Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2007, Article ID 84186, 15 pages
doi:10.1155/2007/84186

Research Article

A Maximum Likelihood Estimation of Vocal-Tract-Related Filter Characteristics for Single Channel Speech Separation

Mohammad H. Radfar,[1] Richard M. Dansereau,[2] and Abolghasem Sayadiyan[1]

[1] *Department of Electrical Engineering, Amirkabir University, Tehran 15875-4413, Iran*
[2] *Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada K1S 5B6*

Received 3 March 2006; Revised 13 September 2006; Accepted 27 September 2006

Recommended by Lin-Shan Lee

We present a new technique for separating two speech signals from a single recording. The proposed method bridges the gap between *underdetermined blind source separation* techniques and those techniques that model the human auditory system, that is, *computational auditory scene analysis* (CASA). For this purpose, we decompose the speech signal into the excitation signal and the vocal-tract-related filter and then estimate the components from the mixed speech using a hybrid model. We first express the probability density function (PDF) of the mixed speech's log spectral vectors in terms of the PDFs of the underlying speech signal's vocal-tract-related filters. Then, the mean vectors of PDFs of the vocal-tract-related filters are obtained using a *maximum likelihood* estimator given the mixed signal. Finally, the estimated vocal-tract-related filters along with the extracted fundamental frequencies are used to reconstruct estimates of the individual speech signals. The proposed technique effectively adds vocal-tract-related filter characteristics as a new cue to CASA models using a new grouping technique based on an underdetermined blind source separation. We compare our model with both an underdetermined blind source separation and a CASA method. The experimental results show that our model outperforms both techniques in terms of SNR improvement and the percentage of crosstalk suppression.

1. INTRODUCTION

Single channel speech separation (SCSS) is a challenging topic that has been approached by two primary methods: *blind source separation* (BSS) [1–4] and *computational auditory scene analysis* (CASA) [5–13]. Although many techniques have so far been proposed in the context of BSS or CASA [12–28], little work has been done to connect these two topics. In this paper, our goal is to take advantage of both approaches in a hybrid probabilistic-deterministic framework.

Single channel speech separation is considered an underdetermined problem in the BSS context since the number of observations is less than the number of sources. In this special case, common BSS with independent component analysis (ICA) techniques fails to separate sources [1–4] due to the noninvertibility of the mixing matrix. It is, therefore, inevitable that the blind constraint on sources be reduced and ultimately rely on some a priori knowledge of sources. The SCSS techniques that use a priori knowledge of speakers to separate the mixed speech can be grouped into two classes: time domain and frequency domain.

In time domain SCSS techniques [14–18], each source is decomposed into independent basis functions in the training phase. The basis functions of each source are learned from a training data set generally based on ICA approaches. Then the trained basis functions along with the constraint imposed by linearity are used to estimate the individual speech signals via a maximum likelihood optimization. While these SCSS techniques perform well when the speech signal is mixed with other sounds, such as music, when the mixture consists of two speech signals, separability reduces significantly since the learnt basis functions of two speakers have a high degree of overlap. In frequency domain techniques [19–23], first a statistical model is fitted to the spectral vectors of each speaker. Then, the two speaker models are combined to model the mixed signal. Finally, in the test phase, underlying speech signals are estimated based on some criteria (e.g., minimum mean square error, likelihood ratio).

The other mainstream techniques for SCSS are CASA-based approaches which exploit psychoacoustic clues for separation [5–13]. In CASA methods, after an appropriate transform (such as the short-time Fourier transform (STFT) [9] or the gammatone filter bank [29]), the mixed signal is segmented into time-frequency cells; then based on some criteria, namely, fundamental frequency, onset, offset, position, and continuity, the cells that are believed to belong to one source are grouped. CASA models suffer from two main problems. First, the current methods are unable to separate unvoiced speech and second, the formant information is not included in the discriminative cues for separation.

Besides the above techniques, there have been other attempts that are categorized as neither BSS nor CASA. In [26], a work was presented based on neural networks and an extension of the Kalman filter. In [27, 28], a generalized Wiener filter and an autoregressive model have been applied for general signal separation, respectively. Though the techniques have a mathematical depth that is worth further exploration, no comprehensive results have been reported on the performance of these systems on speech signals.

Underdetermined BSS methods are usually designed without considering the characteristics of the speech signal. Speech signals can be modeled as an excitation signal filtered by a vocal-tract-related filter. In this paper, we develop a technique that extracts the excitation signals based on a CASA model and estimates the vocal-tract-related filters based on a probabilistic approach from the mixed speech. The model, in fact, adds vocal-tract-related filter characteristics as a new cue along with harmonicity cues. There have been a number of powerful techniques for extracting the fundamental frequencies of underlying speakers from the mixed speech [30–35]. Therefore, we focus on estimating the vocal-tract-related filters of the underlying signals based on maximum likelihood (ML) optimization. For this purpose, we first express the probability density function (PDF) of the mixed signal's log spectral vectors in terms of the PDFs of the underlying signal's vocal-tract-related filters. Then the mean vectors of the PDFs for the vocal-tract-related filters are estimated in a maximum likelihood framework. Finally, the estimated mean vectors along with the extracted fundamental frequencies are used to reconstruct the underlying speech signals. We compare our model with a frequency domain method and a CASA approach. Experimental results, conducted on ten different speakers, show that our model outperforms the two individual approaches in terms of signal-to-noise ratio (SNR) and the percentage of crosstalk suppression.

The remainder of this paper is organized as follows. In Section 2, we start with a preliminary study of the basic concepts of underdetermined BSS and CASA models. The discussions in that section manifest the pros and cons of these techniques and the basic motivations for the proposed method. In Section 3, we review the model and present the overall functionality of the proposed model. The source-filter modeling of speech signals is discussed in Section 4. Harmonicity detection is discussed in Section 5 where we extract the fundamental frequencies of the underlying speech signals from the mixture. In Section 6, we show how to obtain the statistical distributions of vocal-tract-related filters in the training phase. This procedure is performed by fitting a mixture of Gaussian densities to the space feature. Estimating the PDF of the log spectral vector for the mixed speech in terms of the PDFs of the underlying signal vocal-tract-related filters as well as the resulting ML estimator is given in Section 7 with related mathematical definitions. Experimental results are reported in Section 8 and, finally, conclusions are discussed in Section 9.

2. PRELIMINARY STUDY

2.1. Underdetermined BSS

In the BSS context, the separation of I source speech signals when we have access to J observation signals can be formulated as

$$\mathbf{Y}^t = \mathbf{A}\mathbf{X}^t, \tag{1}$$

where $\mathbf{Y}^t = [\mathbf{y}_1^t, \dots, \mathbf{y}_j^t, \dots, \mathbf{y}_J^t]^T$ and $\mathbf{X}^t = [\mathbf{x}_1^t, \dots, \mathbf{x}_i^t, \dots, \mathbf{x}_I^t]^T$ and $\mathbf{A} = [a_{j,i}]_{J\times I}$ is a $(J \times I)$ instantaneous mixing matrix which shows the relative position of the sources from the observations. Also, vectors $\mathbf{y}_j^t = \{y_j^t(n)\}_{n=1}^N$ and $\mathbf{x}_j^t = \{x_j^t(n)\}_{n=1}^N$, for $j = 1, 2, \dots, J$ and $i = 1, 2, \dots, I$, represent N-dimensional vectors of the jth observation and ith source signals, respectively.[1] 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 to or greater than the number of sources $(J \geq I)$, the solution to the separation problem is simply obtained by estimating the inverse of the mixing matrix, that is, $\mathbf{W} = \mathbf{A}^{-1}$, and left multiplying both sides of (1) by $\mathbf{W}$. Many solutions have so far been proposed for determining the mixing matrix and quite satisfactory results have been reported [1–4].

However, when the number of observations is less than the number of sources $(J < I)$, the mixing matrix is noninvertible 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 *underdetermined* BSS and has recently become a hot topic in the signal processing realm.

In this paper, we deal with underdetermined BSS in which we assume $J = 1$ and $I = 2$, that is,

$$\mathbf{y}^t = \mathbf{x}_1^t + \mathbf{x}_2^t, \tag{2}$$

where without loss of generality we assume that the elements of the mixing matrix $(\mathbf{A} = [a_{11} \;\; a_{12}])$ are included in the source signals as they do not provide us with useful information for the separation process. Generally for underdetermined BSS, a priori knowledge of source signals is used in the

[1] It should be noted that throughout the paper the time domain vectors are obtained by applying a smoothing window (e.g., Hamming window) of length N on the source and observation signals.

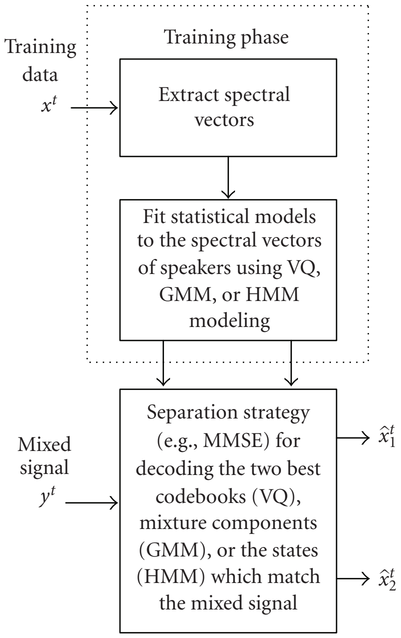

Training phase
Training data x^t
Extract spectral vectors
Fit statistical models to the spectral vectors of speakers using VQ, GMM, or HMM modeling
Mixed signal y^t
Separation strategy (e.g., MMSE) for decoding the two best codebooks (VQ), mixture components (GMM), or the states (HMM) which match the mixed signal
$\hat{x}_1^t$
$\hat{x}_2^t$

FIGURE 1: A schematic of underdetermined BSS techniques.

form of the statistical models of the sources. Figure 1 shows a general schematic for underdetermined BSS techniques in the frequency domain. The process consists of two phases: the training phase and test phase. In the training phase, the feature space of each speaker is modeled using common statistical modeling techniques (e.g., VQ, GMM, and HMM). Then in the test phase, we decode the codevector (when VQ is used), the mixture component (when GMM is used), or the state (when HMM is used) of the two models that when mixed satisfy a minimum distortion criterion compared to the observed mixed signal's feature vector. In these models, three components play important roles in the system's performance:

(i) selected feature,
(ii) statistical model,
(iii) separation strategy.

