Determining Mixing Parameters From Multispeaker Data Using Speech-Specific Information

B. Yegnanarayana, *Senior Member, IEEE*, R. Kumara Swamy, *Member, IEEE*, and K. Sri Rama Murty

***Abstract*—In this paper, we propose an approach for processing multispeaker speech signals collected simultaneously using a pair of spatially separated microphones in a real room environment. Spatial separation of microphones results in a fixed time-delay of arrival of speech signals from a given speaker at the pair of microphones. These time-delays are estimated by exploiting the impulse-like characteristic of excitation during speech production. The differences in the time-delays for different speakers are used to determine the number of speakers from the mixed multispeaker speech signals. There is difference in the signal levels due to differences in the distances between the speaker and each of the microphones. The differences in the signal levels dictate the values of the mixing parameters. Knowledge of speech production, especially the excitation source characteristics, is used to derive an approximate weight function for locating the regions specific to a given speaker. The scatter plots of the weighted and delay-compensated mixed speech signals are used to estimate the mixing parameters. The proposed method is applied on the data collected in actual laboratory environment for an underdetermined case, where the number of speakers is more than the number of microphones. Enhancement of speech due to a speaker is also examined using the information of the time-delays and the mixing parameters, and is evaluated using objective measures proposed in the literature.**

***Index Terms*—Excitation source, mixing parameters, multispeaker data, speaker localization, time-delay estimation.**

I. Introduction

ONE of the interesting and scientifically challenging problems is to separate signals generated by different sources from the mixture of signals picked up by spatially distributed sensors, when there is limited or no knowledge of the nature of the signals and the number of sources. While the motivation for such problems is the so called *cocktail party problem* [1], the problem is addressed as a blind source separation (BSS) problem [2] with possible applications in other areas as well, such as in telecommunications and medical signal processing. When the signals from spatially distributed sources are mixed in a real room environment, the signal at any of the sensors in the room is a mixture of the convolved source signals. Each source signal at a sensor is convolved with the impulse response of the acoustic path from the source to the receiver (sensor). For a BSS problem with p sensors and q sources, the mixed signal at the ith sensor is given by

$$x_i[n] = \sum_{j=1}^{q} \sum_{m} a_{ij}[m] s_j[n-m] + v_i[n], \quad i \in \{1, 2, \ldots, p\}. \tag{1}$$

Here $x_i[n]$ is the mixed signal at the ith sensor, $s_j[n]$ is the signal generated from the jth source, $v_i[n]$ is the additive noise at the ith sensor and $a_{ij}[n]$ represents the impulse response of the acoustic path between jth source and ith sensor. There are $p \times q$ such responses for a room with p spatially distributed sensors and q spatially distributed sources. Given N observations of the $x_i[n]$, $n = 1, 2, \ldots, N$, the problem in BSS is to determine the number of sources, and the individual sources $s_j[n]$, $j = 1, 2, \ldots, q$. This case of source signal separation problem is called *convolutive* BSS problem. If the impulse response of the room is varying with time, then it becomes a case of nonstationary response convolutive BSS problem. In general, the impulse response of the room between two points is assumed to be stationary.

The separated signals are generally obtained using linear filters $w_{ij}[n]$. The estimated source signals $\hat{s}_j[n]$ are given by

$$\hat{s}_j[n] = \sum_{i=1}^{p} \sum_{m} w_{ij}[m] x_i[n-m], \quad j \in \{1, 2, \ldots, q\}. \tag{2}$$

The objective is to derive the deconvolving filters $w_{ij}[n]$, usually of finite-impulse response (FIR) type. The problem of source separation from convolutive mixtures can be treated as inversion of impulse response of the acoustic path imposed on each sensor. Several methods have been proposed, both in time-domain and frequency-domain, to achieve source separation from convolutive mixtures.

Time-domain approaches consider the impulse response along the acoustic path as an FIR filter, and estimate the filter coefficients in the time-domain [3], [4]. Though these methods are successful, the number of sources must be exactly known, as they are extracted simultaneously. Moreover, the convergence of these methods depends critically on the initial values of the filter coefficients and on the selection of the step size for updating them.

Frequency-domain approaches transform the problem of convolutive mixtures into an instantaneous mixture, and solve the instantaneous BSS problem simultaneously for individual frequencies [5], [6]. Since the filter parameters in the frequency-domain are mutually orthogonal, updating one parameter does not influence the rest. Though the mutual independence of the

Manuscript received July 02, 2008; revised January 06, 2009. Current version published July 06, 2009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hiroshi Sawada.

B. Yegnanarayana is with the International Institute of Information Technology, Hyderabad 500 032, India (e-mail: yegna@iiit.ac.in).

R. K. Swamy is with the Department of Electronics and Communications Engineering, Siddaganga Institute of Technology, Tumkur 572103, India (e-mail: hyrkswamy@gmail.com).

K. S. R. Murty is with the Department of Computer Science and Engineering, Indian Institute of Technology-Madras, Chennai 600 036, India (e-mail: ksrmurty@gmail.com).

Digital Object Identifier 10.1109/TASL.2009.2016230

frequency bins is attractive, the frequency domain methods introduce scaling and permutation problems [7]. A robust method for solving the permutation problem of the frequency-domain BSS was proposed by Sawada *et al.* [8].

The linear filter methods (2) do not work well for underdetermined cases $(p < q)$. Instead, methods based on time–frequency masking [9], [10] or l_1-norm minimization [11] are widely used to separate such underdetermined mixtures. The disjointness of the source signals in the time–frequency representations is exploited for extracting them from mixed signals. Most of the exiting techniques for underdetermined convolutive BSS rely on time difference of arrival of source signals at the microphones. Hence, they work effectively under low reverberant conditions. However, under reverberant conditions, time-delay estimation becomes unreliable, and hence such techniques do not work well. Sawada *et al.* proposed a two stage hierarchical time–frequency masking for separation of convolutive mixtures under reverberant conditions [12].

Several source separation methods were developed specifically for the separation of speech signals [13], [14]. In most cases, the problem of underdetermined representation $(p < q)$ was assumed, and specific characteristics of speech were exploited to achieve the separation of speech signals. In [15], a least squares optimization method was proposed to estimate the mixing matrix and the sources, assuming that the cross-correlation functions at multiple times give sufficient constraints for the unknown channels. Attempts are also made to exploit the model of speech production for speech signal separation for the case of underdetermined representation [13], [16]. For noise-free case, a method for signal separation and estimation of the mixing matrix has been proposed under the following assumptions [17]. 1) Source signals are zero mean and mutually independent, and at most one of the sources is Gaussian. 2) The source signals are sufficiently sparse (no overlapping intervals). 3) The mixing matrix is a constant matrix with full rank.

By exploiting the speech-specific characteristics, one can avoid estimation of the impulse response of the room. Also, it may be possible to relax several assumptions usually made in developing solutions to the convolutive BSS problem. In this process, it is also likely that complete characterization of the mixing/demixing process may not be feasible. Thus instead of separation of signals, only enhancement of source signals over others may be possible. In this paper, we address some issues related to the separation of mixed speech signals collected by a pair of microphones in a real live room environment, when people are speaking simultaneously. The speakers are located at different distances from the microphones. No assumptions are made on the number of speakers, or on the way they speak. The average (over frequency) reverberation time of the recording environment is about 0.5 s. The speakers are located about 1–2 m from the two microphones, and the direct component of speech at the microphones is significantly large most of the time, compared to the reverberation component and the environmental noise component. The reverberant component is assumed to be due to diffused sound, not due to any specular reflections/echoes. The reverberant component and the environment noise are clubbed into a single noise component $v_i[n]$ in the mixed signal $x_i[n]$ at each microphone. Therefore, the mixed signal at the ith microphone can be written in terms of the direct component a_{ij} from each of the jth source and a noise component as given as follows:

$$x_i[n] = \sum_{j=1}^{q} a_{ij} s_j[n - \tau_{ij}] + v_i[n], \quad i \in \{1, 2, \ldots, p\} \quad (3)$$

where τ_{ij} is the delay in samples.

The proposed signal processing approach depends on: 1) exploiting the impulse sequence characteristic in the excitation signal of the speech production, 2) the time-delay between the direct components at the two microphones due to a speaker, 3) the lower strengths of impulses in the reverberant and noise components, 4) the increased strength of the impulses due to coherent addition of the strengths of the impulse sequence due to a speaker, and 5) simultaneously decreased strengths of impulses of the other speakers due to incoherent addition. In the process of this analysis, we will show that the parameters of the mixing matrix need not be known completely. We will also show that the parameters (amplitudes and delays) cannot be independent in a mixing environment of a real room. It is also interesting to note that the reverberation and ambient noise do not affect critically as long as there is a significant direct component at the microphones.

