Solving global permutation ambiguity of time domain BSS using speaker specific features of speech signals

Vahid Khanagha, Ali Khanagha

Solving global permutation ambiguity of time domain BSS using speaker specific features of speech signals

Vahid Khanagha
Iran University of Science and Technology
Tehran, Iran
vkhanagha@ee.iust.ac.ir

Ali Khanagha
Amirkabir University of Technology
Tehran, Iran
khanaghaa@ripi.ir

Abstract—**Multidimensional localization of multiple sources using BSS based TDOA estimators, requires the solution of global permutation ambiguity before fusing several TDOA estimations. Since the separation quality of BSS isn't always perfect, it is not easy to decide which TDOA belongs to which source. Here we study the possibility of using several speaker specific features of speech signal in order to recognize perceptually dominant sources in each one of moderately separated outputs of BSS algorithm. We compare the feasibility of different features in terms of validity rate of decisions and computational complexity.**

I. INTRODUCTION

Fusing several TDOA estimates from multiple microphone arrays is a well known method for multidimensional localization of speech sources. Since traditional methods such as generalized cross correlation (GCC) are incapable of estimating TDOAs of multiple active sources in reverberant environments, the use of time domain BSS in TDOA estimation has introduced [1]. The BSS based TDOA estimation, not only provides robust TDOA estimates for multiple sources, but also separates original signals from the observed mixtures which can be used as a clue for the fusion of estimated TDOAs in order to accomplish spatial localization of active sources.

In fact, even if traditional methods, like GCC, were capable of precise TDOA estimation of multiple simultaneous sources, they still couldn't decide which TDOA from each array corresponds to which TDOA from other arrays. Although Some heuristic methods have been reported, like the outlier elimination in [2] and Inter-aural Time and level difference based method in [3], but they often suffer from algorithm complexity and the lack of generality. on the other hand, the use of time domain BSS for TDOA estimation provides the separated sources as a by-product which is ultimately valuable in relating different TDOAs from several arrays to each other and solving the global permutation ambiguity. With perfect separation quality, we can relate outputs of different BSS algorithms executed on different arrays with simple cross correlations as reported in [4]. But, usually speech separation quality is not perfect and we often achieve just a moderate amount of gain in Signal to Interference Ratio (SIR). In practice, for a 2×2 mixture situation, both sources would be present on both outputs but only one of them would be perceptually dominant (to some extent). But this dominancy is not enough to let us use cross correlations for solving permutation ambiguity.

Here, we assume the 2×2 mixture scenario and propose to use speaker specific features to qualify this relative dominancy for some moderately separated outputs of BSS algorithm. Our preliminary motivation for such approach was the fact that BSS is usually a pre-processing step for following actual speech processors such as speaker identification and speaker verification systems where these features are necessary to be extracted; based on source-filter model of speech production, we use some of the well known speaker specific features such as mel frequency cepsteral coefficients (MFCC), Perceptual linear prediction cepsteral coefficients (PLPCC), formant frequencies (F1, F2, F3, F4) and dynamic periodicity/aperiodicity information of speech frames. We also compare theses feature sets from computational complexity point of view.

II. BSS BASED TDOA ESTIMATION

Fig. 1. Shows a typical block diagram for mixing and unmixing systems, in which h_{ij} and w_{ij} are stand for the impulse response of corresponding FIR filters. It is easy to see that the prefect point of separation is met when following relations hold:

$$w_{11} = \alpha_1 h_{22} \qquad w_{12} = -\alpha_1 h_{21} \tag{1}$$

$$w_{21} = -\alpha_2 h_{12} \qquad w_{22} = \alpha_2 h_{11} \tag{2}$$

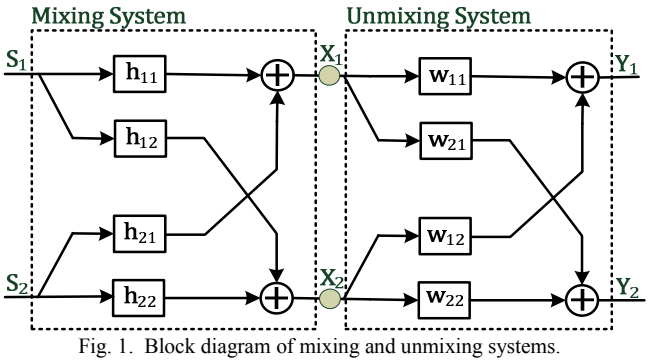

Fig. 1. Block diagram of mixing and unmixing systems.

