ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

TDOA ESTIMATION FOR MULTIPLE SPEAKERS IN UNDERDETERMINED CASE

Mariem Bouafif [1], *Zied Lachiri* [1,2]

[1] LSTS-SIFI Laboratory, National Engineering School of Tunis, BP53, Campus Universitaire, 1002, le Belvedere, Tunis, Tunisia
[2] Depart. of Physic and Instrumentation, National Institute of Applied Sciences and Technology, BP 676, Centre Urbain, 1080, Cedex, Tunis, Tunisia
mariem.bouafif@gmail.com, zied.lachiri@enit.rnu.tn

10.21437/Interspeech.2012-477

Abstract

In this paper we address the issue of estimating the time delay of arrival in underdetermined case. We develop a method using the excitation characteristics of the speech production. This method is based on the cross correlation of the Hilbert Envelops of linear prediction residuals derived from two microphones signals. The method has been applied to real data obtained by recording many sources captured by a pair of microphones. Experiments show that reverberation distorts the input signals, each reverberation causes an extra peak in the cross-correlation. This makes it difficult to determine which peak is the central time-delay peak and which are just reverberation sidelobes. An alternative time delay estimation method has been implemented and compared to spectrum angular methods.

Index Terms: TDOA, Linear prediction, Hilbert Envelop.

1. Introduction

Several approaches for the time delay of arrival (TDOA) estimation have been addressed since it's basically used for determining number of speakers, underdetermined sound localization and separation.

These methods may be broadly classified into three categories: The first approach is based on converting the observed phase difference into a TDOA in each time-frequency bin and building a histogram of the TDOAs whose peaks point to the sources [1, 2, 3]. The second approach is the clustering method, based on alternately clustering the time frequency bins into sources and updating the source TDOAs according to the observed phase differences [4]. The third approach called angular spectrum [5, 6, 7] based on building a function of TDOA whose peak(s) indicate the TDOA(s) compatible with the observed phase difference in each time-frequency bin, and summing this function over all bins. Existing methods suffer many problems. In fact a limitation of current angular spectrum-based methods is that they essentially assign the same weight to all observed phase differences, whether they result from the direct sound of single source or from a mixture of direct and reverberated sound and/or several sources, and they necessitate a prior guess of the number of speakers. On the other hand, the clustering method necessitates an initial guess of the number of sources and their TDOA's due to local optima.

In the following, we consider the first approach which is applicable to small microphones spacing and does not necessitates neither a prior guess of number of sources, either their TDOA's.
The proposed approach based on the excitation component of the speech signal. A Hilbert Envelop (HE) of the linear prediction residual is first derived from the tow mixed signals captured by tow sensors, followed by a preprocessing to emphasize instants of significant excitation. Finally the number of speakers and their time delays of arrival are synthesized over a cross-correlation between the modified and preprocessed Hilbert envelopes of linear prediction residual of the tow mixed signals.
The remainder of the paper is organized as follows. TDOA's and speakers' number estimation using excitation sources components in underdetermined case is reviewed in Section 2. Comparison of the proposed method against different angular spectrum methods simulated on TIMIT database, and evaluation of the proposed method on SISEC2010 database are provided in section3. Conclusions are drawn in Section 4.

2. TDOA estimation based on excitation source components

We can describe the problem using the Short-Time Fourier Transform (STFT).

$$X(t,f) = \sum_{n=1}^{N} \tau_n(f) S_n(t,f) + B(t,f) \qquad (1)$$

Where $X(t,f) = [X_1(t,f)\, X_2(t,f)]^T$ is the STFT of the observed signals at the two microphones, $S_n(t,f)$ is the n-th source signal in time frame t and frequency bin f, and τ_n is the TDOA of the n-th source signal. The mixture $X(t,f)$ can be modeled as the sum of n delayed sources and reverberation $B(t,f)$.
Determining the n delays using the excitation component of the speech signal is better than using the speech signal itself [8, 9]. For this, we propose the use of residual derived from the speech signal by linear prediction (LP) analysis [10]. The time delay can be estimated from the cross-correlation of the HE's of LP residuals of the mixed speech signals collected at the tow microphones [11].