Among these components, the selected feature has a direct influence on the statistical model and the separation strategy used for separation. In previous works [19–23], log spectra (the log magnitude of the short-time Fourier transform) have been mainly used as the selected feature. In [36], we have shown that the log spectrum exhibits poor performance when the separation system is used in a speaker-independent scenario (i.e., training is not on speakers in the mixed signal). This drawback of the selected feature limits remarkably the usefulness of underdetermined BSS techniques in practical situations. In Section 3, we propose an approach to mitigate this drawback for the speaker-independent case.

Before we elaborate on the proposed approach in the next subsection, we review the fundamental concepts of computational auditory scene analysis technique which is a component of the proposed technique.

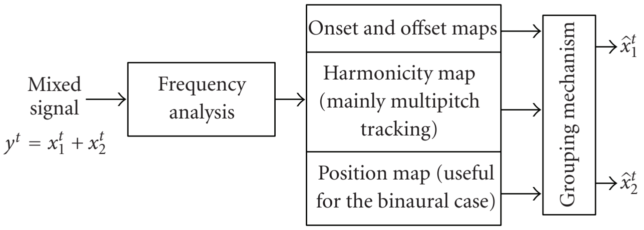


FIGURE 2: Basic operations in CASA models.

2.2. Computational auditory scene analysis

The human auditory system is able to pick out one conversation from among dozens in a crowded room. This is a capability that no artificial system can currently match. Many efforts have been carried out to mimic this fantastic ability. There are rich literatures [5–13] on how the human auditory system solves an auditory scene analysis (ASA). However, less work has been done to implement this knowledge using advanced machine learning approaches. Figure 2 shows a block diagram of the performed operations which attempt to replicate the human auditory system when it receives the sounds from different sources. These procedures were first dubbed by Bregman [5] as *computational auditory scene analysis*. In the first stage, the mixture sound is segmented into the time-frequency cells. Segmentation is performed using either the short-time Fourier transform (STFT) [9] or the gammatone filter bank [29]. The segments are then grouped based on cues which are mainly onset, offset, harmonicity, and position cues [11]. The position cue is a criterion which differs between two sounds received from different directions and distances. Therefore, this discriminative feature is not useful for the SCSS problem where the speakers are assumed to speak from the same position. Starts and ends of vowel and plosive sounds are among the other cues which can be applied for grouping purposes [6]. However, no comprehensive approach has been proposed to take into account the onset and offset cues except a recently proposed approach in [37].

Perhaps the most important cue for grouping the time-frequency segments is the harmonicity cue [38]. Voiced speech signals have a periodic nature which can be used as a discriminative feature when speech signals with different periods are mixed. Thus, the primary goal is to develop algorithms by which we extract the fundamental frequency of the underlying signals. This topic is commonly referred to as *multipitch tracking* and a wide variety of techniques has so far been proposed [29–33, 39–46]. After determining the fundamental frequencies of the underlying signals, the time-frequency cells which lie within the extracted fundamental frequencies or their harmonics are grouped into two speech signals.

The techniques based on CASA suffer from two problems. First, these techniques are not able to separate unvoiced segments and almost in all reported results one or both underlying signals are fully voiced [13, 47]. Second, the vocal-tract-related filter characteristics are not included in the discriminative cues for separation. In other words, in CASA techniques the role of the excitation signal is more important than the vocal tract shape. In the next section, we propose an approach to include the vocal tract shapes of the underlying signals as a discriminative feature along with the harmonicity cues.

3. MODEL OVERVIEW

In the previous section, we reviewed the two different approaches for the separation of two speech signals received from one microphone. In this section, we propose a new technique which can be viewed as the integration of underdetermined BSS with a limited form of CASA.

As shown in Figure 3, the technique can be regarded as a new CASA system in which the vocal-tract-related filter characteristics, which are obtained during a training phase, are included into a CASA model. Introducing the new cue (vocal-tract-related filter characteristics) into the system necessitates a new grouping procedure in which both vocal-tract-related filter and fundamental frequency information should be used for separation, a task which is accomplished using methods from underdetermined BSS techniques.

Figure 4 shows the proposed algorithm in detail. The whole process can be described in the following stages.

(1) Training phase:

- (i) from a large training data set consisting of a wide variety of speakers extract the log spectral envelop vectors (vocal-tract-related filter) based on the method described in [48],
- (ii) fit a Gaussian mixture model (GMM) to the obtained log spectral envelop vectors.

(2) Test phase:

- (i) extract the fundamental frequencies of the underlying signals from the mixture signal using the method described in Section 5,
- (ii) generate the excitation signals using the method described in Appendix A,
- (iii) add the two obtained log excitation vectors to the mean vectors of the Gaussian mixture,
- (iv) decode the two Gaussian mixture's mean vectors which satisfy the maximum likelihood criterion (23) described in Section 7,
- (v) recover the underlying signals using the decoded mean vectors, excitation signals, and the phase of the mixed signal.

This architecture has several distinctive attributes. From the CASA model standpoint, we add a new important cue into the system. In this way, we apply the vocal tract information to separate the speech sources as opposed to current CASA models which use vocal cord information to separate the sounds. As an underdetermined BSS technique, the approach can separate the speech signal even if it comes from unknown speakers. In other words, the system is speaker-independent in contrast with current underdetermined blind source separation techniques that use a priori knowledge of the speakers. This attribute results from separating the vocal-tract-related filter from the excitation signal, which is a speaker-dependent characteristic of the speech signal. It should be noted that from the training data set we obtained one speaker-independent Gaussian mixture model which is then used for both speakers as opposed to approaches that require training data for each of the speakers.

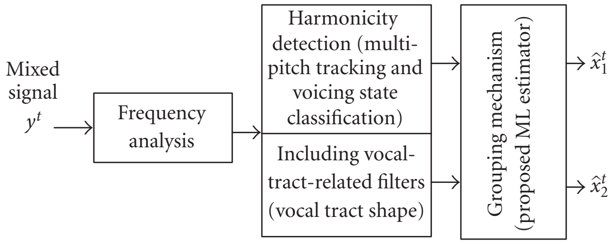

Mixed signal
y^t
Frequency analysis
Harmonicity detection (multi-pitch tracking and voicing state classification)
Including vocal-tract-related filters (vocal tract shape)
Grouping mechanism (proposed ML estimator)
$\hat{x}_1^t$
$\hat{x}_2^t$

FIGURE 3: A new CASA model (proposed model) which includes the vocal-tract-related filters along with harmonicity cues for separation.

In the following sections, we first present the concept of source-filter modeling which is the basic framework built on for the proposed method. Then the components of the proposed technique are described in more details. In the remaining sections these components are training phase, multipitch detection, and maximum likelihood estimation in which we formulate the proposed approach. In particular, we follow the procedure for obtaining the maximum likelihood estimator by which we are able to estimate the vocal-tract-related filters of the underlying signals from the mixture signal.

4. SOURCE-FILTER MODELING OF SPEECH SIGNALS

In the process of speech production, an excitation signal produced by the vocal cord is shaped by the vocal tract. From the signal processing standpoint, the process can be implemented using a convolution operation between the vocal-cord-related signal and the vocal-tract-related filter. Thus, for our case, we have

$$\mathbf{x}_i^t = \mathbf{e}_i^t * \mathbf{h}_i^t, \quad i \in \{1,2\}, \tag{3}$$

where $\mathbf{e}_i^t = \{e_i^t(n)\}_{n=1}^N$ and $\mathbf{h}_i^t = \{h_i^t(n)\}_{n=1}^N$, respectively, represent the excitation signal and vocal-tract-related filter of the ith speaker computed within the analysis window of length N. Also, $*$ denotes the convolution operation. Accordingly, in the frequency domain we have

$$\mathbf{x}_i^f = \mathbf{e}_i^f \times \mathbf{h}_i^f, \tag{4}$$

where $\mathbf{x}_i^f = \{x_i^f(d)\}_{d=1}^D$, $\mathbf{e}_i^f = \{e_i^f(d)\}_{d=1}^D$, and $\mathbf{h}_i^f = \{h_i^f(d)\}_{d=1}^D$ represent the D-point discrete Fourier transform (DFT) of $\mathbf{x}_i^t$, $\mathbf{e}_i^t$, and $\mathbf{h}_i^t$, respectively. The superscript f indicates that the signal is in the frequency domain. In this paper, the main analysis is performed in the log frequency domain. Thus transferring the DFT vectors to the log frequency domain gives

$$\mathbf{x}_i = \mathbf{e}_i + \mathbf{h}_i, \tag{5}$$

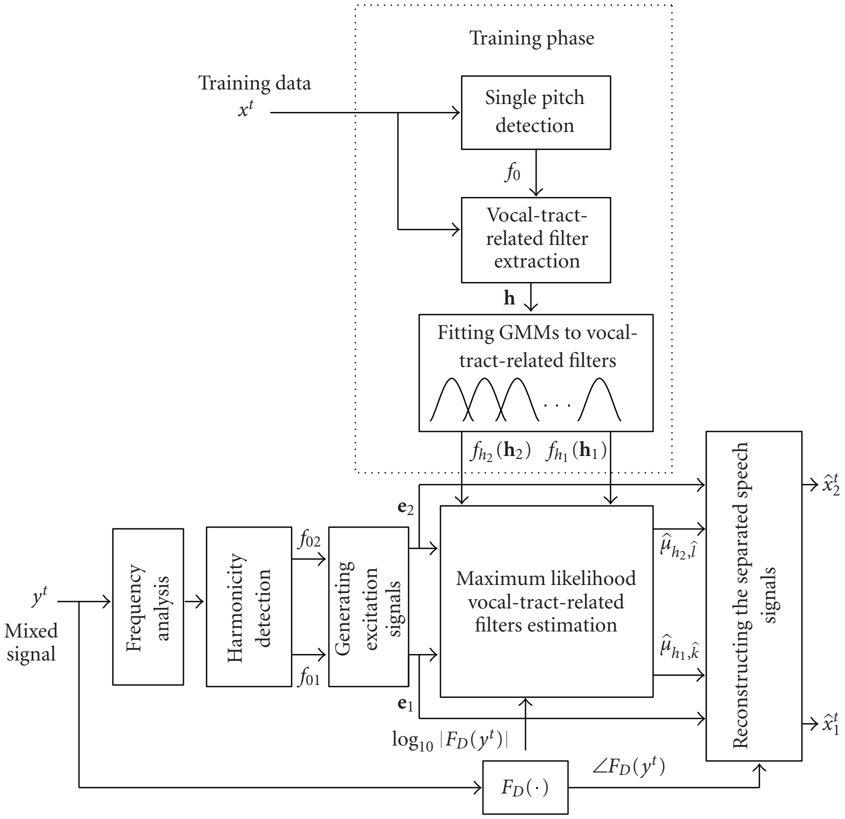

Training phase
Training data
x^t
Single pitch detection
f_0
Vocal-tract-related filter extraction
$\mathbf{h}$
Fitting GMMs to vocal-tract-related filters
$f_{h_2}(\mathbf{h}_2)$ $f_{h_1}(\mathbf{h}_1)$
y^t
Mixed signal
Frequency analysis
Harmonicity detection
f_{02}
f_{01}
Generating excitation signals
$\mathbf{e}_2$
$\mathbf{e}_1$
Maximum likelihood vocal-tract-related filters estimation
$\hat{\mu}_{h_2,\hat{l}}$
$\hat{\mu}_{h_1,\hat{k}}$
$\log_{10}|F_D(y^t)|$
$F_D(\cdot)$
$\angle F_D(y^t)$
Reconstructing the separated speech signals
$\hat{x}_2^t$
$\hat{x}_1^t$

Figure 4: A block diagram of the proposed model.