Some of these issues were addressed in studies on source localization, especially, the estimation of the time delays. The methods used in this paper for delay estimation are based on exploiting the Hilbert envelope of the linear prediction residual of speech. The advantages of using Hilbert envelope of the LP residual is discussed for a single source case in [18] by comparing with some of the state-of-the-art methods including the generalized cross-correlation method [19].

The focus in this paper is to exploit the speech-specific characteristics to address the following issues related to the separation of mixed speech signals: 1) determining the time-delays, 2) deriving the mixing parameters, and 3) enhancement of speech due to one speaker relative to others using the information about the time-delays and the mixing parameters. The organization of this paper is as follows. In Section II, the basis for the proposed signal processing approach for convolutive BSS of multispeaker data is presented. The specific signal processing operations on multispeaker data are described in Section III. A method for deriving the mixing parameters using time-delay and weight function is proposed in Section IV. Some experimental results are presented in Section V to demonstrate the effectiveness of the proposed method. In Section VI, we discuss a method to enhance speech of the target speakers using the estimated time-delays and the mixing parameters. In Section VII, we discuss the results of objective evaluation of the proposed method using synthetic anechoic mixtures and real room recordings, and comparison with the existing methods. Finally, Section VIII gives a summary and discussion on some research issues arising out of this study.

II. Basis for the Proposed Approach for Processing Multispeaker Data

In the current scenario, speech is collected from multiple speakers speaking simultaneously using a pair of microphones. In this case, the direct components of each speaker at the pair of microphones have some time-delay between them due

to differences in distances. There is also a difference in the signal levels of the direct components, due to differences in the distances. This contributes to different amplitude values of the signals collected at the two microphones. The multispeaker signals collected at the two microphones can be written as

$$x_1[n] = \sum_{j=1}^{q} a_{1j} s_j[n - \tau_{1j}] + v_1[n], \quad n = 1, 2, \ldots, N \tag{4a}$$

and

$$x_2[n] = \sum_{j=1}^{q} a_{2j} s_j[n - \tau_{2j}] + v_2[n], \quad n = 1, 2, \ldots, N \tag{4b}$$

where the time index n refers to the sampling instant, and N refers to the total number of samples. Note that, because of the time-delays, the values of the mixing parameters, given by

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1j} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2j} & \cdots & a_{2q} \end{bmatrix} \tag{5}$$

cannot be arbitrary. The distances of a speaker from the microphones dictate the corresponding time-delays and the signal amplitudes, which in turn force some relation among the mixing parameters. This also indicates that if an arbitrary choice is made for the mixing parameters and the time-delays for simulating multispeaker data, such signals may not represent mixed signals that occur in practice.

We propose primarily a signal processing approach for estimating the mixing parameters. Since the relative positions of the excitation impulses due to a speaker remain unchanged at different microphones, the cross-correlation function of the processed mixed signals (highlighting these impulses) can be used to estimate the time-delay at a pair of microphones. The high signal-to-noise ratio (SNR) regions around these impulses due to a given speaker are relatively less affected by the speech due to other speakers. Hence, the multispeaker signals around the high SNR regions can be exploited for estimating the mixing parameters.

Consider a situation where the signal from an impulse generator is collected at a pair of spatially separated microphones at distances d_1 and d_2. In such a situation, the measured distances of the impulse generator from the microphones are related to the observed amplitudes (a_1 and a_2) at the microphones by

$$\frac{d_1}{d_2} = \frac{a_2}{a_1} = r. \tag{6}$$

Since

$$d_1 - d_2 = \tau c \tag{7}$$

where c is the velocity of sound in air and τ time-delay of arrival between the two microphones, we obtain from (6) and (7)

$$d_1 = \frac{\tau r c}{r - 1} \tag{8a}$$

and

$$d_2 = \frac{\tau c}{r - 1}. \tag{8b}$$

Here, the time-delay τ can be estimated by finding the cross-correlation between the two microphone signals, and the amplitude ratio r can be computed by observing the amplitudes of the impulses at both the microphones. The situation considered here is an ideal case, where the sound source is an impulse generator. The rest of the paper deals with applying the theoretical basis developed here to multispeaker speech signals collected in a laboratory environment. This task mainly involves determining the number of speakers, estimating their respective time-delays, localizing the speakers with respect to the microphones, and thereby estimating the mixing parameters.

III. Time-Delay Estimation

In a multispeaker multimicrophone scenario, assuming that the speakers are stationary with respect to the microphones, there exists a fixed time-delay of arrival of the speech signals (between every pair of microphones) from a given speaker. The time-delays corresponding to different speakers can be estimated using the cross-correlation function of the multispeaker signals. Positions of dominant peaks in the cross-correlation function denote the time-delays due to all the speakers at the pair of microphones. However, the cross-correlation function of the multispeaker signals does not show prominent peaks at the time-delays. This is mainly because of the damped-sinusoid-like components in the speech signal due to resonances of the vocal tract, and also because of the effects of reverberation and noise. These effects can be reduced by exploiting the characteristics of the excitation source of speech [20].

During the production of voiced speech, the vocal tract system is excited by a quasiperiodic sequence of impulse-like excitations. These impulse-like excitations occur at the instants of glottal closure (GCI) within each pitch period. In the vicinity of these impulses, the speech signal exhibits a high SNR relative to the other regions. In order to highlight the high SNR regions in the speech signal, linear prediction (LP) residual is derived from the speech signal using the autocorrelation method [21]. The LP residual removes second-order correlations among the samples of the signal, and produces large amplitude fluctuations around the instants of significant excitation. The LP residual corresponds to an estimate of the excitation source of the speech signal. Note that the LP analysis of a mixed speech signal also produces uncorrelated samples in the LP residual, where large amplitude residual samples approximately correspond to the excitation part in the mixed signal. The cross-correlation function of the LP residual signals from the two microphone mixed speech signals is not likely to yield strong peaks because of the large amplitude fluctuations of random polarity around the GCIs, as shown in Fig. 1(b). The high SNR regions around the GCIs can be highlighted by computing the Hilbert envelope (HE) of the LP residual [22]. The Hilbert envelope $h[n]$ of the LP residual signal $e[n]$ is given by

$$h[n] = \sqrt{e^2[n] + e_h^2[n]} \tag{9}$$

where $e_h[n]$ is the Hilbert transform of $e[n]$ [23]. The HE of the LP residual is shown in Fig. 1(c).

The cross-correlation function of the HEs of the LP residual signals derived from the multispeaker signals is used to estimate the time-delays [24]. Apart from the large amplitudes around the instants of significant excitation, the HE contains a large number

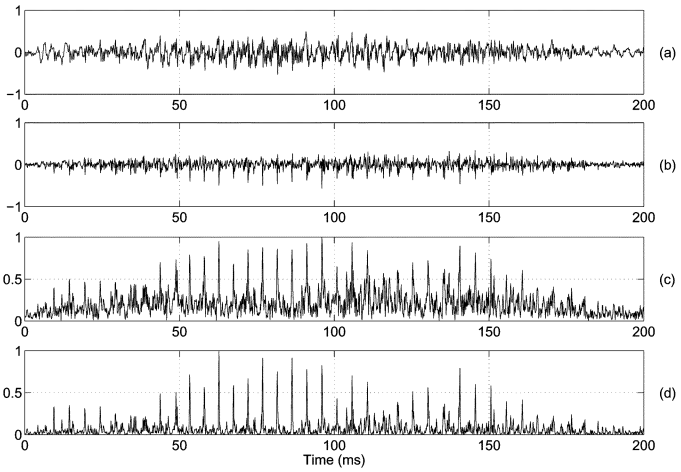

Fig. 1. (a) A 200-ms segment of mixed speech signal, (b) its LP residual, (c) HE of LP residual, and (d) HE after emphasizing the GCIs.