In LP analysis, each sample is predicted as a linear combination of the past p samples, where *p* is the order of prediction. The error between the speech sample and It's predicted value denoted *e(n)*.

$$e(n) = \hat{S}(n) - S(n) \quad (2)$$

Where $S(n)$ is the speech signal sample at n-th instants, and $\hat{S}(n)$ is it's predicted one.
The Hilbert envelop of the LP residual *e(n)* is :

$$h(n) = \sqrt{e^2(n) + e_h{}^2(n)} \quad (3)$$

Where $e_h(n)$ is the Hilbert transform of the LP residual *e(n)*.
The HE's of the LP residual of the tow microphones signals are shown in Figure 1 (a), and Figure 1 (b) for the case of three speeches played simultaneously at three loudspeakers in a Dolby studio with an average reverberation of 300ms. Speeches were taken from the TIMIT database. The two microphones are spaced at 1m apart and the loudspeakers are in the same horizontal plane as microphones. A 12-th order LP analysis is performed on the speech signal recorded at a 44.1 kHz sampling rate.
Number of speakers and there TDOAs could be determined by the location of spurious peaks in cross-correlation between HE's of LP residual of the tow channels Figure1 (c). In fact the large number of small values in the HE's of LP residual of each channel may result in spurious peaks the in cross-correlation, that make difficult to distinguish between lobs defining speakers and others which are just the effect of reverberation. We need to preprocess the signals before estimating time delays by emphasizing the instants of significant excitation in HE's as follow:

$$g_i[n] = \frac{h_i^2[n]}{\frac{1}{2M+1}\sum_{m=n-M}^{n+M} h_i[m]} \quad (4)$$

Where *i = 1,2,* $h_1[n]$ and $h_2[n]$ are the HE's of the LP residual derived from the tow channel, and M = 88 samples at 44.1 kHz.
The scope here is that a region of 2ms (equivalent to 88 samples at 44.1 kHz) around the instants of significant excitation represents regions of high SNR in the speech. Squaring the HE's sample values bring down the dynamic range. And dividing it by the running mean suppress the spurious peaks in HE's. Emphasizing peaks can be seen in Figure 1 (d) and Figure 1 (e).
The modified HE's could be more processed by subtracting the minimum as follows:

$$m_i[n] = g_i[n] = - \min(g_1[n], g_2[n]) \;] \quad (5)$$

It ensures that there are no strong peaks in cross-correlation function at zero lag. The improvement in the resolution can be seen in Figure 1 (g) and Figure 1 (h), where we have processed HE's along 500ms at a sampling rate of 44.1 kHz. This choice was found to give the best performance experimentally.
The location of the peaks with the respect to the center point (zero lag) in the cross correlation function between the two microphones signal correspond to the time delays of arrival of each speaker during 500ms Figure 1 (i).

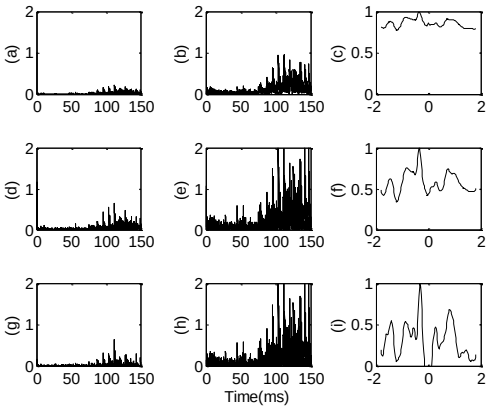

Figure 1 : *(a) and (b) are HE's of LP residual of Mic1 and Mic2 signals. (d) and (e) are g1[n] and g2[n]. (g) and (h) are m1[n] and m2[n]. (c) is the normalized cross-correlation of (a) and (b). (f) is the normalized cross-correlation of (d) and (e), and (i) is the normalized cross-correlation of (g) and (h).*

This delay is computed from the cross-correlation function of successive frames of 500ms shifted by 20ms all over mixed speech. The occurred number of each delay (in term of number of samples) is computed along the mixed speech as shown in Figure 2.