Table 1: Definition of signals which are frequently used.

Signal	Time	DFT	log |DFT|
Observation signal	$\mathbf{y}^t$	$\mathbf{y}^f$	$\mathbf{y}$
Source signal $i \in \{1, 2\}$	$\mathbf{x}_i^t$	$\mathbf{x}_i^f$	$\mathbf{x}_i$
Vocal-tract-related filter	$\mathbf{h}_i^t$	$\mathbf{h}_i^f$	$\mathbf{h}_i$
Excitation signal	$\mathbf{e}_i^t$	$\mathbf{e}_i^f$	$\mathbf{e}_i$
Estimated source signal	$\hat{\mathbf{x}}_i^t$	$\hat{\mathbf{x}}_i^f$	$\hat{\mathbf{x}}_i$

where $\mathbf{x}_i = \log_{10}|\mathbf{x}_i^f| = \{x_i(d)\}_{d=1}^D$, $\mathbf{h}_i = \log_{10}|\mathbf{h}_i^f| = \{h_i(d)\}_{d=1}^D$, and $\mathbf{e}_i = \log_{10}|\mathbf{e}_i^f| = \{e_i(d)\}_{d=1}^D$ denote the log spectral vectors corresponding to $\mathbf{x}_i^f$, $\mathbf{e}_i^f$, and $\mathbf{h}_i^f$, respectively and $|\cdot|$ is the magnitude operation. Since these signals are frequently used hereafter, we present definitions and the symbols representing these signals in Table 1.

Harmonic modeling [48] and linear predictive coding [49] are frequently used to decompose the speech signal into the excitation signal and vocal-tract-related filter. In harmonic modeling (the approach we use in this paper), the envelope of the log spectrum represents the vocal-tract-related filter, that is, $\mathbf{h}_i$. In addition, a windowed impulse train is used to represent the excitation signal. For voiced frames, the period of the impulse train is set to the extracted fundamental frequency while for the unvoiced frames the period of the impulse train is set to 100 Hz [48] (see Appendix A for more details).

We use (5) in Section 7 to derive the maximum likelihood estimator in which the PDF of $\mathbf{y}$ is expressed in terms of the PDFs of the $\mathbf{h}_i$s'. Therefore, it is necessary to obtain $\mathbf{e}_i$ and the PDF of $\mathbf{h}_i$. The excitation signal $\mathbf{e}_i$ is constructed using voicing state and the fundamental frequencies of the underlying speech signals which are determined using the multipitch detection algorithm described in the next section. The PDF of $\mathbf{h}_i$ is also obtained in the training phase as described in Section 6.

5. MULTIPITCH DETECTION

The task of the multipitch detection stage is to extract the fundamental frequencies of the underlying signals from the mixture. Different methods have been proposed for this task [30–35, 39–43] which are mainly based on either the normalized cross-correlation [50] or comb filtering [51]. In order to improve the robustness of the detection stage, some algorithms include preprocessing techniques based on principles of the human auditory perception system [29, 52, 53]. In these algorithms, after passing the mixed signal through a bank of filters, the filter's outputs (for low-frequency channels) and the envelop of the filter's output (for high-frequency channels) are fed to the periodicity detection stage [31, 33].

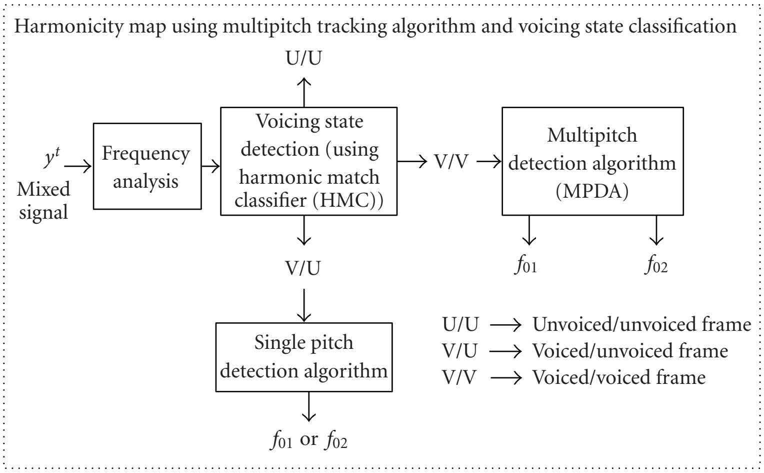

Harmonicity map using multipitch tracking algorithm and voicing state classification
U/U
y^t
Mixed signal
Frequency analysis
Voicing state detection (using harmonic match classifier (HMC))
V/V
Multipitch detection algorithm (MPDA)
f_{01}
f_{02}
V/U
Single pitch detection algorithm
f_{01} or f_{02}
U/U ⟶ Unvoiced/unvoiced frame
V/U ⟶ Voiced/unvoiced frame
V/V ⟶ Voiced/voiced frame

FIGURE 5: The modified multipitch detection algorithm in which a voicing classifier is added to detect the fundamental frequencies in general case.

The comb filter-based periodicity detection algorithms estimate the underlying fundamental frequencies in two stages [30, 35, 41, 42]. At the first stage, the fundamental frequency of one of the underlying signals is determined using a comb filter. Then the harmonics of the measured fundamental frequencies is suppressed in the mixed signal and the residual signal is again fed to the comb filter to determine the fundamental frequency of the second speaker. Chazan et al. [30] proposed an iterative multipitch estimation approach using a nonlinear comb filter. Their technique applies a nonlinear comb filter to capture all quasiperiodic harmonics in the speech bandwidth such that their method led to better results than previously proposed techniques in which a comb filter is used. In this paper, we use the method proposed by Chazan et al. [30] for the multipitch detection stage.

One shortcoming of multipitch detection algorithms is that they have been designed for the case in which one or both concurrent sounds are fully voiced. However, speech signals are generally categorized into voiced (V) or unvoiced (U) segments.[2] Consequently, the mixed speech with two speakers contains U/U, V/U, and V/V segments. This means that in order to have a reliable multipitch detection algorithm, we should first determine the voicing state of the mixed signal's analysis segment. In order to generalize Chazan's multipitch detection system, we augment a voicing state classifier to the multipitch detection system. By doing this, we first determine the state of the underlying signals, then either multipitch detection (when state is V/V) or single pitch detection (when state is V/U) or no action is applied on the mixed signal's analysis segment. Figure 5 shows a schematic of the generalized multipitch detection algorithm.

Several voicing state classifiers have been proposed, namely, using the spectral autocorrelation peak valley ratio (SAPVR) criterion [54], nonlinear speech processing [55], wavelet analysis [56], Bayesian classifiers [57], and harmonic matching classifier (HMC) [58]. In this paper, we use the HMC technique [58] for voicing classification. In this way, we obtain a generalized multipitch tracking algorithm. In a separate study [59], we evaluated the performance of this generalized multipitch tracking using a wide variety of mixed signals. On average, this technique is able to detect the fundamental frequencies of the underlying signals in the mixture with gross error rate equal to 18%. In particular, we noticed that most errors occurred when the fundamental frequencies of the underlying signals are within the range $f_{0_1} = f_{0_2} \pm 15\,\text{Hz}$. It should be noted that tracking fundamental frequencies in the mixed signal when they are close is still a challenging problem [31].

6. TRAINING PHASE

In the training phase, we model the spectral envelop vectors ($\mathbf{h}_i$) using a mixture of Gaussian probability distributions known as Gaussian mixture model (GMM). We first extract the spectral envelop vectors from a large training database. The database contains speech files from both genders with different ages. The procedure for extracting the spectral envelop vectors is similar to that described in [48] (see Section 8.1 for more details). As mentioned earlier, we use a training database which contains the speech signal of different speakers so that we can generalize the algorithm. This approach means that we use one statistical model for the two speakers' log spectral envelop vectors. We, however, use different notations for the two speakers' log spectral envelop vectors in order to not confuse them. In the following, we model the PDF of the log spectral vectors of the vocal-tract-related filter for the ith speaker by a mixture of K_i Gaussian densities in the following form:

$$f_{\mathbf{h}_i}(\mathbf{h}_i) \triangleq \sum_{k=1}^{K_i} c_{h_i,k}\mathcal{N}(\mathbf{h}_i, \mu_{h_i,k}, \mathbf{U}_{h_i,k}), \quad i \in \{1,2\}, \qquad (6)$$

[2] Generally, it is also desired to detect the silence segment, but in this paper we consider the silence segments as a special case of unvoiced segments.

here $c_{h_i,k}$ represents the a priori probability for the kth Gaussian in the mixture and satisfies $\sum_k c_{h_i,k} = 1$ and

$$\mathcal{N}(\mathbf{h}_i, \mu_{h_i}, \mathbf{U}_{h_i}) \triangleq \frac{\exp\left(-(1/2)(\mathbf{h}_i-\mu_{h_i})^T \mathbf{U}_{h_i}^{-1}(\mathbf{h}_i-\mu_{h_i})\right)}{\sqrt{(2\pi)^D\,|\mathbf{U}_{h_i}|}} \tag{7}$$

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

$$f_{\mathbf{h}_i}(\mathbf{h}_i) \triangleq \sum_{k=1}^{K_i} c_{h_i,k} \prod_{d=1}^{D} \frac{\exp\left(-(1/2)\left((h_i(d)-\mu_{h_i,k}(d))/\sigma_{h_i,k}(d)\right)^2\right)}{\sqrt{2\pi}\,\sigma_{h_i,k}(d)}, \tag{8}$$

where $h_i(d)$, $\mu_{h_i,k}(d)$, and $\sigma^2_{h_i,k}(d)$ are the dth component of $\mathbf{h}_i$, dth component of the mean vector, and the dth element on the diagonal of the covariance matrix $\mathbf{U}_{h_i,k}$, respectively.

