boldsymbol{x}_i^t(m)$. It should be noted that it is common to use the phase of the STFT of the mixed signal for reconstructing the individual signals [14]-[16], [27] as it has no palpable effect on the quality of the separated signals. Recently it has been shown that the phase of the short-time Fourier transform has valuable perceptual information when the speech signal is analyzed with a window of long duration, i.e., >1 sec. [46]. To the best of our knowledge no technique has been proposed to extract the individual phase values from the mixed phase.

5. EXPERIMENTAL RESULTS

In this section we report the results obtained from the performance evaluation of the proposed system. We use the corpus introduced in [48] which consists of 34 speakers, each of whom uttered 500 sentences. We randomly choose 15 speakers among the 34 speakers for our experiments. For each speaker, 400 out of 500 sentences are selected for the training phase where a 10 bit-codebook is extracted for the identification stage in the following manner. The sampling rate is decreased to 8 kHz from the original 25 khz rate. The training data are first pre-emphasized using a first order filter with α=0.97 and then windowed with a Hamming window of duration of 32 msec at a frame rate of 10 msec. Thereafter, 30 mel cepstral coefficients (excluding the first one) are extracted from each analysis frame. The lowest and highest band edges of mel filters was set to 50 Hz and 3200 Hz, respectively. The set of extracted MFCC vectors for each speaker are quantized to 1024 clusters whose centers are used for the identification process. Vector quantization is performed using the well-known LBG algorithm [42] with binary splitting initialization. For the test phase, we randomly select 200 sentences (not within the training data set) from 15 chosen speakers. Then, 100 mixed speech signals are created by digitally adding the underlying speech signals with the Target-to-Interference (TIR) ratios set to -9, -6, -3, 0, +3, +6, and +9 dB. TIR represents the ratio between the energies of the two underlying speech signals in terms of dB. We randomly select one of the speech signals as the target and the other as the interference.

We first conduct experiments in order to evaluate the performance of the classification stage. In order to put the results into perspective, we compare the voicing state classification results of our model with that of Wu *et al*. [47]. The technique proposed in [47] is, in fact, a multi-pitch tracking system which detects not only the underlying speakers' pitch values, but also voicing states. We set the Wu's multi-pitch tracking parameters from the package they provided and compare the results with our approach. Table 2 shows how the voiced and unvoiced frames interact in the 100 mixed speech signals. As the table shows 13.92, 39.19, and 46.89 percent of the mixed frames belong to U-U, U-V, and V-V states, respectively. In Tables 3 and 4 we present the confusion matrix for the classification results obtained from our approach and the method proposed in [47], respectively. Each diagonal entry of the matrix shows the number of paired frames that are classified correctly and off-diagonal entries show the number of misclassified paired frames. From Tables 3 and 4 we can observe that in our system with respect to the Wu's approach, the classification performance for the U-U, U-V, and V-V states, on average, have improved 8, 2, and 20 percentage respectively. The most difficult task in both our system and Wu's approach is to recognize the U-V frames from V-V frames, though our system has significantly decreased the error rate for this case. We noticed that two sources of error occur in the classifier. The first one is mainly due to the transitional regions where determining the voicing state, even in the single speech case, is a difficult task. The second source of misclassification happens when the pitch values of the underlying signals lie within the same range. It should be noted that, to the best of our knowledge, no method has so far been proposed to handle these circumstances.

Table 2. Interaction of states in the data base

U-U	U-V	V-V
2204(13.92%)	6206(39.19%)	7424(46.98%)

Table 3. Confusion matrix of voicing state classification for the method proposed in [47]

	U-U	U-V	V-V
U-U	2174(98.64%)	572(09.22%)	223(03.00%)
U-V	20(00.91%)	4762(76.73%)	1856(25.00%)
V-V	10(00.45%)	872(14.05%)	5345(72.00%)

Table 4. Confusion matrix of voicing state classification for the proposed method in this paper