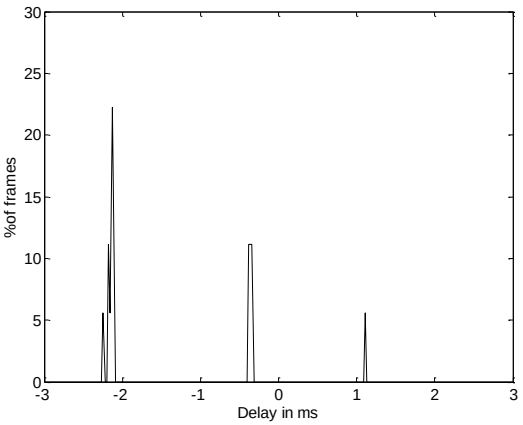

Figure 2: *Percentage of number of frames of each speaker as function of delays*

3. Experimental results

3.1. Experimental conditions

In the beginning, we evaluate the proposed method and angular spectrum methods comparing them to the real measured TDOA, and then comparing them on term of F-measure criterion. Experiments were conducted using multi-speaker signals containing three speakers collected from TIMIT database and played simultaneously throw loudspeakers in a Dolby studio 5.1. The mixed speech data were sampled at 44.1 kHz and collected using tow cardioids microphones separated by a distance of 1m. The loudspeakers were positioned at different distances as shown in Figure 3. This arrangement was made to ensure that each speaker produces different delays at tow microphones.
The proposed method is then evaluated on term of F-measure to test its robustness against a varying number of speakers.

Sources were taken from the underdetermined database of the 2010 Signal Separation Evaluation Campaign (SiSEC) [12].
A mixed speech containing two to six sources, six reverberation times (from 50ms to 750ms), four microphones spacing (from 5cm to 1m), four distances between the sources and the center of the microphones pair(from 20 ms to 2ms), and Three sources types (male, female, and music) sampled at 16kHz.

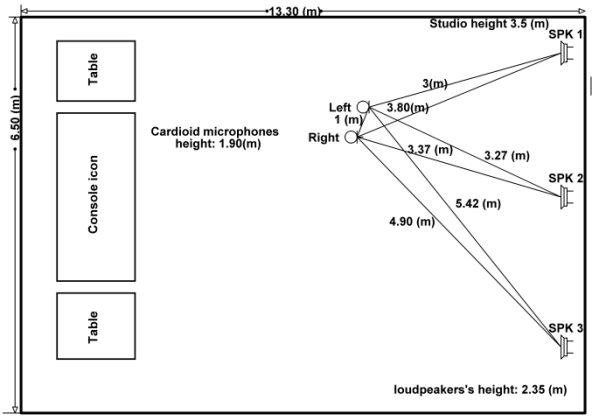

Figure 3: *The configuration of loudspeakers and microphones in a Dolby 5.1 studio.*

3.2. Evaluation on TIMIT database

The real TDOA for each speaker must be in the range $[\frac{-d}{c}, \frac{d}{c}]$, where d is the distance between microphones and c is the speed of sound in air ($c = 340{,}29\, ms^{-1}$).
Table1 lists the real time delay τ computed from the measured distances d_{i1} and d_{i2} and the estimated time delay $\hat{\tau}$ obtained from the proposed algorithm and four existing angular spectrum methods [13]: GCC-PHAT [7], MUSIC [14], MVDR [15], and DS [16].
The proposed method is then evaluated by using another evaluation criterion depending on the number of highest peaks and the number of speakers. An estimated time delay $\hat{\tau}_i$ is considered to be a correct estimate of a true time delay τ_i if $|\frac{c}{d}(\tau_i - \hat{\tau}_i)| \leq \mu$ with μ set in our experiments to 0.05.
If we select the TDOAs corresponding to the highest peaks of the angular spectrum, we can evaluate methods by defining the F-measure criterion [17]:

$$F(J) = 2\,\frac{R(J) * P(J)}{R(J) + P(J)} \qquad (6)$$

Where $R(J) = \frac{I_j}{N}$ is the recall and $P(J) = \frac{I_j}{J}$ is the precision. We note $I_j$the number of correct TDOAs, and N the number of speakers.