In this way, we have the statistical distributions of the vocal-tract-related filters as a priori knowledge. These distributions are then used in the ML estimator.

7. MAXIMUM LIKELIHOOD ESTIMATOR

After fitting a statistical model to the log spectral envelop vectors and generating the excitation signals using obtained fundamental frequencies in the multipitch tracking stage, we are now ready to estimate the vocal-tract-related filters of the underlying signals. In this section, we first express the PDF of the mixed signal's log spectral vectors in terms of the PDFs of the log spectral vectors for the vocal-tract-related filters of the underlying signals. We then obtain an estimate of the underlying signals' vocal-tract-related filters using the obtained PDF in a maximum likelihood framework. In Table 2, notations and definitions which are frequently used in this section are summarized.

To begin, we should first obtain a relation between the log spectral vector of the mixed signal and those of the underlying signals. From the *mixture-maximization* approximation [60], we know

$$\begin{aligned}\mathbf{y} &\approx \operatorname{Max}(\mathbf{x}_1,\mathbf{x}_2)\\ &= \left[\max\left(x_1(1),x_2(1)\right),\dots,\max\left(x_1(d),x_2(d)\right),\dots,\right.\\ &\qquad \left.\max\left(x_1(D),x_2(D)\right)\right]^T,\end{aligned} \tag{9}$$

where $\mathbf{y} = \log_{10}|\mathbf{y}^f|$, $\mathbf{x}_1 = \log_{10}|\mathbf{x}_1^f|$, and $\mathbf{x}_2 = \log_{10}|\mathbf{x}_2^f|$, and $\max(\cdot,\cdot)$ returns the larger element. Equation (9) implies that the log spectrum of the mixed signal is almost exactly the elementwise maximum of the log spectrum of the two underlying signals.

Table 2: Symbols with definitions.

Symbol	Description
I	number of sources $i \in \{1,2\}$
D	order of DFT
K	number of Gaussian mixtures
$f_{\mathbf{s}}(\mathbf{s})$	PDF of signal $\mathbf{s} \in \{\mathbf{x}_i, \mathbf{h}_i, \text{or } \mathbf{y}\}$
$F_{\mathbf{s}}(\mathbf{s})$	CDF of signal $\mathbf{s} \in \{\mathbf{x}_i, \mathbf{h}_i, \text{or } \mathbf{y}\}$
$\mu_{h_i,k}$	kth Gaussian mixture's mean vector
$\sigma^2_{h_i,k}$	kth Gaussian mixture's variance vector

To begin, we first express the PDF of $\mathbf{x}_i$ in terms of the PDF of $\mathbf{h}_i$ given $\mathbf{e}_i$. Clearly,

$$f_{\mathbf{x}_i}(\mathbf{x}_i) = f_{\mathbf{h}_i}(\mathbf{x}_i - \mathbf{e}_i), \quad i \in \{1,2\}, \tag{10}$$

which is the result of (5) and the assumption that $\mathbf{e}_i$ is a deterministic signal (we obtained $\mathbf{e}_i$ through multipitch detection and through generating the excitation signals). Thus the PDF of $\mathbf{x}_i$, for $i \in \{1,2\}$, is identical to the PDF of $\mathbf{h}_i$ except with a shift in the mean vector equal to $\mathbf{e}_i$. The relation between the cumulative distribution function (CDF) of $\mathbf{x}_i$ and those of $\mathbf{h}_i$ is also related in a way similar to (10), that is,

$$F_{\mathbf{x}_i}(\mathbf{x}_i) = F_{\mathbf{h}_i}(\mathbf{x}_i - \mathbf{e}_i), \tag{11}$$

where

$$F_{\mathbf{h}_i}(\sigma) = \int_{-\infty}^{\sigma} f_{\mathbf{h}_i}(\xi)\,d\xi, \quad i \in \{1,2\}, \tag{12}$$

in which σ is an arbitrary vector.

From (9), the cumulative distribution function (CDF) of the mixed log spectral vectors $F_{\mathbf{y}}(\mathbf{y})$ is given by

$$F_{\mathbf{y}}(\mathbf{y}) = F_{\mathbf{x}_1\mathbf{x}_2}(\mathbf{y},\mathbf{y}), \tag{13}$$

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

$$F_{\mathbf{y}}(\mathbf{y}) = F_{\mathbf{x}_1}(\mathbf{y}) \times F_{\mathbf{x}_2}(\mathbf{y}). \tag{14}$$

Thus $f_{\mathbf{y}}(\mathbf{y})$ is obtained by differentiating both sides of (14) to give

$$f_{\mathbf{y}}(\mathbf{y}) = f_{\mathbf{x}_1}(\mathbf{y}) \cdot F_{\mathbf{x}_2}(\mathbf{y}) + f_{\mathbf{x}_2}(\mathbf{y}) \cdot F_{\mathbf{x}_1}(\mathbf{y}). \tag{15}$$

Using (10) and (11) it follows that

$$\begin{aligned}f_{\mathbf{y}}(\mathbf{y}) &= f_{\mathbf{h}_1}(\mathbf{y}-\mathbf{e}_1)\\ &\quad \cdot F_{\mathbf{h}_2}(\mathbf{y}-\mathbf{e}_2) + f_{\mathbf{h}_2}(\mathbf{y}-\mathbf{e}_2) \cdot F_{\mathbf{h}_1}(\mathbf{y}-\mathbf{e}_1).\end{aligned} \tag{16}$$

The CDF to express $F_{\mathbf{h}_i}(\mathbf{y}-\mathbf{e}_i)$ is obtained by substituting $f_{\mathbf{h}_i}(\mathbf{h}_i)$ from (8) into (12) to give

$$\begin{aligned} F_{\mathbf{h}_i}(\mathbf{y}-\mathbf{e}_i) &\\ = \int_{-\infty}^{\mathbf{y}-\mathbf{e}_i} f_{\mathbf{h}_i}(\xi)d\xi &= \int_{-\infty}^{y(d)-e_i(d)} \sum_{k=1}^{K_i} c_{h_i,k} \prod_{d=1}^{D} \\ &\times \left[\frac{1}{\sigma_{h_i,k}(d)\sqrt{2\pi}} \times \exp\left(-\frac{1}{2}\left(\frac{\xi_d-\mu_{h_i,k}(d)}{\sigma_{h_i,k}(d)}\right)^2\right)\right]d\xi_d. \end{aligned} \tag{17}$$

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

$$\begin{aligned} F_{\mathbf{h}_i}(\mathbf{y}-\mathbf{e}_i) &\\ = \sum_{k=1}^{K_i} c_{h_i,k} \prod_{d=1}^{D} &\left[\frac{1}{\sigma_{h_i,k}(d)\sqrt{2\pi}}\right. \\ &\left.\times \int_{-\infty}^{y(d)-e_i(d)} \exp\left(-\frac{1}{2}\left(\frac{\xi_d-\mu_{h_i,k}(d)}{\sigma_{h_i,k}(d)}\right)^2\right)d\xi_d\right]. \end{aligned} \tag{18}$$

The term in the bracket in (18) is often expressed in terms of the *error function*

$$\operatorname{erf}(\alpha) = \frac{1}{\sqrt{2\pi}}\int_0^{\alpha} \exp\left(-\frac{1}{2}\nu^2\right)d\nu. \tag{19}$$

Thus, we conclude that

$$F_{\mathbf{h}_i}(\mathbf{y}-\mathbf{e}_i) = \sum_{k=1}^{K_i} c_{h_i,k} \prod_{d=1}^{D}\left[\operatorname{erf}\left(z_{h_i,k}(d)\right)+\frac{1}{2}\right], \tag{20}$$

where

$$z_{h_i,k}(d) = \frac{y(d)-e_i(d)-\mu_{h_i,k}(d)}{\sigma_{h_i,k}(d)}, \quad i \in \{1,2\}. \tag{21}$$

Finally, we obtain the PDF of the log spectral vectors of the mixed signal by substituting (10) and (20) into (16) to give

$$\begin{aligned} f_{\mathbf{y}}(\mathbf{y}) = \sum_{k=1}^{K_1}\sum_{l=1}^{K_2} c_{h_1,k}c_{h_2,l} &\\ \times \Bigg(\prod_{d=1}^{D}\bigg[&\left(2\pi\sigma_{h_1,k}^2(d)\right)^{-1/2} \times \left(\operatorname{erf}\left(z_{h_2,l}(d)\right)+\frac{1}{2}\right) \\ &\times \exp\left(-\frac{1}{2}z_{h_1,l}^2(d)\right)\bigg] \\ &+\prod_{d=1}^{D}\bigg[\left(2\pi\sigma_{h_2,l}^2(d)\right)^{-1/2} \times \left(\operatorname{erf}\left(z_{h_1,k}(d)\right)+\frac{1}{2}\right) \\ &\times \exp\left(-\frac{1}{2}z_{h_2,l}^2(d)\right)\bigg]\Bigg). \end{aligned} \tag{22}$$

Equation (22) gives the PDF of log spectral vectors for the mixed signal in terms of the mean and variance of the log spectral vectors for the vocal-tract-related filters of the underlying signals.

