Determining Number of Speakers From Multispeaker Speech Signals Using Excitation Source Information

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

***Abstract*—In this letter, we address the issue of determining the number of speakers from multispeaker speech signals collected simultaneously using a pair of spatially separated microphones. The spatial separation of the microphones results in time delay of arrival of speech signals from a given speaker. The differences in the time delays for different speakers are exploited to determine the number of speakers from the multispeaker signals. The key idea is that for a given speaker, the relative spacings of the instants of significant excitation of the vocal tract system remain unchanged in the direct components of the speech signals at the two microphones. The time delays can be estimated from the cross-correlation of the Hilbert envelopes of the linear prediction residuals of the multispeaker signals collected at the two microphones.**

***Index Terms*—Excitation source, Hilbert envelope, linear prediction residual, multispeaker signals, time-delay estimation, underdetermined case.**

I. Introduction

ONE of the important problems in signal processing is to estimate the number of sources from multisensor data. In the case of multispeaker data, the problem is to determine the number of speakers, and then localize and track the speakers from the signals collected using a number of spatially distributed microphones. It is also necessary to separate speech of the individual speakers from the multispeaker signals. Solutions to these problems are needed, especially for signals collected in a practical environment, such as in a room with background noise and reverberation. Several approaches, mostly theoretical, were proposed in the literature for the detection of the number of sources whose mixed signals are collected by an array of passive sensors. One approach is based on the eigenvalues of the covariance matrix of the observation vector [1], [2]. A nested sequence of hypothesis tests was proposed to implement this approach. But this method uses subjective judgement for deciding the threshold level of the likelihood statistic ratio for accepting a hypothesis. To avoid the use of subjective thresholds, Wax and Kailath suggested the use of a minimum description length (MDL) criterion to estimate the number of sources [3]. The MDL estimator can be interpreted as a test for determining the multiplicity of the smallest eigenvalues [4]. Methods based on multiplicity of the smallest eigenvalue are not robust, if there are deviations from the assumed model of the additive noise process. Robustness is improved by exploiting some type of prior knowledge [5]. But the methods based on prior knowledge, e.g., array steering vectors, have high computational complexity requiring multidimensional numerical search. Robust estimators for specific types of deviations from the assumed model are reported in the literature. In particular, the situation when the sensor noise levels are spatially inhomogeneous is considered in [5] to estimate the number of sources by using an information theoretic criterion.

Most of the methods proposed so far assume the following model for the mixed signal vector. The model, consisting of N observations collected at p sensors from q sources, is given by

$$\mathbf{x}[n] = \mathbf{A}\mathbf{s}[n] + \mathbf{v}[n], \quad n = 1, 2, \ldots, N \tag{1}$$

where

$$\begin{aligned}\mathbf{x}[n] &= [x_1[n]\ \ x_2[n] \cdots x_i[n] \cdots x_p[n]]^{\mathrm{T}} \\ \mathbf{s}[n] &= [s_1[n]\ \ s_2[n] \cdots s_j[n] \cdots s_q[n]]^{\mathrm{T}} \\ \mathbf{v}[n] &= [v_1[n]\ \ v_2[n] \cdots v_i[n] \cdots v_p[n]]^{\mathrm{T}} \\ \mathbf{A} &= [a_{ij}]_{p \times q}.\end{aligned}$$