To use angular spectrum methods, we need to fix an optimal number of peaks equal to the known number of speakers to have the best F-measure. If we consider the number of speakers equal to the real number of peaks and we compare results with the proposed algorithm, we show in Table 2 that the last one provides the best F-measure.

Table 2. *Comparison of the proposed algorithm against angular spectrum methods on term of F-measure for the case of 3 speakers (TIMIT database).*

	Proposed algorithm	GCC-PHAT	MVDR	MUSIC	DS
F(J)	0.726	0.6	0.5	0.4	0.5

3.3. Evaluation on (SISEC) database

In this section we investigate the behavior of the F-measure as a function of the microphone spacing, the reverberation time, and the number of selected peaks using a varying number of speakers.
Figure 4-6 show the average of F-measure for different configurations as a function of reverberation time RT60, distance between microphones, and number of speakers respectively.
The proposed method performs well for a reverberation smaller than 300ms. For higher reverberation the F-measure slightly decreases because it affects the number of peaks by adding other sidelobs.
For different number of speakers, the algorithm seems to be performing better for a moderate and large distance between microphones (30cm, and 1m). It's clear for a minimum reverberation time the highest peaks are corresponding to the number of speakers.

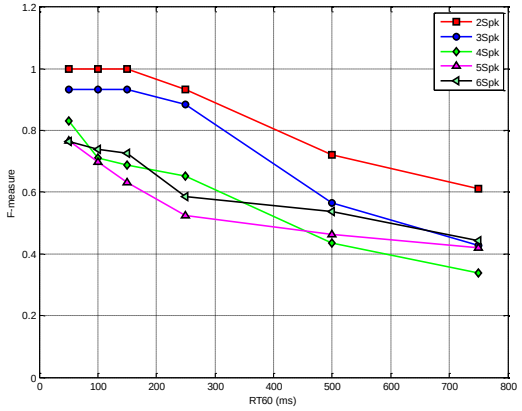

Figure 4 : *Average F-measure as a function of reverberation.*

Table 1. *Comparison of estimated $\hat{\tau}_i$ for different algorithms against the real τ_i*

No of spk	ith spk	d_{i1}	d_{i2}	Real τ	Proposed algorithm	GCC-PHAT	MVDR	MUSIC	DS
3	1	3	3.80	-2.35	-2.15	-2.138	-0.356	2.105	-0.356
	2	3.27	3.37	-0.29	-0.36	-0.356	0.097	1.425	0.097
	3	5.42	4.90	1.5	1.12	1.036	-0.097	0.356	1.036

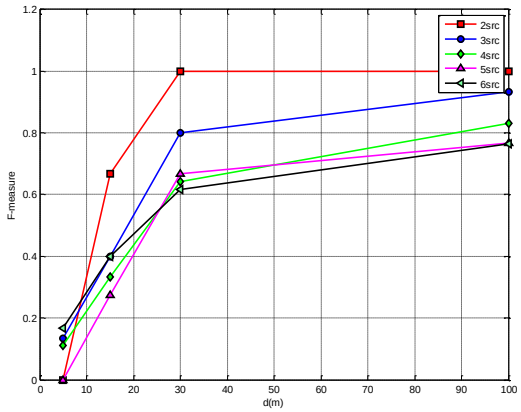

Figure 5: *Average F-measure as a function of the distance between microphones for different number of speakers.*

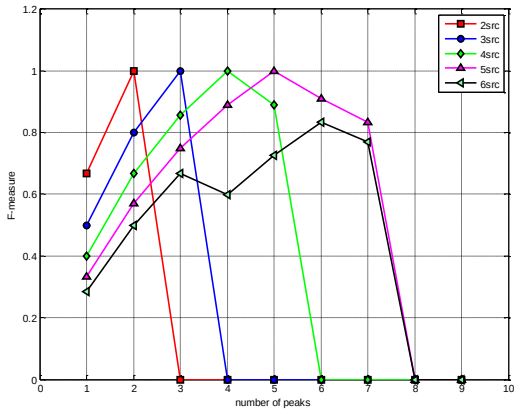

Figure 6 : *Average F-measure as a function of number of peaks for a 50 ms reverberation.*

4. Conclusion

In this paper, we proposed a method for estimation of time-delays and number of speakers in underdetermined case using the excitation source components of speech production.