Now we apply $f_{\mathbf{y}}(\mathbf{y})$ 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 kth Gaussian in $f_{\mathbf{h}_1}(\mathbf{h}_1;\lambda_{h_1})$ and the lth Gaussian in $f_{\mathbf{h}_2}(\mathbf{h}_2;\lambda_{h_2})$ such that $f_{\mathbf{y}}(\mathbf{y})$ is maximized. The estimator is given by

$$\{\hat{k},\hat{l}\}_{\mathrm{ML}} = \arg\max_{\theta_{k,l}} f_{\mathbf{y}}(\mathbf{y} \mid \theta_{k,l}), \tag{23}$$

where

$$\theta_{k,l} = \{\mu_{h_1,k},\mu_{h_2,l},\sigma_{h_1,k},\sigma_{h_2,l}\}. \tag{24}$$

The estimated mean vectors are then used to reconstruct the log spectral vectors of the underlying signals. Using (5), we have

$$\begin{aligned} \hat{\mathbf{x}}_1 &= \hat{\mu}_{h_1,\hat{k}} + \mathbf{e}_1, \\ \hat{\mathbf{x}}_2 &= \hat{\mu}_{h_2,\hat{l}} + \mathbf{e}_2, \end{aligned} \tag{25}$$

where $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ are the estimated log spectral vectors for speaker one and speaker two, respectively. Finally, the estimated signals are obtained in the time domain by

$$\hat{\mathbf{x}}_i^t = \mathcal{F}_{\mathcal{D}}{}^{-1}\left(10^{\hat{\mathbf{x}}_i}\cdot\varphi_{\mathbf{y}}\right), \quad i \in \{1,2\}, \tag{26}$$

where $\mathcal{F}_{\mathcal{D}}{}^{-1}$ denotes the inverse Fourier transform and $\varphi_{\mathbf{y}}$ is the phase of the Fourier transform of the windowed mixed signal, that is, $\varphi_{\mathbf{y}} = \angle\mathbf{y}^f$. In this way, we obtain an estimate of $\mathbf{x}_i^t$ in a maximum likelihood sense. It should be noted that it is common to use the phase of the STFT of the mixed signal for reconstructing the individual signals [13, 19–21] 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, that is, > 1 second [61]. To the best of our knowledge no technique has been proposed to extract the individual phase values from the mixed phase. In the following section we evaluate the performance of the estimator by conducting experiments on mixed signals.

8. EXPERIMENTAL RESULTS AND COMPARISONS

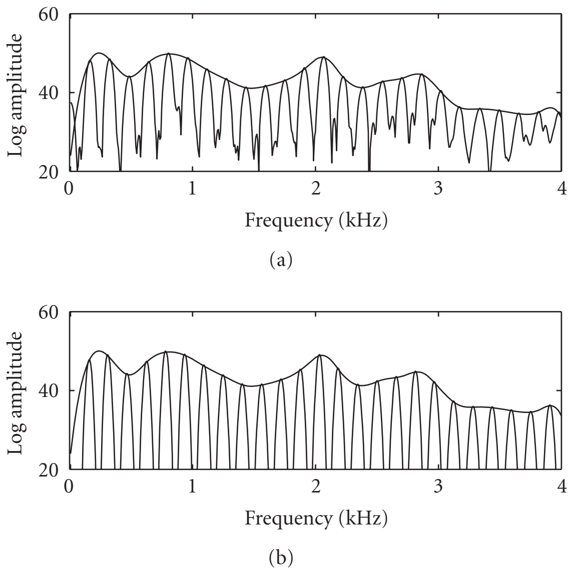
In order to evaluate the performance of our proposed technique, we conducted the following experiments. We first explain the procedure for extracting vocal-tract-related filters in the training phase; then we describe three different separation models with which we compare our model. The techniques are the ideal binary mask, MAXVQ model, and harmonic magnitude suppression (HMS). The ideal binary mask model (see Appendix B for more details) is an upper bound for SCSS systems. Comparing our results with the ideal case shows the gap between the proposed system and an ideal case. The HMS method, which is categorized as a

CASA model, uses the harmonicity cues for separation. In this way, we compare our model with a model which uses one cue (harmonicity cue) instead of our model which uses harmonicity as well as vocal-tract-related filters for separation. The MAXVQ separation technique is an underdetermined BSS method which uses the quantized log spectral vectors as a priori knowledge to separate the speech signal. Thus, we compare our model with both a CASA model and an underdetermined BSS technique. After introducing the feature extraction procedure and models, the results in terms of the obtained SNR and the percentage of crosstalk suppression are reported.

8.1. Feature extraction

We used one hour of speech signals from fifteen speakers. Five speakers among the fifteen speakers were used for the training phase and the remaining ten speakers were used for the testing phase. Throughout all experiments, a Hamming window with a duration of 32 milliseconds and a frame rate of 10 milliseconds was used for short-time processing of the data. The segments are transformed into the frequency domain using a 1024-point discrete Fourier transform ($D = 1024$), resulting in spectral vectors of dimension 512 (symmetry was discarded).

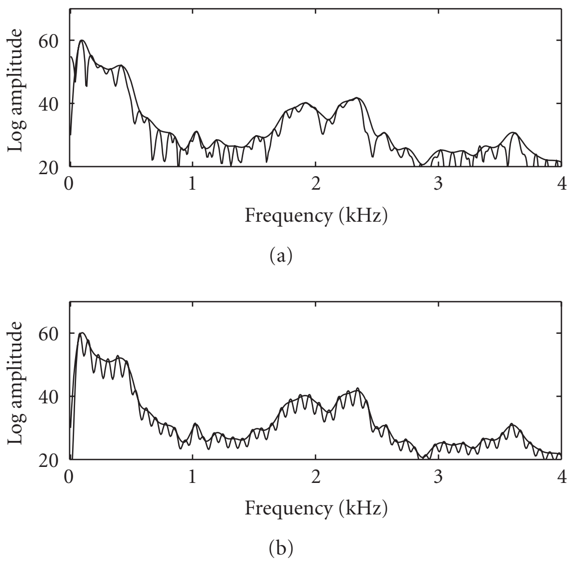
In the training phase, we must extract the log spectral vectors of the vocal-tract-related filters (envelop spectra) of the speech segments. The envelop spectra are obtained by a method proposed by Paul [62] and further developed by McAulay and Quatieri [48]. In this method, first all peaks in a given spectrum vector are marked, and then the peaks whose occurrences are close to the fundamental frequencies and their harmonics are held and the remaining peaks are discarded. Finally, a curve is fitted to the selected peaks using cubic spline interpolation [63]. This process requires the fundamental frequency of the processed segment, so we use the pitch detection algorithm described in [64] to extract the pitch information. It should be noted that during the unvoiced segments no fundamental frequency exists, but as shown in [48], we can use an impulse train with fundamental frequency of 100 Hz as a reasonable approximation. This dense sampling of the unvoiced envelop spectra holds nearly all information contained in the unvoiced segments. As mentioned in Section 4, the spectrum vector $\mathbf{x}_i$ can be decomposed into the vocal-tract-related filter $\mathbf{h}_i$ and the excitation signal $\mathbf{e}_i$, two components by which our algorithm is developed. Figures 6 and 7 show an example of the original and synthesized spectra for a voiced segment and an unvoiced segment, respectively. In Figures 6(a) and 7(a) the original spectra and envelop are shown, while in Figures 6(b) and 7(b) the synthesized spectra are shown which are the results of multiplying the vocal-tract-related filter $\mathbf{h}_i$ by the excitation signal $\mathbf{e}_i$. In these figures, the extracted envelop vector $\mathbf{h}_i$ (vocal-tract-related filter) is superimposed on the corresponding spectrum $\mathbf{x}_i$. The resulting envelop vectors have a dimension of 512 which makes the training phase computationally intensive. As it was shown in [48], due to the smooth nature of the envelop vectors, the envelop vector can be downsampled by a factor of 8 to reduce the dimension to 64.



(a)

(b)

Figure 6: Analysis and synthesis of the spectrum for a voiced segment: (a) envelop superimposed on the original spectrum and (b) envelop superimposed on the synthesized spectrum.



(a)

(b)

Figure 7: Analysis and synthesis of the spectrum for an unvoiced segment: (a) envelop superimposed on the original spectrum and (b) envelop superimposed on the synthesized spectrum.

After extracting the envelop vectors, we fit a 1024-component mixture Gaussian density $f_{\mathbf{h}_i}(\mathbf{h}_i)$ to the training data set. Initial values for the mean vectors μ_i of the Gaussian mixtures are obtained using a 10-bit codebook [65, 66] and the components are assumed to have the same probability. As mentioned earlier, we compare our model with three methods. In the following, we present a short description of these models.

8.2. *Ideal binary mask*

An ideal model, known as the *ideal binary mask* [67], is used for the first comparison (see Appendix B for more details). This method is in fact an upper bound that an SCSS system can reach. Although the performance of current separation techniques is far from that of the ideal binary approach, including the ideal results in experiments reveals how much the current techniques must be improved to approach the desired model if $\mathbf{x}_1$ and $\mathbf{x}_2$ were known in an a priori fashion.

8.3. *MAXVQ technique*

We also compare our model with a technique known as MAXVQ [23] which is an SCSS technique based on the underdetermined BSS principle. The technique is spiritually similar to the ideal binary mask except that the actual spectra are replaced by an N-bit codebook (we use $N = 1024$) of the quantized spectrum vectors for modeling the feature space of each speaker. The objective is to find the codevectors that when mixed satisfy a minimum distortion criterion compared to the observed mixed speech's feature vector. MAXVQ is in fact a simplified version of the HMM-based speech separation techniques [19, 20] in which two parallel HMMs are used to decode the desired states of individual HMMs. In the MAXVQ model, the intraframe constraint imposed by HMM modeling is removed to reduce computational complexity. We chose this technique since it is similar to our model but with two major differences: first, no decomposition is performed such that spectrum vectors are directly used for separation, and second, the inferring strategy is different from our model in which the ML vocal-tract-related filter estimator is used.

8.4. *Harmonic magnitude suppression model*

Since in our model fundamental frequencies are used along with envelop vectors, we compare our model with a technique in which fundamental frequencies solely are used for separation. For this purpose, we use the so-called *harmonic magnitude suppression* [42, 68] technique. In the HMS method, two comb filters are constructed using the obtained fundamental frequencies obtained by using a multipitch detection tracking algorithm. The product of the mixed spectrum with the corresponding comb filter of each speaker is the output of the system. In this way we, in fact, suppress the peaks in log spectrum whose locations correspond to the fundamental frequency and all related harmonics to recover the separated signals. For extracting the fundamental frequencies of two speakers from the mixture, we use the multipitch tracking algorithm described in Section 5.