Here, $x_i[n]$ is the mixed signal at the i^{th} sensor, $s_j[n]$ is the signal generated from the j^{th} source, $v_i[n]$ is the additive noise at the i^{th} sensor, $\mathbf{A}$ is the mixing matrix, N is the number of observations, and the superscript T indicates the transpose operation. The j^{th} column vector of the mixing matrix $\mathbf{A}([a_{1j}\ a_{2j}\ \ldots\ a_{pj}]^{\mathrm{T}})$ gives the array response associated with the j^{th} source signal. The i^{th} row vector of the mixing matrix $\mathbf{A}([a_{i1}\ a_{i2}\ \ldots\ a_{iq}])$ gives the mixing weights for the source signals collected at the i^{th} sensor. For determining the number of sources, three cases are considered: overdetermined case $(p > q)$, well-determined case $(p = q)$, and underdetermined case $(p < q)$. For an overdetermined case $(p > q)$, the number of source signals is determined from the multiplicity of the smallest eigenvalue of the covariance matrix of the observation vector $\mathbf{x}[n]$ [3], [5]. The well-determined case $(p = q)$ is commonly addressed using the independent component analysis (ICA) formulation [6], [7]. For an underdetermined case $(p < q)$, the number of sources can be determined by assuming sparseness of the sources and a constant mixing matrix with full column rank [8]. It is important to note that most of the studies on estimating the number of sources use artificially generated mixed signals according to the model in (1). Practical signals such as multispeaker signals collected from a number of speakers speaking simultaneously have much more variability due to noise and reverberation, besides delay and decay of

Manuscript received August 14, 2006; revised November 12, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Patrick A. Naylor.

R. K. Swamy and K. S. R. Murty are with the Department of Computer Science and Engineering, Indian Institute of Technology-Madras, Chennai 600 036, India (e-mail: kswamy@cs.iitm.ernet.in; ksrm@cs.iitm.ernet.in).

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

Digital Object Identifier 10.1109/LSP.2006.891333

the direct sound due to distance of the microphone from the speaker.

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 speech signals (between every pair of microphones) for 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 of the multispeaker signals give the time delays due to all the speakers at the pair of microphones. However, in general the cross-correlation function of the multispeaker signals does not show unambiguous prominent peaks at the time delays. This is mainly because of the damped sinusoidal 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. In particular, the speech signal exhibits relatively high signal-to-noise ratio (SNR) and high signal-to-reverberation ratio (SRR), in the vicinity of time instants of significant excitations of the vocal tract. In Section II, we discuss preprocessing of multispeaker signals to emphasize the regions of high SNR and SRR. In Section III a method for estimating the time delays, and thereby determining the number of speakers is explained. Experimental results for determining the number of speakers from multispeaker signals are presented in Section IV. Section V gives a summary and the conclusions of these studies.

II. Preprocessing of Multispeaker Speech Signal

During the production of voiced speech, the vocal tract system is excited by a quasi-periodic sequence of impulse-like excitations [9]. These significant excitations occur at the instants of glottal closure (GCI) within each pitch period. The relative positions of these instants of significant excitation in the direct component of the speech signal remain unchanged at each of the microphones for a given speaker. These sequences differ only by a fixed delay corresponding to the relative distances of the microphones from the speaker [10]. Moreover, in the vicinity of the instants of significant excitations, the speech signal exhibits a high SNR relative to the other regions, due to damping of the impulse response of the vocal tract system. While the reflected components and noise may also contribute to some high SNR regions, their relative positions will be different in the signals collected at the two microphones. Hence, the coherence of the high SNR regions in the direct components of the signals at the two microphones can be exploited for estimating the time delay.

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 [11]. The LP residual removes the 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. The cross-correlation function of the LP residual signals from the two microphone signals is not likely to yield strong peaks, as the large amplitude fluctuations will be 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 [12]. The Hilbert envelope $h[n]$ of the LP residual signal $e[n]$ is given by

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

where $e_h[n]$ is the Hilbert transform of $e[n]$ [13]. The HE of the LP residual is shown in Fig. 1(c). The HEs of the LP residuals of the multispeaker signals are used to estimate the time delays.