	U-U	U-V	***V-V***
U-U	2002(90.83%)	869(14.00%)	608(08.19%)
U-V	180(08.17%)	4592(74.00%)	2955(39.80%)
V-V	22(01.00%)	745(12.00%)	3861(52.01%)

In order to evaluate the accuracy of the identification stage, we measure the correct speaker identification rate of the system for two groups of input features. First, with the MFCC coefficients of the U-V frames obtained from the classification stage; and second with the MFCC coefficients of all frames (without classification). 100 test mixed signals are fed to the speaker ID stage and the percentages of times in which the target and interference are correctly identified are computed. Figures 4 and 5 show the correct speaker ID rate for the target and interferences with and without the U-V frame extraction. We observe that the correct identification rate obtained from the U-V features, on average, outperforms that of the non-classified features.

We do, however, note that at high or low TIR values where the target or interference speakers' energy is remarkably greater than the other, the performance for non-classified features reaches that of the U-V features. This improvement can be justified as follows. From speaker recognition techniques we know that

the speaker ID rate has a direct relation with the length of the test utterance such that the more test speech applied, the better the identification performance obtained [49]. Let $n_{s_1}^c$ be the number of detected mixed frames in which speaker one is in the V state and speaker two in the U state. Also let $n_{s_1}^o$ be the number of original voiced frames of speaker one. If we assume that speaker one is the target signal, then when SSR increases, $n_{s_1}^c \to n_{s_1}^o$. Thus the performance of the system with/without classification becomes identical. The same justification can be made for the interference signal. Accordingly, as the TIR increases or decreases from zero the target or interference becomes the dominant speaker (i.e. $n_{s_1}^c \to n_{s_1}^o$), and consequently the performance of with/without classification approaches the same values.

