Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 4, April 2015, pg.747 – 753

RESEARCH ARTICLE

Speakers Determination and Isolation from Multispeaker Speech Signal

Ms. Asharani V R[1], Mrs. Anitha G[2], Dr. Mohamed Rafi[3]

[1]Department of Computer Science, UBDTCE, Davangere, Visvesvaraya Technological University, India

[2]Department of Computer Science, UBDTCE, Davangere, Visvesvaraya Technological University, India

[3]Department of Computer Science, UBDTCE, Davangere, Visvesvaraya Technological University, India

[1] asharanivr0691@gmail.com, [2] anithapoojar@yahoo.con, [3] rafimohamed17@gmail.com

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 multi speaker signals. The key idea is that for a given speaker, the relative spacing's 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 multi speaker signals collected at the two microphones.

Keywords— Excitation Source, Hilbert Envelope, Linear Prediction Residual, Multispeaker speech signal, time-delay estimation

I. Introduction

A The Main problem in signal processing is to estimate the number of sources from multisensor data. Same problem is arise 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.

In particular, the situation when the sensor noise levels are spatially inhomogeneous is considered in 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 observations collected at sensors from sources, is given by

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

where

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

Here, xi[n] is the mixed signal at the ith sensor, sj[n] is the signal generated from the jth source, vi[n] is the additive noise at the ith sensor, A is the mixing matrix, N is the number of observations, and the superscript T indicates the transpose operation. The jth column vector of the mixing matrix A ([a1i, a2i.aij] T) gives the array response associated with the jth source signal. The ith row vector of the mixing matrix A ([a1i, a2i, aij] T) gives the mixing weights for the source signals collected at ith the 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 x[n]. The well-determined case(p=q) is commonly addressed using the independent component analysis (ICA) formulation . 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 .

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 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.

II. Literature Survey

***Q. Lv and X.-D. Zhang*:** We are proposed a unified method for blind separation of sparse sources with unknown source number; blind source separation is of high interest in practical applications. Some approaches are proposed for the unknown number of sources of the BSS problem. However, they only consider the over determined case with the number of sensors more

than the number of sources. To implement the practical BSS without prior assumption on the number of sources, we propose a new BSS method. It uses the unsupervised robust C prototypes algorithm to estimate the mixing matrix and then makes the estimation of sources. Simulations using speech signals confirm the validity of the proposed method.

Advantages:

The implement the practical BSS without prior assumption on the number of sources, separation is of high interest in practical applications.

Disadvantages:

They only consider the over determined case with the number of sensors more than the number of sources.

***B. Yegnanarayana, S. R. M. Prasanna, R. Duraiswami, and D. Zotkin*:** We are proposed processing of reverberant speech for time-delay estimation, in this paper, we present a method of extracting the time-delay between speeches signals collected at two microphone locations. Time-delay estimation from microphone outputs is the first step for many sound localization algorithms, and also for enhancement of speech. For time-delay estimation, speech signals are normally processed using short-time spectral information. The spectral features are affected by degradations in speech caused by noise and reverberation. Features corresponding to the excitation source of the speech production mechanism are robust to such degradations. We show that these source features can be extracted reliably from the speech signal. The time-delay estimate can be obtained using the features extracted even from short segments of speech from a pair of microphones. The proposed method for time-delay estimation is found to perform better than the generalized cross-correlation approach. A method for enhancement of speech is also proposed using the knowledge of the time-delay and the information of the excitation source.

Advantages:

The proposed method estimates the delays accurately, whereas the GCC shows significant variations in the estimated delays. Ideally all the points should lie along a vertical line at the delay value.

Disadvantages:

Compensating for the spectral side effects or enhancement of the speech spectral components have met with limited success only provide, a vertical line at the delay value. So the spread of points from the vertical line indicates degradation in the performance of the method.

***M. Wax and T. Kailath*:** We are proposed Detection of signals by information theoretic Criteria, Detecting the number of sources is a well-known and a well-investigated problem. In this problem, the number of sources impinging on an array of sensors is to be estimated. The common approach for solving this problem is to use an information theoretic criterion like the minimum description length, or the Akaike information criterion. Although it has been gaining much popularity and has been used in a variety of problems, the performance of information theoretic criteria-based estimators for the unknown number of sources has not been sufficiently studied, yet. In the context of array processing, the performance of such estimators was analyzed only for the special case of Gaussian sources where no prior knowledge of the array structure, if given, is used. Based on the theory of misspecified models, this paper presents a general asymptotic analysis of the performance of any information theoretic criterion-based estimator, and especially of the MDL estimator. In particular, the performance of the MDL estimator, which assumes Gaussian sources and structured array when applied to Gaussian sources, is analyzed. In addition, it is shown that the performance of a certain MDL estimator is not very sensitive to the actual distribution of the source signals. However, appropriate use of prior knowledge about the array geometry can lead to significant improvement in the performance of the MDL estimator. Simulation results show good fit between the empirical and the theoretical results.

Advantages:

The MDL estimator has significant advantage over the GMDL estimator. The GMDL estimator needs four times more snapshots in order to perform as well as the MDL estimator.

Disadvantages:

The disadvantage in using the GMDL estimator for digital signals is related to the number of sources that can be detected.