of smaller positive values also, which may result in spurious peaks in the cross-correlation function. Therefore, the regions around the instants of significant excitation can be further emphasized by dividing the square of each sample of the HE by the moving average of the HE computed over a short window (about 4 ms, i.e., less than the average pitch period) around the sample. The computation of the preprocessed HEs is as follows [20]:

$$g_i[n] = \frac{h_i^2[n]}{\frac{1}{2M+1}\sum_{m=n-M}^{n+M} h_i[m]}, \qquad i \in \{1, 2, \ldots, p\} \quad (10)$$

where $h_i[n]$ is the HE of the LP residual of the multispeaker signal collected at the ith microphone, $g_i[n]$ is the corresponding preprocessed HE, $2M + 1$ is the number of samples corresponding to a duration of 4 ms, and p is the number of microphones. The effect of emphasizing the regions around the instants of significant excitation is shown in Fig. 1(d) for the HE shown in Fig. 1(c). The cross-correlation function $r_{12}[l]$ between the preprocessed HEs $g_1[n]$ and $g_2[n]$ is computed as [20]

$$r_{12}[l] = \frac{\sum_{n=z}^{N-|k|-1} g_1[n]g_2[n-l]}{\sqrt{\sum_{n=z}^{N-|k|-1} g_1^2[n] \sum_{n=z}^{N-|k|-1} g_2^2[n-l]}}, \quad l = 0, \pm 1, \pm 2, \ldots, \pm L \quad (11)$$

where $z = l, k = 0$ for $l \geq 0$, and $z = 0, k = l$ for $l < 0$, and N is the length of the segments of the HE. Here, both the vectors are normalized to unit magnitude for every sample shift before computing the cross-correlation. The cross-correlation function is computed over an interval of $2L + 1$ lags, where $2L + 1$ corresponds to an interval greater than the largest expected delay. The largest expected time-delay can be estimated from the approximate positions of the speakers and the microphones in the room. The locations of the peaks with respect to the origin (zero lag) of the cross-correlation function correspond to the time-delays between the microphone signals for all the speakers. The number of prominent peaks should correspond to the number

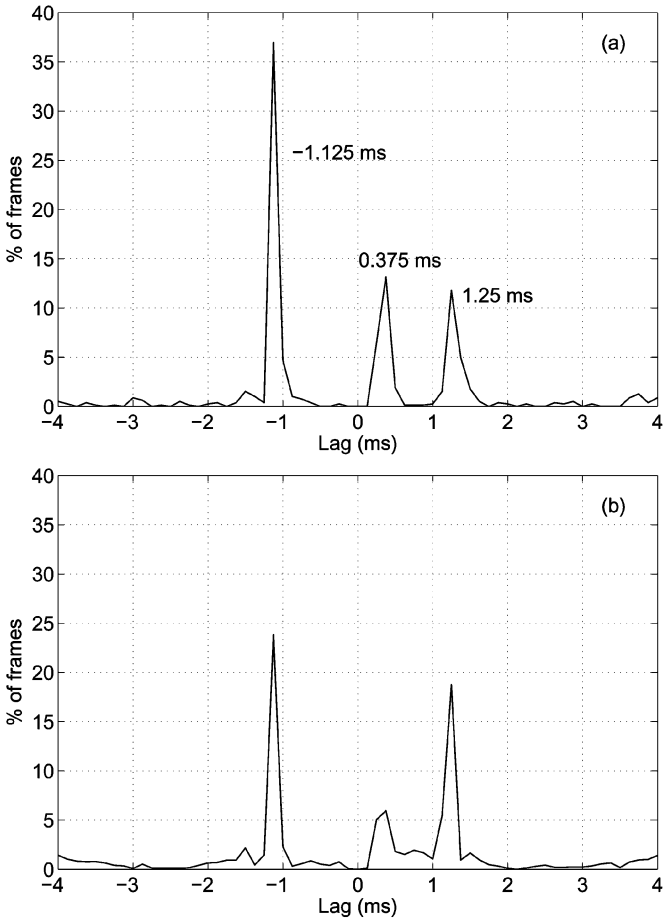

Fig. 2. Percentage of frames for each delay for a three-speakers case computed using (a) Hilbert envelope of LP residual of speech signal and (b) speech signal directly. The location the peak gives the time-delay, and the height of the peak gives an indication of prominence of the corresponding speaker.

of speakers. However, in practice, this is not always true because of the following reasons. 1) All speakers may not contribute to voiced sounds in the segments used for computing the cross-correlation function. 2) There could be spurious peaks in the cross-correlation function, which may not correspond to the time-delay due to a speaker. Hence, we rely only on the delay due to the most prominent peak in the cross-correlation function. This delay is computed from the cross-correlation function of successive frames of 50 ms duration shifted by 5 ms. Since different regions of the speech signal may provide evidence for the delays corresponding to different speakers, the number of frames corresponding to each delay is accumulated over the entire data. This helps in determining the number of speakers as well as their respective delays. Thus, by collecting the number of frames corresponding to each delay over the entire data, there will be large evidence for the delays corresponding to the individual speakers. Fig. 2(a) shows the percentage of the frames for each delay, for a recording consisting of speech from three speakers. The histogram plot obtained by using the cross-correlation of the multispeaker speech signals directly is given in Fig. 2(b) for comparison. The plots show that emphasizing the regions around the significant excitation gives better estimation of the time-delays. The locations of the peaks in the histogram indicate the time-delays due to different speakers [20]. Thus, the number of peaks in the histogram indicates the number of

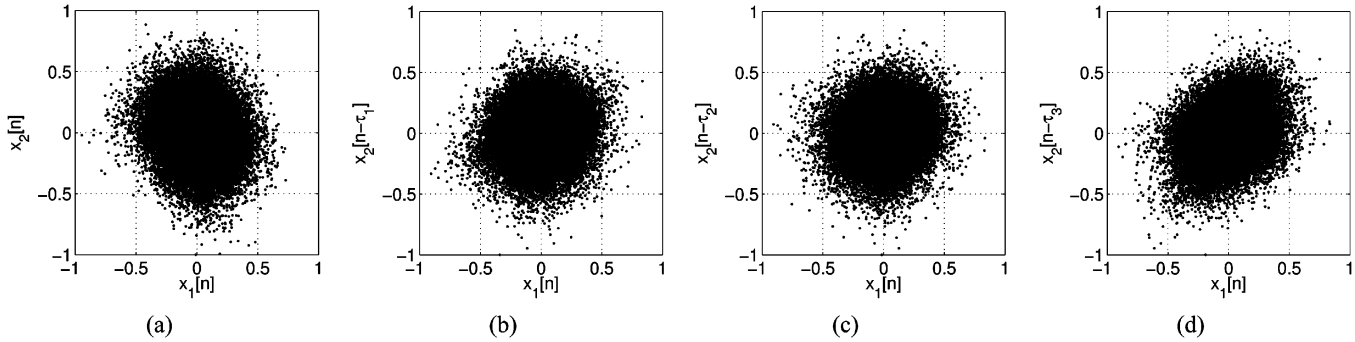

Fig. 3. (a) Scatter plot of mixed speech signals. Delay-compensated scatter plots of (b) *Spkr-1*, (c) *Spkr-2*, and (d) *Spkr-3*.

speakers, and the heights of the peaks show the relative prominence of each speaker in the conversation. The estimation of delays is based on the assumption that each speaker speaks at least for reasonable percentage of time. The minor peaks are due to random peaks in the correlation functions and occur in very small ($< 5\%$) percentage.

IV. Estimation of Mixing Parameters

The distances of a given speaker from the two microphones dictate the time-delay of arrival and also the ratio of amplitudes of the direct components of the speech signals at the microphones. The relation between the distances, the time-delay of arrival, and the amplitudes of the signals was discussed in Section II, for an ideal case. In a multispeaker case too, these relationships (due to all the speakers) are preserved, though it is not evident from the speech signals themselves. Without loss of generality, the multispeaker signals given in (4) can be rewritten as

$$x_1[n] = \sum_{j=1}^{q} s_j[n] + v_1[n] \tag{12a}$$

and

$$x_2[n] = \sum_{j=1}^{q} r_j s_j[n - \tau_j] + v_2[n] \tag{12b}$$

where $\tau_j = \tau_{1j} - \tau_{2j}$ is the differential time-delay of arrival, and $r_j = a_{2j}/a_{1j}$ is the relative attenuation of the jth source signal at the pair of microphones. The time-delay τ_j and the relative attenuation r_j can be computed from the measured distances, or can be estimated from the observed signals. A method to estimate the time-delay $\hat{\tau}_j$ is already discussed in Section III. In this section, a method to estimate the relative attenuation constants r_j from the observed multispeaker speech signals is presented. Here, the data collected from three speakers using a pair of microphones is considered for illustration.