8.5. *Results*

For the testing phase, ten speech files are selected from the ten test speakers (one sentence from each speaker) and mixed in pairs to produce five mixed signals for the testing phase. We chose the speech files for the speakers independent and outside of the training data set to evaluate the independency of our model from speakers. The test utterances are mixed with aggregate signal-to-signal ratio adjusted to 0 dB.

Table 3: SNR results (dB).

Mix	Sep	(a)	(b)	(c)	(d)
$f_1 + m_6$[†]	f_1	11.54	6.80	3.56	4.16
$f_1 + m_6$	m_6	11.57	7.76	3.19	5.25
$f_2 + m_7$	f_2	11.61	6.75	2.33	3.53
$f_2 + m_7$	m_7	11.08	6.73	3.28	5.04
$f_3 + m_8$	f_3	11.21	6.98	2.86	5.39
$f_3 + m_8$	m_8	10.91	6.60	2.59	4.53
$f_4 + f_9$	f_4	8.86	5.60	3.33	3.81
$f_4 + f_9$	f_9	9.95	5.21	3.49	3.80
$m_5 + m_{10}$	m_5	10.17	5.61	2.38	3.75
$m_5 + m_{10}$	m_{10}	10.05	5.90	1.70	4.12
	Ave[‡]	10.7	6.40	2.88	4.33

[(a)] Ideal binary mask (upper bound for separation) [67].
[(b)] Proposed method.
[(c)] MAXVQ method [23].
[(d)] HMS [42, 68].
[†] f_i and m_j show speech signals of ith female and jth male speakers, respectively.
[‡] Averaged SNR over the ten speech files.

In order to quantify the degree of the separability, we chose two criteria: (i) the SNR between the separated and original signals in the time domain and (ii) the percentage of crosstalk suppression [13]. The SNR for the separated speech signal of the ith speaker is defined as

$$\mathrm{SNR}_i = 10 \cdot \log_{10}\left[\frac{\sum_n \left(x_i^t(n)\right)^2}{\sum_n \left(x_i^t(n) - \hat{x}_i^t(n)\right)^2}\right], \quad n = 1, 2, \ldots, \aleph, \tag{27}$$

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

The second criterion is the percentage of crosstalk suppression, P_i, which quantifies the degree of suppression of interference (crosstalk) in the separated signals (for more details see Appendix C).

The SNR and the percentage of crosstalk suppression are reported in Tables 3 and 4, respectively. The first column in each table represents the mixed speech file pairs, the second column represents the resulting separated speech file from the mixture. In Table 3, the SNR obtained using (a) the ideal binary mask approach, (b) our proposed method, (c) MAXVQ technique, and (d) HMS method is given in columns three to six, respectively. The last row shows the SNR averaged over the ten separated speech files. Analogous to Table 3, Table 4 instead shows the percentage of crosstalk suppression (P_i) for each separated speech file.

As the results in Tables 3 and 4 show, our model significantly outperforms the MAXVQ and HMS techniques both in terms of SNR and the percentage of crosstalk suppression. However, there is a significant gap between our model and the ideal binary mask case. On average, an improvement of 3.52 dB for SNR and an improvement of 28% in suppressing crosstalk are obtained using our method. The results

Table 4: Percentage of crosstalk suppression (%).

Mix	Sep	(a)	(b)	(c)	(d)
$f_1 + m_6^{\dagger}$	f_1	100	85.4	73.8	65.6
$f_1 + m_6$	m_6	100	85.5	75.8	68.2
$f_2 + m_7$	f_2	100	84.3	69.4	63.5
$f_2 + m_7$	m_7	100	89.7	73.1	66.5
$f_3 + m_8$	f_3	100	83.2	70.5	62.7
$f_3 + m_8$	m_8	100	81.4	72.6	61.4
$f_4 + f_9$	f_4	100	83.8	70.2	69.6
$f_4 + f_9$	f_9	100	87.2	67.6	65.3
$m_5 + m_{10}$	m_5	100	80.3	76.4	67.5
$m_5 + m_{10}$	m_{10}	100	83.3	75.1	66.6
	Ave‡	100	84.3	72.3	65.4

(a) Ideal binary mask (upper bound for separation) [67].
(b) Proposed method.
(c) MAXVQ method [23].
(d) HMS [42, 68].
† f_i and m_j show speech signals of ith female and jth male speakers, respectively.
‡ Averaged P_i values over the ten speech files.

obtained for the HMS method show that in terms of SNR, that the HMS method outperforms MAXVQ, although the HMS approach suffers severely from crosstalk.

Figure 8 shows an example of two separated speech signals from their mixture using the proposed technique. The speech signals in the upper panels, (a) and (b), are the original signals of a male speaker and a female speaker. The middle panel (c) shows their mixture and the signals in the bottom panels, (d) and (e), are the corresponding separated signals. As Figure 8 shows, the proposed separation model separates the individual signals accurately. In particular, notice the portion where a voiced segment overlaps with an unvoiced segment (around 1.2 seconds). Though it is very difficult to extract unvoiced segments in such a situation, our proposed algorithm extracts unvoiced segments even with an energy lower than the voiced segment. The performance of our model can be improved by increasing the number of Gaussian mixtures. However, this in turn increases the decoding time.

9. CONCLUSIONS

We have presented a maximum likelihood approach to perform separation of two speech signals from a mixture. The problem is too ill conditioned (due to the noninvertibility of mixing matrix) to be solved using the common BSS techniques. Therefore, we use a special case of BSS methods that rely on a priori knowledge of speakers. In contrast with previous methods, we take into account the characteristics of the speech signal where each component of the speech signal, vocal-tract-related filter, and excitation signal is separated using two different strategies, namely, a probabilistic approach and a deterministic approach. Using this hybrid model we have linked the underdetermined BSS techniques with CASA algorithms. The separability of our model outperforms the two techniques that use either the underdetermined BSS or CASA approaches. These results reveal that in order to obtain a better separation technique we should incorporate the speech signal characteristic into probabilistic models as we did for the special case for fundamental frequencies. In this paper, we only incorporate the harmonicity cue in our model. We believe that further work should be done to first incorporate the other cues presented in CASA models (onset, offset cues) into the system and second extend the model such that it includes the dynamic characteristics of the speech signal into the model using an HMM model.

APPENDICES

A. CONSTRUCTING THE EXCITATION SIGNALS

The vocal-cord-related signal, which is commonly referred to as the excitation signal, is in the form of an impulse train and a white noise process for voiced speech signals and unvoiced speech signals, respectively, and represented by

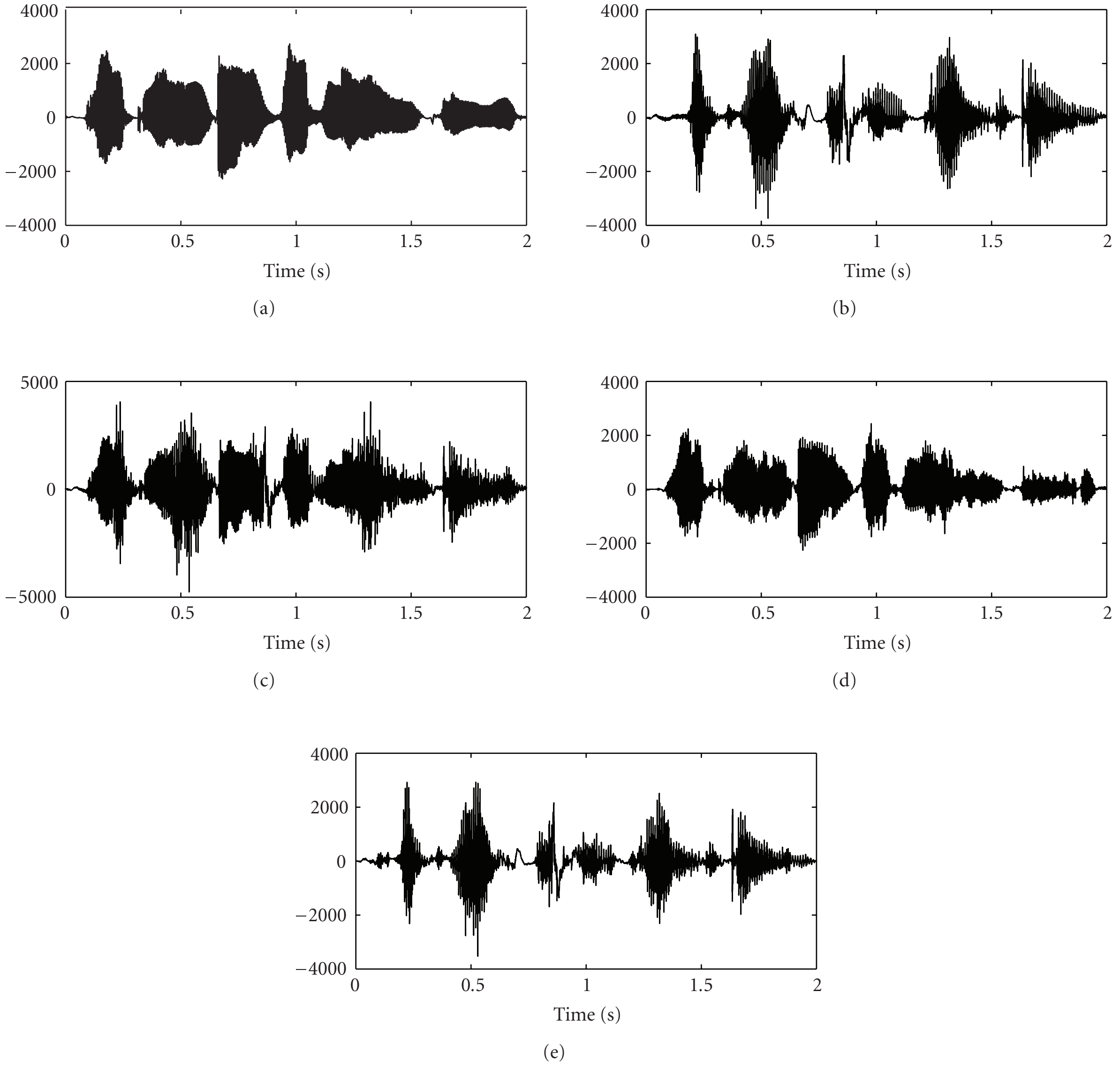
$$e^t(n) = \begin{cases} \displaystyle\sum_{\tau=-\infty}^{+\infty} \delta\left(n - \frac{\tau}{f_0}\right) w^t(n), & \text{voiced speech}, \\ \mathcal{N}(0,\sigma^2)\, w^t(n), & \text{unvoiced speech}, \end{cases} \qquad n = 1,2,\ldots,N, \tag{A.1}$$