III. Determining the Number of Speakers

The cross-correlation function of the HEs of the LP residual signals derived from the multispeaker signals is used to determine the number of speakers. Apart from the large amplitudes around the instants of significant excitation, the HE also contains a large number of small positive values, which may result in spurious peaks in the cross-correlation function. The regions around the instants of significant excitation are further emphasized by dividing the square of each sample of HE by the moving average of the HE computed over a short window around the sample. The computation of the preprocessed HE is as follows:

$$g_i[n] = \frac{h_i^2[n]}{\frac{1}{2M+1}\sum_{m=n-M}^{n+M} h_i[m]}, \quad i \in \{1, 2, \ldots, p\} \tag{3}$$

where $g_i[n]$ is the preprocessed HE of the LP residual of multispeaker signal collected at the i^{th} microphone, M is the number of samples corresponding to 4 ms duration, 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 given in Fig. 1(c). In this paper, we consider multispeaker signals collected using a pair of microphones, and hence $p = 2$. The cross-correlation function $r_{12}[l]$ between the preprocessed HEs $g_1[n]$ and $g_2[n]$ is computed as

$$r_{12}[l] = \frac{\sum_{n=z}^{N-|k|-1} g_1[n] g_2[n-l]}{\sqrt{\sum_{n=z}^{N-|k|-1} g_1^2[n] \sum_{n=z}^{N-|k|-1} g_2^2[n]}}, \quad l = 0, \pm 1, \pm 2, \ldots, \pm L \tag{4}$$

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 delay can be estimated from the approximate positions of the speakers and 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 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 and 2) there could be spurious peaks in the cross-correlation function, which may not correspond to the delay due to a speaker. Hence, we rely only on the delay due to the most prominent peak in the cross-correlation function.

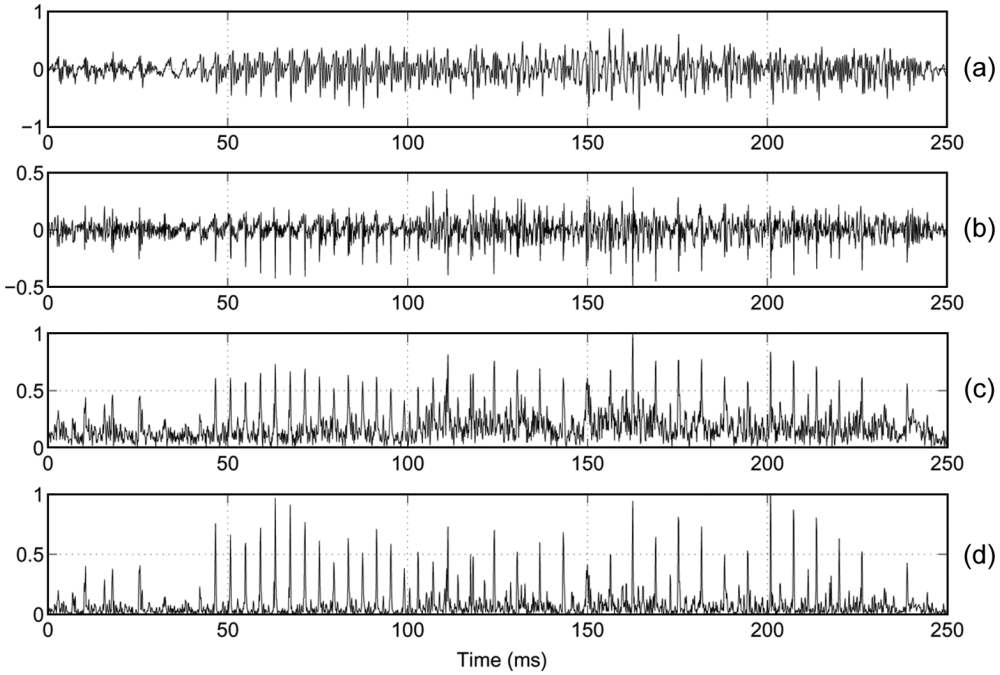

Fig. 1. (a) A 250 ms segment of speech signal, (b) its LP residual, (c) HE of LP residual, and (d) HE after emphasizing the GCIs.