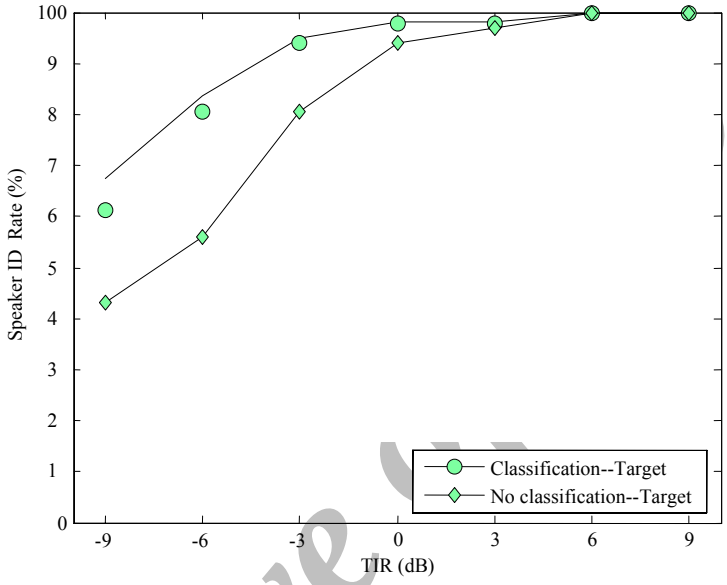

Fig. 4, Correct speaker ID rate versus TIR ratio for target speech signals

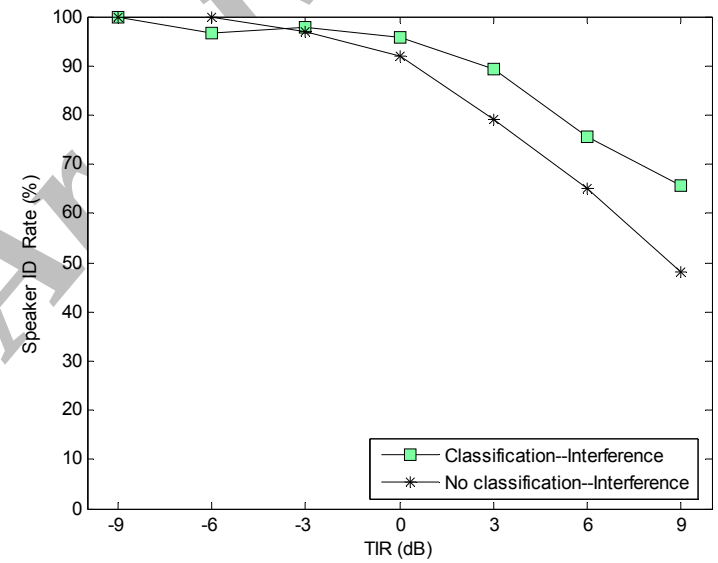

Fig. 5. Correct speaker ID rate versus TIR ratio for interference speech signals

The last and the most important stage of the system is the separation stage. In this stage, we first model the spectral space of each speaker using a mixture of Gaussian densities. The spectral vectors are extracted from the segments obtained by applying a Hamming window with a length of 52 msec at a frame

rate equal to 10 msec. In [50], we showed that for model based single channel speech separation algorithms, this window length leads to the optimal separation performance. Then, a 512-point discrete Fourier transform (D=512) is applied to the windowed segments, resulting in spectral vectors of dimension 256 (symmetric part was discarded). In order to fit a mixture of Gaussian densities to each speaker's feature space, we first tried to apply the Expectation-Maximization approach which is commonly used for GMM training. Unfortunately we encountered two problems that caused the training procedure to be intractable. First, the feature vector's dimension is remarkably higher than that used in other applications (e.g. Speech recognition, identification) where a vector with 20-40 elements is applied. Second, we need to train a GMM with a large number of elements (we use 256 elements) since we want to recover the underlying speech signals from the mean vectors of Gaussians. Hence, we found that for 15 minute training data it is time consuming to train a GMM with the above specifications using the available software. Therefore, we use a semi-continuous GMM model trained in the following manner. We assume that all components are equal probable. In addition, the Gaussians mean vectors are obtained using an 8 bit-codebook and Gaussians covariance matrixes are obtained from computing the sample covariance matrix of each cluster. To further decrease the computational burden, we just use the diagonal components of the sample covariance matrixes.

In order to show the superiority of speaker-dependent separation modeling over speaker-independent separation modeling, we also fit a GMM to the training data of all speakers. We quantify the degree of the separability by computing the SNR between the separated and the original signals in the time domain. The SNR value for the separated speech signal of the i^{th} speaker is defined as

$$\mathrm{SNR}_i = 10\log_{10}\left[\frac{\sum_n x_i^2(n)}{\sum_n (x_i(n)+\hat{x}_i(n))^2}\right] \qquad n=1,2,\ldots,N \tag{29}$$

where $x_i(n)$ and $\hat{x}_i(n)$ are the original and separated speech signals of length N, respectively.

Figures 6 and 7 show the SNR results versus the TIR ratio averaged over 20 separated utterances for the target and interference speeches, respectively. The circled and squared lines show the results for speaker-dependent modeling (multiple database) and speaker-independent modeling (single database), respectively. The results are shown for both target (Fig. 6) and interference (Fig. 7) speeches. From Figs. 6 and 7, we observe that, on average, there is a 3.5 dB performance gain over the speaker-independent scenario. This improvement is remarkable in current single channel speech separation techniques.