If the corresponding sample values $(x_1[n], x_2[n])$ of multispeaker speech signals collected at *Mic-1* and *Mic-2* are plotted as a point in a 2-D plane for each of the time index $(n = 1, 2, \ldots, N)$, then the resulting *scatter plot* shows up as a cluster of points with near circular symmetry as shown in Fig. 3(a), but if the estimated time-delays $\hat{\tau}_1$, $\hat{\tau}_2$ and $\hat{\tau}_3$ of the three speakers are known, then the scatter plots of $(x_1[n], x_2[n-\hat{\tau}_1])$, $(x_1[n], x_2[n-\hat{\tau}_2])$, and $(x_1[n], x_2[n-\hat{\tau}_3])$ are expected to align in distinct directions, one for each speaker, whose slopes give an estimate of the mixing parameters $\hat{r}_1$, $\hat{r}_2$, and $\hat{r}_3$, respectively. The scatter plots of the original speech signals, and the speech signals after compensating for the three delays are shown in Fig. 3. There is lack of anticipated directions in the delay-compensated scatter plots of the individual speakers shown in Fig. 3. The reason is that, even though the signals of a given speaker are aligned, the effect of competing speakers is not suppressed. Moreover, a given speaker may not be speaking over the entire duration of the recording. Hence, it is not a good idea to use the entire data for deriving the scatter plots of the individual speakers. Oriented scatter plot can be obtained for a given speaker by selecting regions that are less affected by the competing speakers.

The characteristics of excitation around the GCIs, and the robustness of the relative spacing of the GCIs in the speech signals collected at a pair of microphones can be exploited for identifying the regions corresponding to a given speaker. Let $g_1[n]$ and $g_2[n]$ be the preprocessed HE sequences of the LP residuals of speech signals collected at *Mic-1* and *Mic-2*, respectively, as given in (10). By aligning the HEs $g_1[n]$ and $g_2[n]$ after compensating for the estimated time-delay $(\hat{\tau}_1)$ corresponding to *Spkr-1*, the GCIs corresponding to that speaker will be in coherence, whereas the GCIs corresponding to the remaining speakers will be incoherent. By considering the minimum of the HE sequences $g_1[n]$ and $g_2[n - \hat{\tau}_1]$, only the HEs around the GCIs corresponding to *Spkr-1* are retained. Note that this operation of retaining minimum ensures that the HE peaks at the GCIs of the other speakers are suppressed. The resulting sequence is the HE sequence specific to *Spkr-1*. In a similar manner, sequences that retain the HEs around the GCIs corresponding to the other speakers can be derived. Let

$$h_{sj}[n] = \min(g_1[n], g_2[n - \hat{\tau}_j]), \qquad j \in \{1, 2, 3\} \tag{13}$$

where $h_{s1}[n]$, $h_{s2}[n]$, and $h_{s3}[n]$ are the sequences in which the HEs around the GCIs corresponding to *Spkr-1*, *Spkr-2*, and *Spkr-3*, respectively, are retained. Fig. 4 shows the HEs of the LP residuals of multispeaker signals collected at the two microphones, and the individual HE sequences specific to each speaker. Fig. 4(a) and (b) consists of the peaks of the HEs around the GCIs from all the speakers, whereas the peaks of HEs for the GCIs corresponding to individual speakers only are retained in the Fig. 4(c)–(e). A weight function is derived from the HE sequence of a given speaker to highlight the regions around the GCIs of that speaker, and simultaneously reduce the effect of the competing speakers.

By setting a threshold θ on the HE sequence of a given speaker, a binary sequence is derived to locate the GCIs corresponding to that speaker. For example, the binary sequence

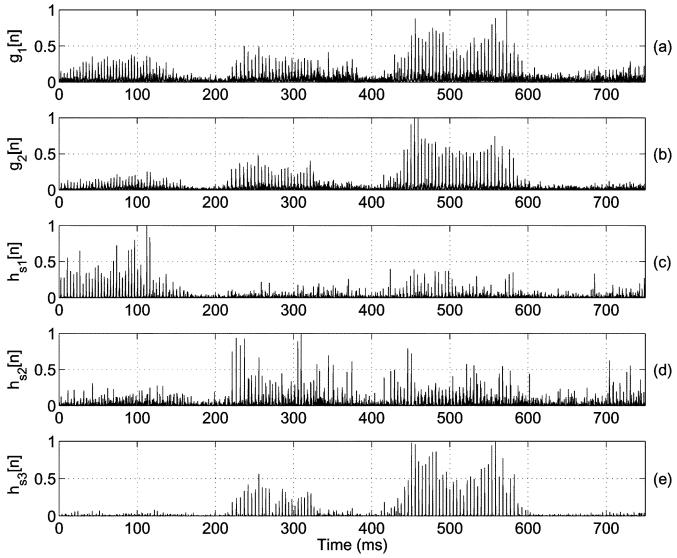

Fig. 4. HE of the LP residual of (a) *Mic-1* signal, (b) *Mic-2* signal and speaker-specific HE sequences of (c) *Spkr-1*, (d) *Spkr-2*, and (e) *Spkr-3*. Notice that the speaker-specific HE sequences are normalized between 0 and 1 for a better view.

$b_{s1}[n]$ corresponding to *Spkr-1* is derived by setting a threshold on the HE sequence $h_{s1}[n]$ of *Spkr-1*. That is

$$b_{s1}[n] = \begin{cases} 1, & \text{for } h_{s1}[n] \geq \theta_1 \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

In this study, we have chosen $\theta_j = 0.25 \max(h_{sj}[n])$, $j = 1, 2, 3$, although this not critical. Even a few samples of the target speaker that are not affected by the competing speakers are adequate for obtaining scatter plot of the target speaker.

Using the approximate locations of the GCIs, a weight function is derived to highlight the regions around the GCIs. The weight function $w_{s1}[n]$ is derived by convolving the binary sequence $b_{s1}[n]$ with a rectangular window $w_r[n]$ of size 2 ms. That is

$$\begin{aligned} w_{s1}[n] &= b_{s1}[n] * w_r[n] \\ &= \sum_{k=0}^{M-1} w_r[k] b_{s1}[n-k] \end{aligned} \tag{15}$$

where the rectangular window $w_r[n]$ is defined as

$$w_r[n] = \begin{cases} 1, & \text{for } 0 \leq n \leq M-1 \\ 0, & \text{otherwise.} \end{cases}$$

Here, M is the number of samples corresponding to 2-ms duration ($M = 16$ at 8 kHz). Fig. 5 shows 125 ms of the HE sequence $h_{s1}[n]$ of *Spkr-1*, the binary sequence $b_{s1}[n]$ derived from $h_{s1}[n]$, and the window function $w_{s1}[n]$ which retains the regions around the GCIs of *Spkr-1*. The size of the rectangular window M is not critical. A very small window results in selecting fewer samples around the GCIs, which may be inadequate to determine the slope of the scatter plot. On the other hand, a larger window size increases the effect of the competing speakers, thereby skewing the orientation of the scatter plot. The weight functions $w_{s2}[n]$ and $w_{s3}[n]$ which retain the GCIs of *Spkr-2* and *Spkr-3*, respectively, are derived in a similar way. These speaker-specific weight functions are used to

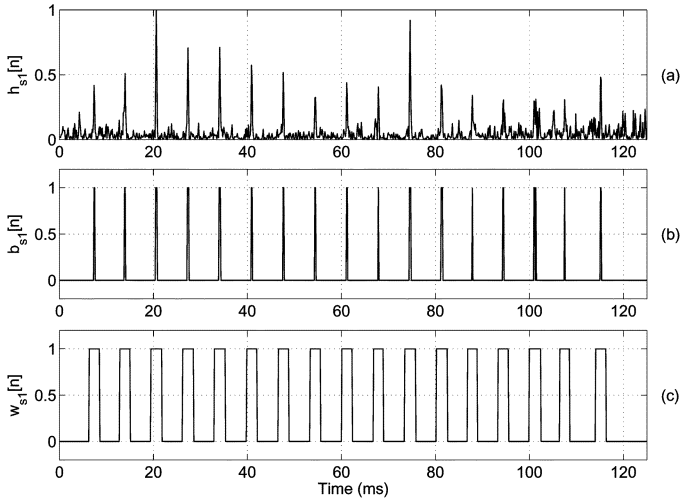

Fig. 5. (a) HE sequence specific to *Spkr-1*, (b) binary sequence indicating the locations of the GCIs, and (c) weight function derived to emphasize the regions around GCIs of *Spkr-1*.

select the regions that are relatively less affected by the competing speakers, and thereby, to obtain the oriented scatter plots for individual speakers.

The scatter plots for the individual speakers are obtained by weighting the multispeaker signals with the speaker-specific weight functions, after compensating for the respective time-delays. Scatter plot for *Spkr-1* is obtained by plotting the corresponding sample values of $(x_1[n]w_{s1}[n], x_2[n-\hat{\tau}_1]w_{s1}[n])$ in a 2-D plane, where $\hat{\tau}_1$ is the estimated time-delay corresponding to *Spkr-1*, and w_{s1} is the weight function specific to *Spkr-1*. The delay-compensated scatter plots of the individual speakers after multiplying with the speaker-specific weight functions are shown in Fig. 6. The scatter plot of the original multispeaker signals is also shown for comparison. The scatter plots of the individual speakers are oriented in distinct anticipated directions.