where $\delta(n)$ denotes the impulse function and f_0 is the fundamental frequency of the voiced speech signal. $\mathcal{N}(0,\sigma^2)$ represents zero mean white Gaussian noise with variance σ^2. Also, $w^t(n)$ represents the analysis window applied for short-term processing. Taking the D-point Fourier transform of both sides of (A.1), gives

$$e^f(d) = \begin{cases} \displaystyle\sum_{m=1}^{M(\omega_0)} w^f(d - m\omega_0), & \text{voiced speech}, \\ \sigma^2, & \text{unvoiced speech}, \end{cases} \qquad d = 1,2,\ldots,D, \tag{A.2}$$

in which $\omega_0 = 2\pi f_0/D$, where f_0 denotes the fundamental frequency of the voiced speech signal. Also, $M(\omega_0)$ represents the number of harmonics in the speech bandwidth and $w^f(d)$ represents the Fourier transform of the analysis window. As shown in [48] for better perceptual results, we can also use a windowed impulse train with a period equal to 100 Hz during the unvoiced speech. Thus the excitation signal can be reexpressed as

$$e^f(d) = \begin{cases} \displaystyle\sum_{m=1}^{M(\omega_0)} w^f(d - m\omega_0), & \text{voiced speech}, \\ \displaystyle\sum_{m=1}^{M(\omega_0^{\text{un}})} w^f(d - m\omega_0^{\text{un}}), & \text{unvoiced speech} \\ & (\omega_0^{\text{un}} = 2\pi 100), \end{cases} \qquad d = 1,2,\ldots,D, \tag{A.3}$$



(a) (b) (c) (d) (e)

Figure 8: An example of the separated speech signals obtained by ML vocal-tract-related filter: (a) original female speaker, (b) original male speaker, (c) mixed speech signal, (d) separated female speaker, and (e) separated male speaker.

where σ^2 is assumed to be included in the vocal-tract-related filter. Therefore, the excitation signal can be interpreted as the windowed impulse with a period equal to the fundamental frequency of the analysis frame for voiced frames and 100 Hz for unvoiced frames.

B. SEPARATION USING THE IDEAL BINARY MASK

In this appendix, we describe the ideal binary mask separation method [67]. We then, based on this method, explain a method for measuring the amount of crosstalk introduced in a separation system. As explained in Section 7, the log spectral vector of the mixed signal and those of the underlying signals are related through (9). Suppose that we have access to the original underlying signals $\mathbf{x}_1$ and $\mathbf{x}_2$. From these log spectral vectors, we construct two ideal binary masks in the form

$$\text{mask}_1^{\text{ideal}}(d) = \begin{cases} 0, & x_1(d) < x_2(d), \\ 1, & x_1(d) \geq x_2(d), \end{cases}$$
$$\text{mask}_2^{\text{ideal}}(d) = \begin{cases} 0, & x_2(d) < x_1(d), \\ 1, & x_2(d) \geq x_1(d), \end{cases} \qquad d = 1, 2, \ldots, D. \tag{B.4}$$

Multiplying the ideal masks by the mixed log spectral vector gives an estimate of the log spectral vectors of the underlying signals, that is,

$$\begin{aligned} \hat{\mathbf{x}}_1^{\text{ideal}} &= \mathbf{y} \times \mathbf{mask}_1^{\text{ideal}}, \\ \hat{\mathbf{x}}_2^{\text{ideal}} &= \mathbf{y} \times \mathbf{mask}_2^{\text{ideal}}. \end{aligned} \tag{B.5}$$

The estimated log spectral vector along with the phase of the mixed signal is used to recover the speech signals in the time domain, thus

$$\hat{\mathbf{x}}_i^{t,\text{ideal}} = \mathcal{F}_{\mathcal{D}}{}^{-1}\left(10^{\hat{\mathbf{x}}_i^{\text{ideal}}} \cdot \varphi_{\mathbf{y}}\right), \quad i \in \{1,2\}, \tag{B.6}$$

where $\mathcal{F}_{\mathcal{D}}{}^{-1}$ denotes the inverse Fourier transform and $\varphi_{\mathbf{y}} = \angle \mathbf{y}^f$ is the phase of the Fourier transform of the windowed mixed signal.

High-quality separated speech signals are obtained using the ideal binary masks. Inspired by this, the goal of many SCSS techniques in both CASA and BSS contexts is to obtain the separated signals whose binary masks is as close as possible to the ideal binary masks. In this way, a criterion to evaluate a separation system is to compare the binary masks obtained from the estimated log spectral vector to the ideal binary masks obtained from the original log spectral vectors. Or equivalently, compare the separated signals to the obtained signals from applying the ideal binary masks.

C. COMPUTING THE CROSSTALK SUPPRESSION RATE

The binary masks are also used to compute the amount of crosstalk introduced in separation algorithms in the following way. Let $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ be the estimated log spectral vectors obtained by a separation system. We construct two binary masks, $\mathbf{mask}_1$ and $\mathbf{mask}_2$, from the estimated log spectral vectors $\hat{\mathbf{x}}_1$ and $\hat{\mathbf{x}}_2$ in a way similar to (B.4). Accepting the ideal binary mask approach as a system that suppresses all crosstalks, two new masks are generated by the routine described as follows. When we apply a binary mask, we in fact suppress all frequency bins that belong to the other speakers and keep the ones we wish to recover. Therefore, when a frequency bin in an ideal binary mask is zero while in the estimated binary mask it is one, it means that the applied separation technique is not able to suppress this frequency bin. Thus the crosstalk mask contains all frequency bins that should be suppressed, but the separation algorithm fails to suppress. Consequently, the signal which is obtained by applying the crosstalk mask is the crosstalk signal. This process can be explained as follows.

Suppose we obtain the crosstalk masks, $\mathbf{mask}_1^{\text{crosstalk}}$ and $\mathbf{mask}_2^{\text{crosstalk}}$. The log spectral crosstalk vectors are obtained from

$$\begin{aligned} \mathbf{x}_1^{\text{crosstalk}} &= \mathbf{y} \times \mathbf{mask}_1^{\text{crosstalk}}, \\ \mathbf{x}_2^{\text{crosstalk}} &= \mathbf{y} \times \mathbf{mask}_2^{\text{crosstalk}} \end{aligned} \tag{C.7}$$

and similar to previous discussions, the windowed crosstalk signals in the time domain are given by

$$\mathbf{x}_i^{t,\text{crosstalk}} = \mathcal{F}_{\mathcal{D}}{}^{-1}\left(10^{\mathbf{x}_i^{\text{crosstalk}}} \cdot \varphi_{\mathbf{y}}\right), \quad i \in \{1,2\}. \tag{C.8}$$

The whole crosstalk utterance can be obtained from the overlap-add method performed on windowed crosstalk segments. Finally, we define the ratio between the energy of the crosstalk signal to the energy of the signal recovered by the ideal binary mask as the crosstalk-to-signal (CTS) ratio, given by

$$\text{CTS}_i = \frac{\sum_n \left(x_i^{t,\text{crosstalk}}(n)\right)^2}{\sum_n \left(\hat{x}_i^{t,\text{ideal}}(n)\right)^2}, \quad n = 1,2,\ldots,\aleph,\ i \in \{1,2\}, \tag{C.9}$$

where $\aleph$ represents the length of the whole utterance. The percentage of crosstalk suppression is then defined as

$$P_i = \left(1 - \text{CTS}_i\right) \times 100\%, \quad i \in \{1,2\}. \tag{C.10}$$

We use this criterion to evaluate the system's performance.

ACKNOWLEDGMENT

The authors wish to thank the Natural Sciences and Engineering Research Council (NSERC) of Canada for funding this project.

REFERENCES

[1] C. Jutten and J. Herault, "Blind separation of sources, part I. An adaptive algorithm based on neuromimetic architecture," *Signal Processing*, vol. 24, no. 1, pp. 1–10, 1991.

[2] P. Comon, "Independent component analysis. A new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, 1994.

[3] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.

[4] S.-I. Amari and J.-F. Cardoso, "Blind source separation-semiparametric statistical approach," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2692–2700, 1997.

[5] A. S. Bregman, *Auditory Scene Analysis*, MIT Press, Cambridge, Mass, USA, 1994.

[6] G. J. Brown and M. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, no. 4, pp. 297–336, 1994.

[7] M. Cooke and D. P. W. Ellis, "The auditory organization of speech and other sources in listeners and computational models," *Speech Communication*, vol. 35, no. 3-4, pp. 141–177, 2001.

[8] D. P. W. Ellis, "Using knowledge to organize sound: the prediction-driven approach to computational auditory scene analysis and its application to speech/nonspeech mixtures," *Speech Communication*, vol. 27, no. 3-4, pp. 281–298, 1999.

[9] T. Nakatani and H. G. Okuno, "Harmonic sound stream segregation using localization and its application to speech stream segregation," *Speech Communication*, vol. 27, no. 3, pp. 209–222, 1999.

[10] G. J. Brown and D. L. Wang, "Separation of speech by computational auditory scene analysis," in *Speech Enhancement: What's New?*, J. Benesty, S. Makino, and J. Chen, Eds., pp. 371–402, Springer, New York, NY, USA, 2005.

[11] C. J. Darwin and R. P. Carlyon, "Auditory grouping," in *The Handbook of Perception and Cognition*, B. C. J. Moore, Ed., vol. 6, chapter Hearing, pp. 387–424, Academic Press, Orlando, Fla, USA, 1995.

[12] D. L. Wang and G. J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 684–697, 1999.

[13] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1135–1150, 2004.

[14] G. J. Jang and T. W. Lee, "A probabilistic approach to single channel blind signal separation," in *Proceedings of Advances in Neural Information Processing Systems (NIPS '02)*, pp. 1173–1180, Vancouver, British Columbia, Canada, December 2002.