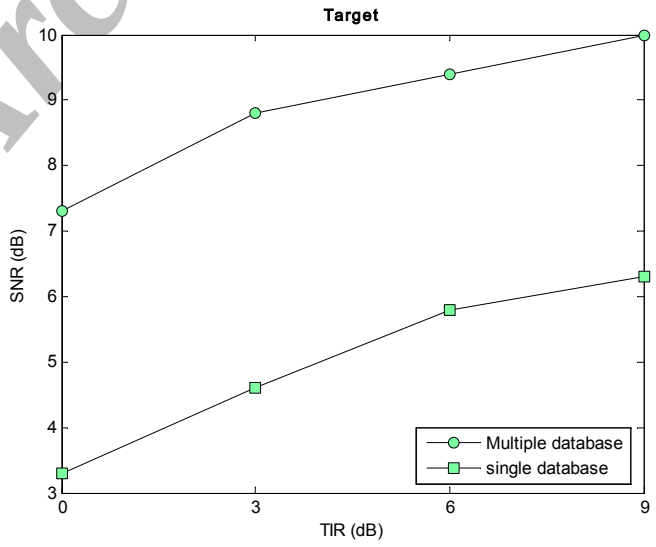

Fig. 6. SNR versus TIR ratio averaged over separated target speech signals

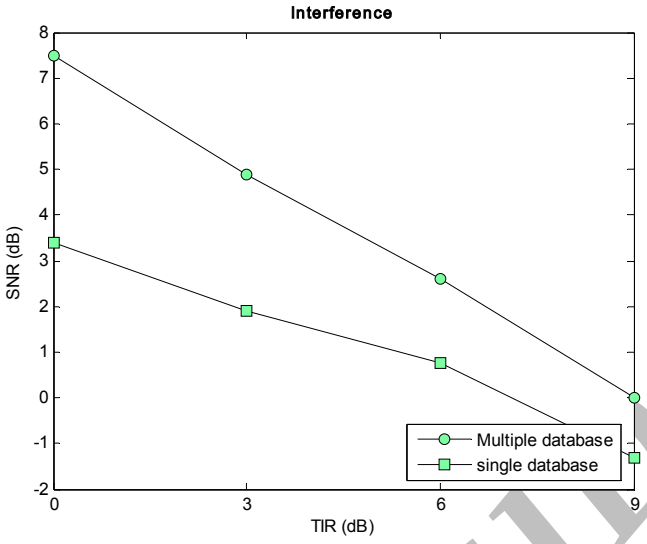

Fig. 7. SNR versus TIR ratio averaged over separated interferences speech signals

6. CONCLUSIONS

In this paper, we have presented a new model-based single channel speech separation technique. This technique can be effectively applied to separate two speech signals from their mixture where the common single channel separation techniques fail to handle the problem. The proposed technique not only preserves the advantages of speaker dependent single channel speech separation algorithms, but is also able to separate the speech signals of an unlimited number of speakers given the speakers' models. The speaker databases can be augmented into the system in an adaptation phase. The proposed technique consists of three stages: classification, identification, and separation. The speakers' identities are first determined using the classification and identification stages. Then, the identified speakers' models are used to separate the underlying signals. The performance of classification, identification and separation were evaluated and compared with current algorithms. The obtained results also support the idea that the human auditory system uses a priori knowledge about the concurrent sounds to separate them. We believe the next step in this research should be to first improve the identification accuracy and second, adapt the system for a new speaker using the prevalent speaker adaptation techniques applied in speech recognition.

Acknowledgment- The authors would like to thank the Iran Ministry of Science and Research and the Natural Sciences and Engineering Research Council (NSERC) of Canada which partially funded this research.

REFERENCES

1. Jutten, C. & Herault, J. (1991). Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing, 24*, 1–10.
2. Common, P. (1994). Independent component analysis, a new concept? *Signal Processing, 36*, 287–314.
3. Bell, A. J. & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation, 7*, 1129–1159.
4. Amari, S. I. & Cardoso, J. F. (1997). Blind source separation–semiparametric statistical approach. *IEEE Trans. Signal Processing, 45*(11), 2692–2700.
5. Belouchrani, A. & Cardoso, J. F. (1994). On the performance of orthogonal source separation algorithm. *in EUSIPCO*, Edinburgh, Scotland, 768–771.