The slope of the scatter plot of the jth speaker gives an estimate of the relative attenuation $\hat{r}_j$. In an ideal situation, the scatter plot of a given speaker should be a straight line, but in practice, the scatter plot deviates from the straight line because of noise and reverberation components, and the effect of the competing speakers. The slope of the first principal component of the data points is assumed to be the slope of the scatter plot. The solid line in each figure shows the direction of the first principal component of the data points. The estimated slope of the scatter plot gives the relative attenuation $\hat{r_j} = e_{2j}/e_{1j}$ of the jth speaker at the microphones, where e_{1j} and e_{2j} are the elements of the eigenvector corresponding to the maximum eigenvalue of the data points of the jth speaker. The distances $\hat{d}_{1j}$ and $\hat{d}_{2j}$ of the jth speaker from the two microphones can be computed from the estimated time-delay $\hat{\tau}_j$ and the relative attenuation $\hat{r}_j$ as shown in (8).

V. Experimental Results

The proposed method was verified on multispeaker data collected from two speakers (well-determined case) and three speakers (underdetermined case) using a pair of microphones. Speech is collected from spatially distributed speakers speaking simultaneously in a laboratory environment having some reverberation. Fig. 7 shows the recording scenario. The multispeaker

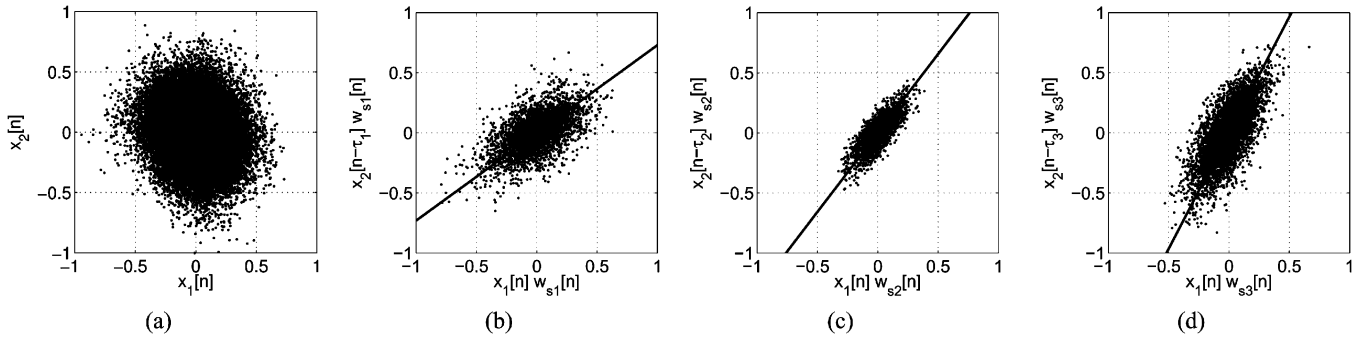

Fig. 6. (a) Scatter plot of mixed speech signals. Delay-compensated scatter plots of (b) *Spkr-1*, (c) *Spkr-2*, and (d) *Spkr-3* after weighing the multispeaker signals with the weight functions specific to individual speakers.

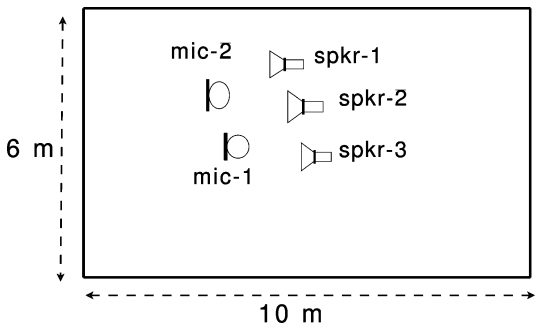

Fig. 7. Recording scenario for two-microphones and three-speakers case. Height of the room is about 3 m. Distance between the microphones is 0.6 m. Microphones are at a height of 1.7 m from ground. All the speakers are also approximately at the same level.

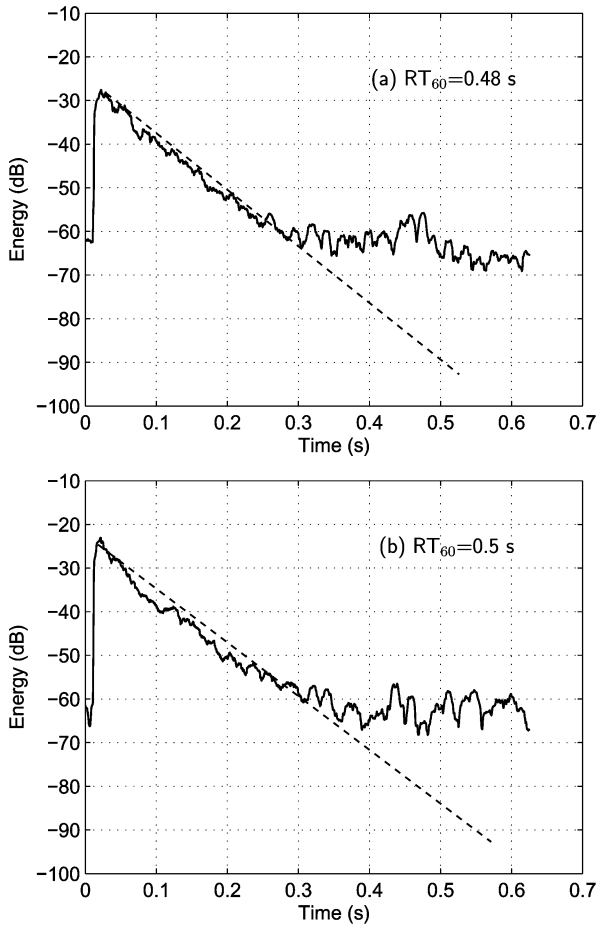

Fig. 8. Estimation of reverberation time from room response. (a) At *Mic-1*. (b) At *Mic-2*.

data was collected simultaneously using a pair microphones separated by about 0.6 m. The speakers were at different locations, around the microphones, at an average distance of 1 m from the microphones. All the speakers were stationary, and spoke simultaneously during the entire duration of recording, resulting in significant overlap. There are five recordings for two-speakers case, and five recordings for three-speakers case. The speakers were chosen from a pool of seven speakers. The duration of each recording was about 10 s. The speech signals were sampled at 8 kHz. During each recording, the distances of the speakers from both the microphones were measured. The reverberation time is estimated (as shown in Fig. 8) by computing the decay of energy of the response signal at each of the microphone positions for an impulse-like excitation from a loudspeaker. The average reverberation time of the room was about 0.5 s.

The multispeaker signals were processed using the proposed method to estimate the mixing parameters. A 10th-order LP analysis (autocorrelation method) was used for deriving the LP residual. The cross-correlation function of the HEs of the LP residuals of the multispeaker signals is used to estimate the time-delays as explained in Section III. As mentioned earlier in Section I, and as illustrated in Section III, the advantage of using the HE of the LP residual is that the HE shows peaks around the GCIs even for multiple speakers case. The cross-correlation function of the HE of the LP residual for the multispeaker signals shows prominent peaks than in the crosscorrelation function computed directly from the speech signal, as in the case of single source case discussed in [18]. The delay compensated scatter plots of the individual speakers are obtained by weighting the multispeaker signals with the speaker-specific weight functions, as explained in Section IV. An estimate of the relative attenuation $\hat{r}_j$ is obtained from the slope of the scatter plot of the jth speaker. The reference value r_j of the relative attenuation is computed from the measured distances of the microphones from the speakers using (6).