We have tested the algorithm on a 3 speakers mixed speech captured by a pair of microphones, and we have compared it with different angular spectrum methods.

The experimental results show that the number of speakers and their time delays estimated by the proposed method are closer to the actual values than those estimated by the angular spectrum methods. The proposed method compute the number of speakers as well as their time delay of arrival, but the angular spectrum methods rely on a known speaker's number and could only compute the time delays of each one.

The only limitation for the proposed method is reverberation, affecting the number of peaks and decreasing the F-measure. One way to improve this method would be to apply dereverberation on the mixed speech before computing TDOA's.

We have evaluated the proposed method by trying to apply it on multi-configurations context. In fact the proposed method performs well in a minimum reverberation time, and for a microphones distance's above 0.6m. Computing F-measure function of a varying number of peaks, the highest peaks are corresponding to the number of speakers.

5. References

[1] O. Yilmaz and S.T. Rickard. Blind separation of speech mixtures via time-frequency masking. IEEE Transactions on Signal Processing, 52(7):1830–1847, 2004.

[2] H. Sawada, S. Araki, R. Mukai and S. Makino. Grouping separated frequency components by estimating propagation model parameters in frequency-domain blind source separation. IEEE Trans. on Audio, Speech, and Language Processing, 15(5):1592–1604, 2007.

[3] S. Arberet, R. Gribonval and F. Bimbot. A robust method to count and locate audio sources in a multichannel underdetermined mixture. IEEE Transactions on Signal Processing, 58(1):121–133, 2010.

[4] O Y. Izumi, N. Ono, and S. Sagayama. Sparseness-based 2ch BSS using the EM algorithm in reverberant environment. In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pages 147–150, 2007.

[5] F. Nesta, P. Svaizer, and M. Omologo. Cumulative state coherence transform for a robust two-channel multiple source localization. In Proc. 8th Int Conf on Independent Component Analysis and Signal Separation, pages 290–297, 2009.

[6] R. Schmidt. Multiple emitter location and signal parameter estimation. IEEE Transactions on Antennas and Propagation, 34(3):276–280, 1986.

[7] C. Knapp and G. Carter. The generalized crosscorrelation method for estimation of time delay. IEEE Transactions on Acoustics, Speech and Signal Processing 24(4):320–327, 1976.

[8] B. Yegnanarayana, S. Prasanna, R. Duraiswamy and D. Zotkin. Processing reverberant speech for time-delay estimation. IEEE transactions on speech and audio on processing, vol. 13, no. 6, November 2005.

[9] B Yegnanarayana, S. Prasanna, KS Rao: Speech enhancement using excitation source information. Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, vol. I

[10] J.Makhoul: Linear prediction: A tutorial review. Proc. IEEE 63 (1975) 561, 580.

[11] S. R. M. Prasanna, Event Based Analysis of Speech, Ph.D. thesis, Dept. of Computer Science and Eng, IIT Madras, India, 2004.

[12] "http://sisec.wiki.irisa.fr/inria".

[13] Charles Blandin, Emmanuel Vincent and Alexey Ozerov: Multi-source TDOA estimation in reverberant audio using angular spectra and clustering. Published in "Signal Processing 92 (2012) 1950-1960".

[14] R. Schmidt. Multiple emitter location and signal parameter estimation. IEEE Transactions on Antennas and Propagation, 34(3), pp. 276–280, 1986.

[15] H. Krim, M. Viberg, Two decades of array signal processing research: the parametric approach, IEEE Signal Processing Magazine 13 (4), pp: 67-94, 1996.

[16] F. Nesta, P. Svaizer, M. Omologo, Cumulative state coherence transform for a robust two-channel multiple source localization, in: Proc. 8th Int. Conf. on Independent Component Analysis and Signal Separation (ICA), 2009, pp. 290–297.

[17] C. J. Van Rijsbergen. Information Retrieval, Butterworths, London, UK, 1979.