[15] C. Fevotte and S. J. Godsill, "A Bayesian approach for blind separation of sparse sources," *IEEE Transaction on Speech and Audio Processing*, vol. 4, no. 99, pp. 1–15, 2005.

[16] M. Girolami, "A variational method for learning sparse and overcomplete representations," *Neural Computation*, vol. 13, no. 11, pp. 2517–2532, 2001.

[17] T.-W. Lee, M. S. Lewicki, M. Girolami, and T. J. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Signal Processing Letters*, vol. 6, no. 4, pp. 87–90, 1999.

[18] T. Beierholm, B. D. Pedersen, and O. Winther, "Low complexity Bayesian single channel source separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 5, pp. 529–532, Montreal, Quebec, Canada, May 2004.

[19] S. Roweis, "One microphone source separation," in *Proceedings of Advances in Neural Information Processing Systems (NIPS '00)*, pp. 793–799, Denver, Colo, USA, October-November 2000.

[20] M. J. Reyes-Gomez, D. P. W. Ellis, and N. Jojic, "Multiband audio modeling for single-channel acoustic source separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04)*, vol. 5, pp. 641–644, Montreal, Quebec, Canada, May 2004.

[21] A. M. Reddy and B. Raj, "A minimum mean squared error estimator for single channel speaker separation," in *Proceedings of the 8th International Conference on Spoken Language Processing (INTERSPEECH '04)*, pp. 2445–2448, Jeju Island, Korea, October 2004.

[22] T. Kristjansson, H. Attias, and J. Hershey, "Single microphone source separation using high resolution signal reconstruction," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 2, pp. 817–820, Montreal, Quebec, Canada, May 2004.

[23] S. T. Rowies, "Factorial models and refiltering for speech separation and denoising," in *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH '03)*, vol. 7, pp. 1009–1012, Geneva, Switzerland, September 2003.

[24] T. Virtanen and A. Klapuri, "Separation of harmonic sound sources using sinusoidal modeling," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '00)*, vol. 2, pp. 765–768, Istanbul, Turkey, June 2000.

[25] T. F. Quatieri and R. G. Danisewicz, "An approach to cochannel talker interference suppression using a sinusoidal model for speech," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, no. 1, pp. 56–69, 1990.

[26] E. A. Wan and A. T. Nelson, "Neural dual extended Kalman filtering: applications in speech enhancement and monaural blind signal separation," in *Proceedings of the 7th IEEE Workshop on Neural Networks for Signal Processing (NNSP '97)*, pp. 466–475, Amelia Island, Fla, USA, September 1997.

[27] J. R. Hopgood and P. J. W. Rayner, "Single channel nonstationary stochastic signal separation using linear time-varying filters," *IEEE Transactions on Signal Processing*, vol. 51, no. 7, pp. 1739–1752, 2003.

[28] R. Balan, A. Jourjine, and J. Rosca, "AR processes and sources can be reconstructed from degenerative mixtures," in *Proceedings of the 1st International Workshop on Independent Component Analysis and Signal Separation (ICA '99)*, pp. 467–472, Aussois, France, January 1999.

[29] J. Rouat, Y. C. Liu, and D. Morissette, "A pitch determination and voiced/unvoiced decision algorithm for noisy speech," *Speech Communication*, vol. 21, no. 3, pp. 191–207, 1997.

[30] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '93)*, vol. 2, pp. 728–731, Minneapolis, Minn, USA, April 1993.

[31] M. Wu, D. L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, 2003.

[32] T. Nishimoto, S. Sagayama, and H. Kameoka, "Multi-pitch trajectory estimation of concurrent speech based on harmonic GMM and nonlinear Kalman filtering," in *Proceedings of the 8th International Conference on Spoken Language Processing (INTERSPEECH '04)*, vol. 1, pp. 2433–2436, Jeju Island, Korea, October 2004.

[33] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 6, pp. 708–716, 2000.

[34] Y.-H. Kwon, D.-J. Park, and B.-C. Ihm, "Simplified pitch detection algorithm of mixed speech signals," in *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS '00)*, vol. 3, pp. 722–725, Geneva, Switzerland, May 2000.

[35] D. P. Morgan, E. B. George, L. T. Lee, and S. M. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 5, pp. 407–424, 1997.

[36] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, "Performance evaluation of three features for model-based single channel speech separation problem," in *Proceedings of the 9th International Conference on Spoken Language Processing (INTERSPEECH '06)*, pp. 2610–2613, Pittsburgh, Pa, USA, September 2006.

[37] G. Hu and D. Wang, "Auditory segmentation based on onset and offset analysis," to appear in *IEEE Transactions on Audio, Speech, and Language Processing*.

[38] D. Ellis, "Model-based scene analysis," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. Wang and G. Brown, Eds., Wiley/IEEE Press, New York, NY, USA, 2006.

[39] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection," *Journal of the Acoustical Society of America*, vol. 60, no. 4, pp. 911–918, 1976.

[40] A. de Cheveigné and H. Kawahara, "Multiple period estimation and pitch perception model," *Speech Communication*, vol. 27, no. 3, pp. 175–185, 1999.

[41] M. Weintraub, "A computational model for separating two simultaneous talkers," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '86)*, vol. 11, pp. 81–84, Tokyo, Japan, April 1986.

[42] B. A. Hanson and D. Y. Wong, "The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering speech," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '84)*, vol. 2, pp. 18A. 5. 1–18A. 5. 4, San Diego, Calif, USA, March 1984.

[43] P. P. Kanjilal and S. Palit, "Extraction of multiple periodic waveforms from noisy data," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '94)*, vol. 2, pp. 361–364, Adelaide, SA, Australia, April 1994.

[44] M. R. Every and J. E. Szymanski, “Separation of synchronous pitched notes by spectral filtering of harmonics,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1845–1856, 2006.

[45] R. C. Maher and J. W. Beauchamp, “Fundamental frequency estimation of musical signals using a two-way mismatch procedure,” *Journal of the Acoustical Society of America*, vol. 95, no. 4, pp. 2254–2263, 1994.

[46] M. Karjalainen and T. Tolonen, “Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’99)*, vol. 2, pp. 929–932, Phoenix, Ariz, USA, March 1999.

[47] M. Cooke, “Modeling auditory processing and organization,” Doctoral thesis, Cambridge University, Cambridge, UK, 1991.

[48] R. J. McAulay and T. F. Quatieri, “Sinusoidal coding,” in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds., Elsevier, New York, NY, USA, 1995.

[49] T. F. Quatieri, *Discrete-Time Speech Signal Processing Principle and Practice*, Prentice-Hall, Englewood Cliffs, NJ, USA, 2001.

[50] E. Yair, Y. Medan, and D. Chazan, “Super resolution pitch determination of speech signals,” *IEEE Transactions on Signal Processing*, vol. 39, no. 1, pp. 40–48, 1991.

[51] P. Martin, “Comparison of pitch detection by cepstrum and spectral comb analysis,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’82)*, vol. 7, pp. 180–183, Paris, France, May 1982.

[52] R. Meddis and M. Hewitt, “Virtual pitch and phase sensitivity of a computer model of the auditory periphery I: pitch identification,” *Journal of the Acoustical Society of America*, vol. 89, no. 6, pp. 2866–2882, 1991.

[53] R. Meddis and L. O’Mard, “A unitary model of pitch perception,” *Journal of the Acoustical Society of America*, vol. 102, no. 3, pp. 1811–1820, 1997.

[54] N. Chandra and R. E. Yantorno, “Usable speech detection using the modified spectral autocorrelation peak to valley ratio using the LPC residual,” in *Proceedings of 4th IASTED International Conference on Signal and Image Processing*, pp. 146–149, Kaua’i Marriott, Hawaii, USA, August 2002.

[55] Y. A. Mahgoub and R. M. Dansereau, “Voicing-state classification of co-channel speech using nonlinear state-space reconstruction,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’05)*, vol. 1, pp. 409–412, Philadelphia, Pa, USA, March 2005.

[56] A. R. Kizhanatham, N. Chandra, and R. E. Yantorno, “Co-channel speech detection approaches using cyclostationarity or wavelet transform,” in *Proceedings of 4th IASTED International Conference on Signal and Image Processing*, Kaua’i Marriott, Hawaii, USA, August 2002.

[57] D. S. Benincasa and M. I. Savic, “Voicing state determination of co-channel speech,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP ’98)*, vol. 2, pp. 1021–1024, Seattle, Wash, USA, May 1998.

[58] M. H. Radfar, R. M. Dansereau, and A. Sayadiyan, “A joint identification-separation technique for single channel speech separation,” in *Proceedings of the 12th IEEE Digital Signal Processing Workshop (DSP ’06)*, pp. 76–81, Grand Teton National Park, Wyo, USA, September 2006.

[59] M. H. Radfar, A. Sayadiyan, and R. M. Dansereau, “A new algorithm for two-talker pitch tracking in single channel paradigm,” in *Proceedings of International Conference on Signal Processing (ICSP ’06)*, Guilin, China, November 2006.

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

[61] K. K. Paliwal and L. D. Alsteris, “On the usefulness of STFT phase spectrum in human listening tests,” *Speech Communication*, vol. 45, no. 2, pp. 153–170, 2005.

[62] D. B. Paul, “The spectral envelope estimation vocoder,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 4, pp. 786–794, 1981.

[63] C. de Boor, *A Practical Guide to Splines*, Springer, New York, NY, USA, 1978.

[64] D. Talkin, “A robust algorithm for pitch tracking (RAPT),” in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds., pp. 495–518, Elsevier, Amsterdam, The Netherlands, 1995.

[65] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic, Norwell, Mass, USA, 1992.

[66] W. C. Chu, “Vector quantization of harmonic magnitudes in speech coding applications - a survey and new technique,” *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 17, pp. 2601–2613, 2004.

[67] D. Wang, “On ideal binary mask as the computational goal of auditory scene analysis,” in *Speech Separation by Humans and Machines*, P. Divenyi, Ed., pp. 181–197, Kluwer Academic, Norwell, Mass, USA, 2005.

[68] J. A. Naylor and S. F. Boll, “Techniques for suppression of an interfering talker in co-channel speech,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP ’87)*, vol. 1, pp. 205–208, Dallas, Tex, USA, April 1987.