Table I presents a comparison of the reference values r_j of the mixing parameters with the estimated values $\hat{r}_j$, for the two-speakers case. The table lists the measured distances (d_{1j}, d_{2j}) of the microphones from the jth speaker, the actual time-delay of arrival τ_j, and the estimated time-delay of arrival $\hat{\tau}_j$. A similar study for the three-speakers case is presented in Table II. The reference values and the estimated values of the mixing parameters are in close agreement, validating the hypothesis outlined

TABLE I
COMPARISON OF ESTIMATED VALUES ($\hat{r}_j$) OF THE MIXING PARAMETERS WITH THE REFERENCE VALUES (r_j) FOR TWO-SPEAKERS CASE. (d_{1j}, d_{2j}) ARE THE DISTANCES OF MICROPHONES FROM THE jTH SPEAKER, τ_j IS THE ACTUAL TIME-DELAY OF ARRIVAL, AND $\hat{\tau}_j$ IS THE ESTIMATED TIME-DELAY OF ARRIVAL

S. No.	*Spkr-j*	d_{1j} (m)	d_{2j} (m)	τ (ms)	$\hat{\tau}_j$ (ms)	r_j	$\hat{r}_j$
1	*Spkr-1*	0.52	0.77	-0.78	-0.875	0.67	0.67
	Spkr-2	0.88	0.54	1.06	1.0	1.63	1.53
2	*Spkr-1*	0.56	0.88	-1.0	-1.0	0.64	0.65
	Spkr-2	0.90	0.47	1.34	1.375	1.91	2.04
3	*Spkr-1*	0.52	0.95	-1.34	-1.125	0.55	0.56
	Spkr-2	0.87	0.56	0.96	0.875	1.55	1.51
4	*Spkr-1*	0.5	0.96	-1.43	-1.375	0.52	0.51
	Spkr-2	0.81	0.41	1.25	1.0	1.97	1.87
5	*Spkr-1*	0.76	1.07	-0.97	-1.0	0.71	0.72
	Spkr-2	0.85	0.69	0.5	0.5	1.23	1.26

TABLE II
COMPARISON OF ESTIMATED VALUES ($\hat{r}_j$) OF THE MIXING PARAMETERS WITH THE REFERENCE VALUES (r_j) FOR THREE-SPEAKERS CASE. (d_{1j}, d_{2j}) ARE THE DISTANCES OF MICROPHONES FROM THE jTH SPEAKER, τ_j IS THE ACTUAL TIME-DELAY OF ARRIVAL, AND $\hat{\tau}_j$ IS THE ESTIMATED TIME-DELAY OF ARRIVAL

S. No.	*Spkr-j*	d_{1j} (m)	d_{2j} (m)	τ (ms)	$\hat{\tau}_j$ (ms)	r_j	$\hat{r}_j$
1	*Spkr-1*	0.49	0.97	-1.4	-1.375	0.50	0.50
	Spkr-2	0.74	0.98	-0.7	-0.625	0.75	0.70
	Spkr-3	0.76	0.46	0.88	0.875	1.65	1.48
2	*Spkr-1*	0.75	1.2 0	-1.32	-1.25	0.62	0.62
	Spkr-2	0.52	0.74	-0.64	-0.5	0.70	0.70
	Spkr-3	0.99	0.61	1.2	1.125	1.62	1.77
3	*Spkr-1*	0.78	0.95	-0.5	-0.5	0.82	0.83
	Spkr-2	0.98	0.83	0.44	0.375	1.18	1.18
	Spkr-3	0.77	0.44	0.97	0.875	1.75	1.69
4	*Spkr-1*	0.36	0.84	-1.41	-1.375	0.43	0.47
	Spkr-2	0.90	0.75	0.45	0.5	1.2	1.40
	Spkr-3	0.91	0.42	1.44	1.375	2.17	2.36
5	*Spkr-1*	0.64	0.97	-0.97	-1.125	0.66	0.57
	Spkr-2	0.91	0.78	0.38	0.375	1.16	1.18
	Spkr-3	0.96	0.53	1.26	1.25	1.81	1.98

in Section II. The results show the effectiveness of the proposed method in estimating the mixing parameters. Minor deviations between the actual and the estimated values can be attributed to the following. 1) Errors in measuring the actual distances, 2) the resolution that can be obtained for a sampling frequency of 8 kHz, 3) movement of the speakers during recording session, and 4) differences in the amplitude responses of the microphones, even after compensation. Note that microphone calibration is not required for determining the slope information from the scatter plots. This is needed only to obtain the actual amplitude ratios for comparison with the amplitude ratios derived by processing the multispeaker data.

VI. COMMENT ON SPEAKER SEPARATION

Separation of source signals from mixed signals is a challenging problem, as the problem does not conform to any standard or known formulations. Though the mixing parameters of (12) are known, the source signals cannot be separated completely because of the time-delays involved. However, it is possible to enhance the target speaker by reducing the effect due to the competing speakers. In this paper, we propose a method to enhance the target speakers in the two-speakers case. In this method, the knowledge of the estimated time-delays and the mixing parameters is used to subtract an estimate of the speech signal of the undesired speaker from the mixed signal. For the two-speakers case, the mixed signals given in (12) can be simplified to

$$x_1[n] = s_1[n] + s_2[n] \tag{16a}$$

and

$$x_2[n] = r_1 s_1[n - \tau_1] + r_2 s_2[n - \tau_2] \tag{16b}$$

where the noise component is ignored. By expressing the above equations in the frequency domain and inverting (complex) mixing matrix in the frequency domain, it is possible to obtain estimates of the component signals corresponding to *Spkr-1* and

TABLE III
COMPARISON OF ESTIMATED VALUES ($\hat{r}_j$) OF THE MIXING PARAMETERS WITH THE REFERENCE VALUES (r_j) FOR TWO-SPEAKERS CASE, ON SYNTHETIC ANECHOIC MIXTURES. (d_{1j}, d_{2j}) ARE THE ASSUMED DISTANCES OF MICROPHONES FROM THE jTH SPEAKER, τ_j IS THE ACTUAL TIME-DELAY OF ARRIVAL AND $\hat{\tau}_j$ IS THE ESTIMATED TIME-DELAY OF ARRIVAL

S. No.	*Spkr-j*	d_{1j} (m)	d_{2j} (m)	Reference		Proposed		Duet	
				τ (ms)	r_j	$\hat{\tau}_j$ (ms)	$\hat{r}_j$	$\hat{\tau}_j$ (ms)	$\hat{r}_j$
1	*Spkr-1*(F)	0.56	1.1	-1.63	0.51	-1.625	0.51	-1.625	0.51
	Spkr-2(M)	0.92	0.73	0.66	1.30	0.625	1.24	0.625	1.31
2	*Spkr-1*(F)	0.98	0.80	0.55	1.23	0.625	1.27	0.625	1.21
	Spkr-2(F)	0.75	0.43	0.97	1.73	1	1.74	1	1.72
3	*Spkr-1*(M)	0.7	1.2	-1.515	0.583	-1.5	0.583	-1.5	0.588
	Spkr-1(M)	0.99	0.61	1.151	1.623	1.125	1.613	1.125	1.61

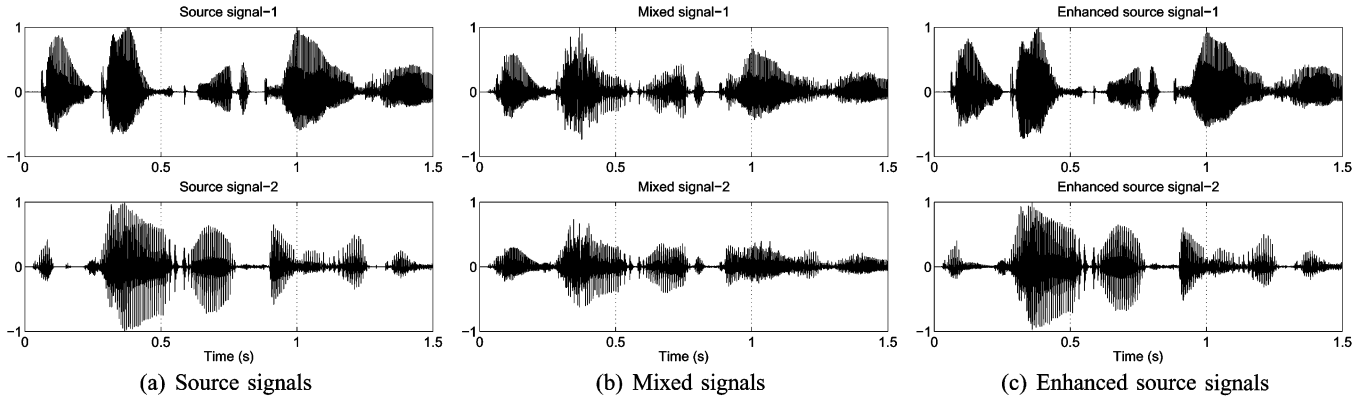

Fig. 9. Illustration of source enhancement capability of the proposed algorithm. (a) Original source signals. (b) Mixed signals. (c) Enhanced source signals obtained from the mixed signals in (b) using the proposed method.