This delay is computed from the cross-correlation function of successive frames of 50 ms duration shifted by 5 ms. Since different regions of 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 the determination of 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 evidence in favor of each delay, for a recording consisting of speech from three speakers. The figure shows three prominent peaks corresponding to the three speakers.

IV. Experimental Results

Experiments were conducted using different multispeaker signals containing three, four, five, and six speakers. Speech data was collected simultaneously using two microphones separated by about 1 m in a laboratory environment, with an average (over the frequency range of 0.5–3.5 kHz) reverberation time of about 0.5 s. All recordings for this study were made under the following practical conditions.

a) The speakers were seated approximately along a circle, at an average distance of about 1.5 m from the microphones. The speakers were seated such that their heads and the microphones were approximately in the same plane.
b) The speakers were positioned in such a way that the delay is different for different speakers. In fact, any random placement of speakers with respect to the microphones satisfies this requirement.
c) It is assumed that the level of the direct component of speech from each speaker at the microphones is significantly higher relative to the noise and reverberation components in the room.

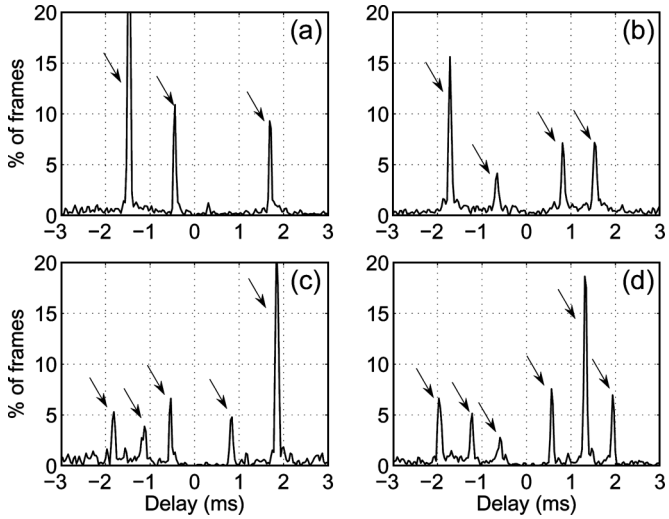

Fig. 2. Percentage of frames for each delay in milliseconds for (a) three speakers, (b) four speakers, (c) five speakers, and (d) six speakers. The arrows indicate the peaks corresponding to different speakers.

d) All the speakers were stationary, and spoke simultaneously during the entire duration of recording, resulting in significant overlap.

The speech signals were sampled at 32 kHz. During each recording, the distances of the speakers from both the microphones were measured. The actual time delay of arrival τ of speech signals at *Mic-1* and *Mic-2* located at distances d_1 and d_2, respectively, from a speaker is given by

$$\tau = \frac{d_1 - d_2}{c} \tag{5}$$

where c is speed of sound in air. A negative time delay (lead) indicates that the speaker is nearer to *Mic-1* relative to *Mic-2*.

TABLE I
COMPARISON OF ESTIMATED TIME DELAYS $\hat{\tau}$ WITH THE TIME DELAYS τ COMPUTED FROM THE MEASURED DISTANCES d_1 AND d_2

# Speakers	Speaker	d_1 (m)	d_2 (m)	τ (ms)	$\hat{\tau}$ (ms)
3	*Spkr-1*	0.45	0.98	-1.5	-1.47
	Spkr-2	0.93	1.11	-0.51	-0.47
	Spkr-3	1.48	0.91	1.63	1.69
4	*Spkr-1*	0.55	1.14	-1.7	-1.72
	Spkr-2	1.01	1.23	-0.63	-0.65
	Spkr-3	1.43	1.17	0.74	0.81
	Spkr-4	1.21	0.68	1.5	1.5
5	*Spkr-1*	0.6	1.24	-1.83	-1.8
	Spkr-2	0.88	1.29	-1.2	-1.13
	Spkr-3	1.30	1.49	-0.54	-0.56
	Spkr-4	1.42	1.14	0.80	0.81
	Spkr-5	1.16	0.54	1.77	1.81
6	*Spkr-1*	0.4	1.08	-1.9	-2
	Spkr-2	0.82	1.29	-1.3	-1.25
	Spkr-3	1.19	1.4	-0.6	-0.59
	Spkr-4	1.39	1.18	0.6	0.56
	Spkr-5	1.42	0.96	1.3	1.31
	Spkr-6	1.4	0.75	1.9	1.94