6. Burel, G. (1992). Blind separation of sources: a nonlinear neural algorithm. *Neural Networks, 5*, 937–947.
7. van der Kouwe, A. J. W., Wang, D. L. & Brown, G. J. (2001). A comparison of auditory and blind separation techniques for speech segregation. *IEEE Trans. Speech and Audio Processing, 9*(3), 189–195.
8. Ellis, D. (2006). *Model-based scene analysis*. in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. W. G. Brown, Ed. Wiley/IEEE Press in press.
9. Jang, G. J. & Lee, T. W. (2003). A probabilistic approach to single channel source separation. *in Proc. Advances in Neural Inform. Process. Systems*, 1173–1180.
10. [10] Fevotte, C. & Godsill, S. J. (2005). A Bayesian approach for blind separation of sparse sources. *IEEE Trans. on Speech and Audio Processing, 4*(99), 1–15.
11. Girolami, M. (2001). A variational method for learning sparse and overcomplete representations. *Neural Computation, 13*(11), 2517–2532.
12. Lee, T. W., Lewicki, M. S., Girolami, M. & Sejnowski, T. J. (1999). Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters, 6*(4), 87– 90.
13. Beierholm, T., Pedersen, B. D. & Winther, O. (2004). Low complexity Bayesian single channel source separation. *in Proc. ICASSP–04, 5*, 529–532.
14. Roweis, S. (2000). One microphone source separation. *in Proc. Neural Inf. Process. Syst.*, 793–799.
15. Reyes-Gomez, M. J., Ellis, D. & Jojic, N. (2004). Multiband audio modeling for single channel acoustic source separation. *Proc. ICASSP–04, 5*, 641–644.
16. Reddy, A. M. & Raj, B. (2004). A minimum mean squared error estimator for single channel speaker separation. *in INTERSPEECH*, 2445–2448.
17. Kristjansson, T., Attias, H. & Hershey, J. (2004). Single microphone source separation using high resolution signal reconstruction. *Proc. ICASSP*–05, 817–820.
18. Rowies, S. T. (2003). Factorial models and refiltering for speech separation and denoising. *EUROSPEECH–03, 7*, 1009–1012.
19. Radfar, M. H., Dansereau, R. M. & Sayadiyan, A. (2006). A novel low complexity VQ-based single channel speech separation technique. into appear in *IEEE International Symposium on Signal Processing and Information Technology*.
20. Wan, E. A. & Nelson, A. (1997). Neural dual extended kalman filtering: Applications in speech enhancement and monaural blind signal separation. *IEEE Proc. Neural Networks for Signal Processing,* 466–475.
21. Hopgood, J. R. & Rayner, P. J. W. (2003). Single channel non-stationary stochastic signal separation using linear time-varying filters. *IEEE Trans. Acoustics, Speech, and Signal Process, 51*(7), 1793–1752.
22. Balan, R., Jourjine, A. & Rosca, J. (1999). Ar processes and sources can be reconstructed from degenerative mixtures. *Proc. ICA-99*, 467–472.
23. Radfar, M. H., Dansereau, R. M. & Sayadiyan, A. (2006). A joint probabilistic-deterministic approach using source-filter modeling of speech signal for single channel speech separation. *Proc. IEEE MLSP-06*, 47–52.
24. Radfar, M. H., Dansereau, R. M. & Sayadiyan, A. (2006). Performance evaluation of three features for model-based single channel speech separation problem. *Interspeech 2006, Intern. Conf. on Spoken Language Processing* (ICSLP'2006 Pittsburgh), 2610–2613.
25. Bregman, A. S. (1994). *Computational auditory scene analysis*. Cambridge MA: MIT Press.
26. Brown, G. J. & Cooke, M. (1994). Computational auditory scene analysis. *Computer Speech and Language, 8*(4), 297–336.
27. Hu, G. & Wang, D. (2004). Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Trans. Neural Networks, 15*(5), 1135–1150.
28. Wang, D. L. & Brown, G. J. (1999). Separation of speech from interfering sounds based on oscillatory correlation. *IEEE Trans. Neural Networks, 10*, 684–697.