Spkr-2. However, for the two-speaker case, these components can be obtained more directly (in the time domain) as follows:

$$\hat{s}_1[n] = x_1[n] - \frac{1}{\hat{r}_2} x_2[n + \hat{\tau}_2] \tag{17a}$$

and

$$\hat{s}_2[n] = x_2[n] - \hat{r}_1 x_1[n - \hat{\tau}_1] \tag{17b}$$

where $\hat{r_1}$, $\hat{r_2}$, $\hat{\tau_1}$, and $\hat{\tau_2}$ are estimated using the proposed analysis. In the next section objective evaluation of the proposed method is given for a two-speaker, two-microphone case.

Extending this method to the three-speakers case requires inversion of the 2×3 mixing matrix which cannot be solved uniquely. An underdetermined system with a known mixing matrix can be solved using l_1 norm minimization solutions. The l_1 norm solutions are used to avoid computing the pseudo inverse of the rectangular mixing matrix. A constrained l_1 norm minimization solution for the underdetermined source separation from a known mixing matrix is proposed in [11]. The mixing parameters estimated using the proposed method can be used along with the l_1 norm minimization techniques to enhance the sources in an underdetermined case.

VII. OBJECTIVE EVALUATION

A. Evaluation on Synthetic Anechoic Mixtures

The source enhancement capability of the proposed algorithm is evaluated on synthetic mixing data. The synthetic mixing data is used because the performance evaluation requires the original sources to be known. The synthetic data is prepared using data from four speakers (two female and two male) taken from the TIMIT database [25]. The source signals are mixed assuming arbitrary distances between the sources and the microphones. The assumed distances of the two sources from the pair of microphones are listed in Table III. The proposed analysis is performed on the mixed signals to estimate the time-delays and

TABLE IV
PERFORMANCE COMPARISON OF THE PROPOSED SOURCE ENHANCEMENT METHOD WITH THE DUET ALGORITHM ON SYNTHETIC ANECHOIC MIXTURES

		Mixed Signals			Proposed method			DUET [9], [26]		
S. No.	*Spkr*	SDR-in	SIR-in	SAR-in	SDR-out	SIR-out	SAR-out	SDR-out	SIR-out	SAR-out
1	*Spkr-1*(F)	6.47	6.47	82.27	26.58	26.61	50.44	12.76	26.82	12.94
	Spkr-2(M)	1.66	1.66	79.49	21.51	22.94	27.05	11.20	28.43	11.29
2	*Spkr-1*(F)	0.04	0.04	80.10	15.10	15.11	69.44	9.11	20.53	9.47
	Spkr-2(F)	2.76	2.76	83.51	19.02	19.02	77.68	10.64	21.15	11.08
3	*Spkr-1*(M)	3.48	3.48	81.67	24.24	24.58	35.44	11.01	28.65	11.09
	Spkr-2(M)	5.36	5.36	81.83	26.74	29.18	30.42	12.25	28.53	12.36

TABLE V
COMPARISON OF ESTIMATED VALUES $(\hat{r}_j)$ OF THE MIXING PARAMETERS WITH THE REFERENCE VALUES (r_j) FOR TWO-SPEAKERS CASE, ON REAL ROOM RECORDINGS. (d_{1j}, d_{2j}) ARE THE MEASURED DISTANCES OF MICROPHONES FROM THE jTH SPEAKER, τ_j IS THE ACTUAL TIME-DELAY OF ARRIVAL AND $\hat{\tau}_j$ IS THE ESTIMATED TIME-DELAY OF ARRIVAL

				Reference		Proposed		Duet	
S. No.	*Spkr-j*	d_{1j} (m)	d_{2j} (m)	τ (ms)	r_j	$\hat{\tau}_j$ (ms)	$\hat{r}_j$	$\hat{\tau}_j$ (ms)	$\hat{r}_j$
1	*Spkr-1*	0.39	0.64	-0.74	0.61	-0.75	0.65	-0.625	0.75
	Spkr-2	0.81	0.55	0.76	1.47	0.75	1.48	0.875	1.46
2	*Spkr-1*	0.44	0.61	-0.5	0.72	-0.5	0.75	-0.5	1.41
	Spkr-2	0.70	0.39	0.91	1.79	0.875	1.75	0.875	1.69

the mixing parameters. A comparison of the estimated relative attenuation $(\hat{r}_j)$ and the estimated differential time-delay $(\hat{\tau}_j)$ with the actual values is provided in Table III. The values are compared with the estimates obtained from the DUET algorithm [9]. A Matlab code given in [26] is used for evaluating the DUET algorithm proposed in [9]. Since in the current scenario, the distance between the microphones is 60 cm, the *big delay DUET* [26] implemented through *histogram tiling* is used in evaluation. As can be seen, for this mixtures, the estimates obtained by both methods are nearly equal to actual parameters used in mixing.

The sources are enhanced from the mixed signals using the estimated mixing parameters and the time-delays as given in (17). Fig. 9 shows segments of original source signals, mixed signals, and the enhanced source signals retrieved from the mixed signals using the proposed method. The enhanced source signals shown in Fig. 9(c) are similar to the original source signals shown in Fig. 9(a), indicating the potential of the proposed algorithm.

The performance of the proposed source enhancement method is evaluated using the measures proposed in [27]. In this evaluation, the estimated source signal is decomposed into a source part (s_{target}), plus error terms corresponding to the interference (e_{interf}) and algorithmic artifacts (e_{artif}). These decomposed components are used to define the following performance measures: The source-to-distortion ratio (SDR)

$$\text{SDR} = 10\log_{10}\frac{||s_{\text{target}}||^2}{||e_{\text{interf}} + e_{\text{artif}}||^2}\ \text{dB} \tag{18}$$

the source-to-interference ratio (SIR)

$$\text{SIR} = 10\log_{10}\frac{||s_{\text{target}}||^2}{||e_{\text{interf}}||^2}\ \text{dB} \tag{19}$$

and the source-to-artifacts ratio (SAR)

$$\text{SAR} = 10\log_{10}\frac{||s_{\text{target}} + e_{\text{interf}}||^2}{||e_{\text{artif}}||^2}\ \text{dB}. \tag{20}$$

The performance of the proposed source enhancement method is evaluated in terms of SDR, SIR, and SAR using the actual sources as references. The Matlab code for performance measurement of BSS algorithms available at [28] is used in this evaluation.

The performance measures SDR, SIR, and SAR obtained on the proposed method for three different cases (listed in Table III) are given in Table IV. The input SDR (SDR-in) is the SDR of a source signal in the mixed signals. The output SDR (SDR-out) is the SDR of a source signal in the enhanced source signal. The performance of the proposed method is compared with the DUET algorithm [9], [26], which exploits the sparsity of the source signals in the time–frequency domain and estimates binary masks to separate them. The proposed method introduces significantly less overall distortion compared to the DUET algorithm. However, the interference measure given

TABLE VI
PERFORMANCE COMPARISON OF THE PROPOSED SOURCE ENHANCEMENT METHOD WITH THE DUET ALGORITHM ON REAL ROOM RECORDINGS

		Mixed Signals			Proposed method			DUET [9], [26]		
S. No.	*Spkr*	SDR-in	SIR-in	SAR-in	SDR-out	SIR-out	SAR-out	SDR-out	SIR-out	SAR-out
1	*Spkr-1*	-6.86	-1.68	-1.37	-2.95	9.43	-2.22	-3.52	8.20	-2.60
	Spkr-2	-4.75	1.66	-1.37	7.65	14.37	8.84	6.45	15.43	7.15
2	*Spkr-1*	-7.93	-0.9	-3.5	-3.51	14.07	-3.26	-4.25	7.08	-3.13
	Spkr-2	-6.97	0.7	-3.49	4.07	14.51	4.64	-2.37	4.30	0.05

by SIR is slightly superior in DUET algorithm. Note that the proposed method introduces significantly less artefacts in comparison with DUET algorithm as indicated by the values of SAR in the Table IV. The enhanced signals are available for listening at http://speech.cs.iitm.ernet.in/~sriram/bssResults/.

B. Evaluation on Real Room Recordings (i.e., Echoic Mixtures)

The real room recordings are done in a typical laboratory environment with room reverberation characteristics as shown in Fig. 8. The two microphones are separated by a distance of 60 cm. Speech from two speakers is recorded one after another using two microphones, as suggested in [29]. The measured distances of each speaker from the pair of microphones are listed in Table V. Speech signals from the pair of speakers are added to obtain the mixed signals. The individual speaker recordings are used as reference signals to evaluate the performance measures discussed in Section VII-A. Since the reference signals are recorded in a live room, the signals are echoic in nature.