***E. Fishler and H. V. Poor*:** We are proposed Estimation of the number of sources in unbalanced arrays via information theoretic criteria, estimating the number of sources impinging on an array of sensors is a well-known and well-investigated problem. A common approach for solving this problem is to use an information theoretic criterion, such as Minimum Description Lengthor the Akaike Information Criterion. The MDL estimator is known to be a consistent estimator, robust against deviations from the Gaussian assumption, and non-robust against deviations from the point source and/or temporally or spatially white additive noise assumptions. Over the years, several alternative estimation algorithms have been proposed and tested. Usually, these algorithms are shown, using computer simulations, to have improved performance over the MDL estimator and to be robust against deviations from the assumed spatial model. Nevertheless. In this paper, a systematic approach toward the problem of robust estimation of the number of sources using information theoretic criteria is taken. An MDL-type estimator that is robust against deviation from assumption of equal noise level across the array is studied. The consistency of this estimator, even when deviations from the equal noise level assumption occur, is proven.

Advantages:

A novel low-complexity implementation method avoiding the need for multidimensional searches is presented as well, making this estimator a favorable choice for practical applications.

Disadvantages:

These robust algorithms have high computational complexity, requiring several multidimensional searches, which is motivated by real-life get problems.

***J. F. Cardoso and B. Laheld*:** We are proposed Equivariant adaptive source separation, Source separation consists in recovering a set of independent signals when only mixtures with unknown conceits are observed. This paper introduces a class of adaptive algorithms for source separation which implements an adaptive version of equivariant estimation and is henceforth called EASI. The EASI algorithms are based on the idea of serial updating: this specie form of matrix updates systematically yields algorithms with a simple, parallelizable structure, for both real and complex mixtures.

Most importantly, the performance of an EASI algorithm does not depend on the mixing matrix. In particular, convergence rates, stability conditions and interference rejection levels depend only on the (normalized) distributions of the source signals. Close form expressions of these quantities are given via an asymptotic performance analysis. This is completed by some numerical experiment sill us trading the activeness of the proposed approach.

Advantages:

It entails very little extra computation with respect to and it does not introduce additional parameter. The range of the output signals allows to properly scaling the non-linearity.

Disadvantages:

Determines an upper limit to the noise level, under which the performance of EASI does not depended, the scaling in determinations inherent to the source separation problem, some parameters have to be arbitrarily fixed.

Proposed method:

The 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. 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 proposed method was demonstrated for the case where the time delays are distinct for each speaker. 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.

III. System Design

During the production of voiced speech, the vocal tract system is excited by a quasi-periodic sequence of impulse-like excitations. 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. 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.

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. 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.The high SNR regions around the GCIs can be highlighted by computing the Hilbert envelope (HE) of the LP residual the HEs of the LP residuals of the multispeaker signals are used to estimate the time delays.

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 in-stants 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

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

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.

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 empha-sized 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:

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

where $g_i[n]$ is the preprocessed HE of the LP residual of multi-speaker 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

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

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 de-lays 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 be-cause of the following reasons: 1) all speakers may not con-tribute 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.

IV. 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, atan 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.

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

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.

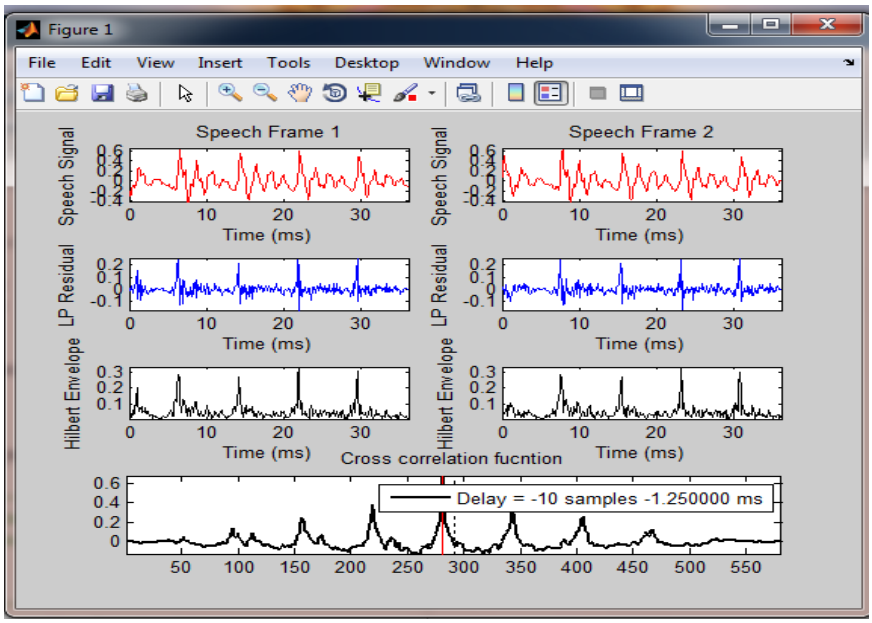

Fig. time delay calculation b/w different speech signal

V. Conclusion

In this letter, we are going to 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. Use of several microphones can also reduce the problem of weak signals of some speakers at a given pair of microphones.

REFERENCES

1. 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.
2. 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.
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. 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.
5. J. F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Process.*, vol. 44, no. 12, pp. 3017–3030, Dec.1996.