29. Virtanen, T. & Klapuri, A. (2000). Separation of harmonic sound sources using sinusoidal modeling. *Proc. ICASSP–2000*, 765–768.
30. Quatieri, T. F. & Danisewicz, R. G. (1990). An approach to co-channel talker interference suppression using a sinusoidal model for speech. *IEEE Trans. Acoustics, Speech, and Signal Process, 38*, 56–69.
31. Radfar, M. H., Sayadiyan, A. & Dansereau, R. M. (2006). Monaural multipitch tracking using joint mean square error harmonic modelling and sinusoidal spectrogram. submitted to *Speech Communication*.
32. Talkin, D. (1995). *Robust pitch tracking. in speech coding and synthesis*. W. Kleijn and K. Paliwal, Eds. Elsevier.
33. Kinnunen, T., Karpov, E. & Franti, P. (2006). Real-time speaker identification and verification. *IEEE Trans. Speech Audio Processing, 14*(1), 277–288.
34. Jialong, H., Li, L. & Palm, G. (1999). A discriminative training algorithm for VQ-based speaker identification. *IEEE Trans. Speech Audio Processing, 7*(3), 353–356.
35. Soong, F., Rosenberg, A., Rabiner, L. & Juang, B. (1985). A vector quantization approach to speaker recognition. *Proc. ICASSP-85, 10*, 387–390.
36. Yantorno, R. E. (1999). Co-channel speech study, Air Force Office of Scientific Research Speech Processing Lab Rome Labs. Report for Summer Research Faculty Program.
37. Chandra, N. & Yantorno, R. E. (2002). Usable speech detection using the modified spectral autocorrelation peak to valley ratio using the LPC residual. *Proc. 4th IASTED–02*, 146–149.
38. Mahgoub, Y. & Dansereau, R. (2005). Voicing-state classification of cochannel speech using nonlinear state-space reconstruction. *Proc. ICASSP–05, 1*, 409–412.
39. Kizhanatham, A. R., Chandra, N. & Yantorno, R. E. (2002). Co-channel speech detection approaches using cyclostationarity or wavelet transform. *Proc. IASTED-02*.
40. Benincasa, D. S. & Savic, M. I. (1998). Voicing state determination of cochannel speech. *Proc. ICASSP–98, 2*, 1021–1024.
41. Shao, Y. & Wang, D. L. (2003). Co-channel speaker identification using usable speech extraction based on multi-pitch tracking. *Proc. ICASSP-03, 2*, 205–208.
42. Gersho, A. & Gray, R. M. (1992). *Vector quantization and signal compression*. Norwell MA: Kluwer Academic.
43. Nadas, A., Nahamoo, D. & Picheny, M. A. (1989). Speech recognition using noise-adaptive prototypes. *IEEE Trans. Acoust. Speech Sig. Process., 37*(10), 1495–1503.
44. Banihashemi, M. H. R. A., Dansereau, R. M. & Sayadiyan, A. (2006). A non-linear minimum mean square error estimator for the mixture-maximization approximation. *Electronic Letters, 42*(12), 75–76.
45. Papoulis, A. (1991). *Probability, random variables, and stochastic processes*. McGraw-Hill.
46. Paliwal, K. K. & Alsteris, L. D. (2005). On the usefulness of stft phase spectrum in human listening tests. *Speech Communication, 45*(2), 153–170.
47. Wu, M., Wang, D. L. & Brown, G. J. (2003). A multipitch tracking algorithm for noisy speech. *IEEE Trans. Acoustics, Speech, and Signal Process, 11*(3), 229–241.
48. Cooke, M. P., Barker, J., Cunningham, S. P. & Shao, X. (2005). An audiovisual corpus for speech perception and automatic speech recognition. JASA, http://www.dcs.shef.ac.uk/spandh/gridcorpus.
49. Campbell, J. & Reynolds, D. A. (1999). Corpora for the evaluation of speaker recognition systems. *Proc. ICASSP-99, 2*, 829–832.
50. Radfar, M. H., Dansereau, R. M. & Sayadiyan, A. On the choice of window size in model-based single channel speech separation. *Proc. of the IEEE Canadian Conf. on Elec. and Comp. Eng, 1*, 981–984.