The multispeaker signals were processed using the proposed method to obtain the time delays. A 16th-order LP analysis 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. The percentage of frames for each delay (in ms) for three, four, five, and six speakers are shown in Fig. 2. The locations of the peaks in the histograms correspond to the time delays due to different speakers. Thus, the number of peaks in the histogram indicates the number of speakers, and the heights of the peaks show the relative prominence of each speaker in the conversation. Table I lists the actual time delay τ obtained from the measured distances d_1 and d_2 (5), and the estimated time delays $\hat{\tau}$ obtained from the histograms. The actual and the estimated time delays are in close agreement, thus indicating the effectiveness of the proposed method in determining the number of speakers and their corresponding time delays from multispeaker signals. The deviation in some cases could be attributed mostly to the inaccuracies in the measurement of distances between speakers and microphones.

V. SUMMARY AND CONCLUSIONS

In this letter, a method for determining the number of speakers from the multispeaker speech signals at two spatially separated microphones is proposed. This method works even for an underdetermined case, where the number of sensors is far less than the number of sources. The proposed method exploits the time delay of arrival of speech signals between the two microphones for a given speaker. The multispeaker speech signals are preprocessed to highlight the regions of significant excitation of the vocal tract system. Since the direct component of signals generally dominates over the reflected or reverberant components, the method can be applied for speech signals collected in a room having some reverberation and background noise. The method fails if the direct components are masked by high levels of ambient noise and reverberation. The proposed method was demonstrated for the case where the time delays are distinct for each speaker. The problems of specific or arbitrary distribution of speakers relative to microphone positions can be overcome by using pairs of several spatially distributed microphones. Use of several microphones can also reduce the problem of weak signals of some speakers at a given pair of microphones. In this study the speakers were stationary during recording sessions. This ensures that the time delays are nearly constant. In situations where the speakers are moving, variation of the time delays must be tracked to determine the number of speakers.

REFERENCES

[1] M. S. Bartlett, "A note on the multiplying factors for various x^2 approximations," *J. Roy. Stat. Soc.*, vol. 16, no. ser B, pp. 296–298, 1954.
[2] D. N. Lawley, "Tests of significance of the latent roots of the covariance and correlation matrices," *Biometrica*, vol. 43, pp. 128–136, 1956.
[3] M. Wax and T. Kailath, "Detection of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 387–392, Apr. 1985.
[4] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai, "On the detection of number of signals in the presence of white noise," *J. Multivariate Anal.*, vol. 20, pp. 1–20, Jan. 1986.
[5] E. Fishler and H. V. Poor, "Estimation of the number of sources in unbalanced arrays via information theoretic criteria," *IEEE Trans. Signal Process.*, vol. 53, no. 9, pp. 3543–3553, Sept. 2005.
[6] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. New York: Wiley Interscience, 2001.
[7] J. F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Process.*, vol. 44, no. 12, pp. 3017–3030, Dec. 1996.
[8] 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.
[9] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Eaglewood Cliffs, NJ: Prentice-Hall, 1993.
[10] 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. 6, pp. 1110–1118, Nov. 2005.
[11] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.
[12] 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. ASSP-27, no. 4, pp. 309–319, Aug. 1979.
[13] A. V. Oppenheim, R. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Upper Saddle River, NJ: Prentice-Hall, 2000.