The proposed method is applied on the mixed signals to estimate the mixing parameters. A comparison is made on the estimated differential time-delays and the relative attenuations with the reference values obtained from the measured distances as given in Table V. The estimated values obtained with the DUET algorithm [9] are also given in the table. Since these are echoic mixtures, the estimates of the differential time-delay and relative attenuation are not accurate in the case of DUET algorithm, whereas the proposed method gives accurate values of these parameters. This is reflected in the results of the source separation also as can be seen in Table VI, where the speech signals from individual speakers are enhanced from the mixed signals using using the estimated mixing parameters. We notice from Table VI that most of the values of the performance measures are better for the proposed method compared to the values obtained using the DUET algorithm.

VIII. SUMMARY AND CONCLUSION

In this paper, we have addressed the issues in processing multispeaker data collected in a real environment. While the problem should be addressed as a convolutive BSS problem, we have shown that the solution to the problem can be explored in a different way, if we can exploit the speech-specific characteristics for processing the multispeaker data. In particular, the sequence of impulse-like excitation in speech production enables us to reduce the problem to 1) estimating the time-delays for estimating the mixing parameters, and 2) deriving the weight function for enhancing the speech of one speaker over others in the mixed signals. The proposed method was tested with the multispeaker data collected in a real live room environment. The performance of the proposed method seem to be significantly better than some of the available methods as demonstrated by the objective evaluation by objective evaluation. It is interesting to note that for normal placement (inside the room) of speakers and microphones, the effects of environment noise and reverberation are not significant. This is because most of the time there will be significant direct sound component at the microphones. This can be assured, in general, by providing several spatially distributed microphones, so that there will be some microphones which will have significant direct sound component, at least for a subset of the speakers. The reverberant sound can be assumed to be diffused, and can be merged with noise in normal live rooms, where the average reverberation time is usually about 0.5 s, and also where there are no strong echoes.

The proposed approach for processing multispeaker data does not involve the difficult problem of determination of the impulse response of the room for convolutive BSS. The approach also does not require a complete characterization of the mixing situation. It is also obvious that one cannot obtain complete separation of the mixed signals in a convolutive BSS problem. One can only achieve some enhancement of a given speaker over other speakers from a mixed signal. By refining the time-delay estimation and by combining the information from more than two microphones in a multimicrophone situation, one may be able to improve the results of enhancement.

REFERENCES

[1] O. M. M. Mitchell, C. A. Ross, and G. H. Yates, "Signal processing for a cocktail party effect," *J. Acoust. Soc. Amer.*, vol. 50, no. 2, pp. 656–660, 1971.

[2] J.-F. Cardoso, "Blind signal separation: Statistical principles," *Proc. IEEE*, vol. 86, no. 10, pp. 2009–2025, Oct. 1998.

[3] S. Amari, S. C. Douglas, A. Chichocki, and H. H. Yang, "Multichannel blind deconvolution and equalization using the natural gradient," in *Proc. 1st IEEE Workshop Signal Process. Adv. Wireless Commun.*, Paris, France, Apr. 1997, pp. 101–102.
[4] S. C. Douglas, H. Sawada, and S. Makino, "Natural gradient multichannel blind deconvolution and speech separation using causal FIR filters," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 92–104, Jan. 2005.
[5] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *NeuroComputing*, vol. 22, no. 1-3, pp. 21–34, Nov. 1998.
[6] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Frequency-domain blind source separation," in *Speech Enhancement*, J. Benesty, S. Makino, and J. Chen, Eds. Berlin, Germany: Springer, 2005, pp. 299–327.
[7] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. Hoboken, NJ: Wiley, 2001.
[8] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 530–538, Sep. 2004.
[9] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time–frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, Jul. 2004.
[10] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors," *Signal Process.*, vol. 87, no. 8, pp. 1833–1847, 2007.
[11] S. Winter, W. Kellermann, H. Sawada, and S. Makino, "MAP-based underdetermined blind source separation of convolutive mixtures by hierarchical clustering and l_1-norm minimization," *EURASIP J. Advances in Signal Processing*, vol. 2007, pp. 12–12, 2007.
[12] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust.*, New Paltz, NY, Oct. 2007, pp. 139–142.
[13] D. Smith, J. Lukasiak, and I. S. Burnett, "An analysis of the limitations of blind signal separation application with speech," *Signal Process.*, vol. 86, no. 2, pp. 353–359, Feb. 2006.
[14] Q. Lv and X.-D. Zhang, "A unified method for blind separation of sparse sources with unknown source number," *IEEE Signal Process. Lett.*, vol. 13, no. 1, pp. 49–51, Jan. 2006.
[15] L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 320–327, May 2000.
[16] 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 Process. Lett.*, vol. 6, no. 4, pp. 87–91, Apr. 1999.
[17] Y. Li, S. Amari, A. Cichocki, D. W. C. Ho, and S. Xie, "Underdetermined blind source separation based on sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 2, pp. 423–437, Feb. 2006.
[18] V. C. Raykar, B. Yegnanarayana, S. R. M. Prasanna, and R. Duraiswami, "Speaker localization using excitation source information in speech," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 751–761, Sep. 2005.
[19] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of timedelay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.
[20] R. K. Swamy, K. S. R. Murty, and B. Yegnanarayana, "Determining number of speakers from multispeaker speech signals using excitation source information," *IEEE Signal Process. Lett.*, vol. 14, no. 7, pp. 481–484, Jul. 2007.
[21] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.
[22] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 4, pp. 309–319, Aug. 1979.
[23] A. V. Oppenheim, R. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Upper Saddle River, NJ: Prentice-Hall, 2000.
[24] B. Yegnanarayana, S. R. M. Prasanna, R. Duraiswami, and D. Zotkin, "Processing of reverberent speech for time-delay estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 1110–1118, Nov. 2005.
[25] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The darpa speech recognition research database: Specifications and status," in *Proc. DARPA Workshop Speech Recognition*, Feb. 1986, pp. 93–99.
[26] S. Rickard, "The duet blind source separation algorithm," in *Blind Speech Separation*, S. Makino, T. W. Lee, and H. Sawada, Eds. Dordrecht, The Netherlands: Springer, 2007, ch. 8, pp. 217–241.
[27] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[28] C. Fevotte, R. Gribonval, and E. V. Vincent, BSS EVAL Toolbox. [Online]. Available: http://www.irisa.fr/metiss/bss_eval/
[29] D. Schobben, K. Torkkola, and P. Smaragdis, "Evaluation of blind signal separation methods," in *Proc. ICA BSS*, Aussois, France, Jan. 1999.

B. Yegnanarayana (M'78–SM'84) received the B.Sc. degree from Andhra University, Waltair, India, in 1961, and the B.E., M.E., and Ph.D. degrees in electrical communication engineering from the Indian Institute of Science (IISc) Bangalore, India, in 1964, 1966, and 1974, respectively.

He is a Professor and Microsoft Chair at the International Institute of Information Technology (IIIT), Hyderabad. Prior to joining IIIT, he was a Professor in the Department of Computer Science and Engineering at the Indian Institute of Technology (IIT), Madras, India, from 1980 to 2006. He was the Chairman of the Department from 1985 to 1989. He was a Visiting Associate Professor of computer science at Carnegie-Mellon University, Pittsburgh, PA, from 1977 to 1980. He was a member of the faculty at the Indian Institute of Science (IISc), Bangalore, from 1966 to 1978. He has supervised 32 M.S. theses and 24 Ph.D. dissertations. His research interests are in signal processing, speech, image processing, and neural networks. He has published over 300 papers in these areas in IEEE journals and other international journals, and in the proceedings of national and international conferences. He is also the author of the book *Artificial Neural Networks* (Prentice-Hall of India, 1999).

Dr. Yegnanarayana was an Associate Editor for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING from 2003 to 2006. He is a Fellow of the Indian National Academy of Engineering, a Fellow of the Indian National Science Academy, and a Fellow of the Indian Academy of Sciences. He was the recipient of the Third IETE Prof. S. V. C. Aiya Memorial Award in 1996. He received the Prof. S. N. Mitra memorial Award for the year 2006 from the Indian National Academy of Engineering.

R. Kumara Swamy (M'07) received the B.E. and M.E. degrees from Bangalore University, Karnataka, India, in 1991 and 1997, respectively, and the Ph.D degree in computer science and engineering from the Indian Institute of Technology Madras, Chennai, India, in 2007.

He is a Professor at Siddaganga Institute of Technology, Tumkur, Karnataka, India. His research interests are in signal processing, speech processing, source separation, and pattern recognition.

K. Sri Rama Murty received the B.Tech degree in electronics and communications engineering from Jawaharlal Nehru Technological University (JNTU), Hyderabad, India, in 2002. He is currently pursuing the Ph.D. degree at the Indian Institute of Technology (IIT) Madras, Chennai, India.

His research interests include signal processing, speech analysis, blind source separation, and pattern recognition.