If we write the TDOA of each source to the sensory array as[1]:

$$\tau_1 = argmax_n|h_{11}(n)| - argmax_n|h_{12}(n)| \quad (3)$$
$$\tau_2 = argmax_n|h_{22}(n)| - argmax_n|h_{21}(n)| \quad (4)$$

Using (1) and (2), it is easy to rewrite (3) and (4) as:

$$\tau_1 = argmax_n|w_{22}(n)| - argmax_n|w_{21}(n)| \quad (5)$$
$$\tau_2 = argmax_n|w_{11}(n)| - argmax_n|w_{12}(n)| \quad (6)$$

This way, we would have the TDOA estimates of both sources along with their separated version. However the separation quality is not perfect, but still the desired source is the perceptually dominant voice in BSS output.

Fig 2. Shows the overall proposed block diagram for Localization of two speakers. Two sets of microphone arrays are used to record sounds. For each array, an independent BSS algorithm is executed and using resulted separation system, a pair of TDOA estimates are calculated. Consequently we have two pairs of TDOA estimates from each array. However as a result of BSS permutation ambiguity we don't know which estimate of each TDOA pair belongs to which one of the sources.

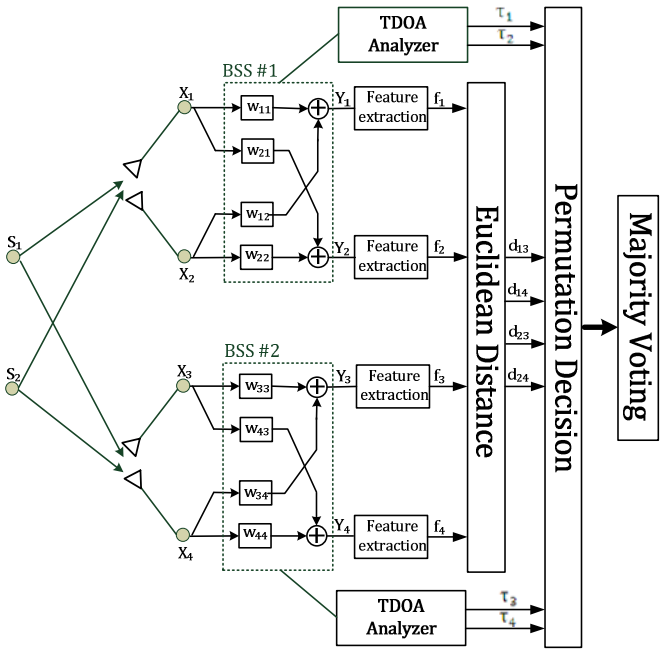

Fig. 2. the proposed block diagram for Localization of two speakers.

For example, if we assume the TDOA estimate τ_1 from array number 1, belongs to source S_1, we can't determine if the other TDOA estimate of S_1 from array number two is τ_3 or τ_4. In fact we need some additional information to decide about the correct pairing of estimated TDOAs from these two arrays. Fortunately BSS itself, provides separated output signals as of this required additional information. In fact, each TDOA estimate τ_i corresponds with BSS output Y_i. Consequently, the correct pair of TDOAs for each speaker, corresponds with two separated outputs which are representative for the same speaker. If the separation quality was ideal, we could pair outputs with similar waveforms with basic time domain correlators. But, usually speech separation quality is not perfect and we often achieve just a moderate amount of gain in Signal to Interference Ratio (SIR). In practice, for a 2×2 mixture situation, both sources would be present on both outputs but only one of them would be perceptually dominant (to some extent). Although human auditory system might percept this dominancy, the domination is not enough to let us use cross correlations for solving this global permutation ambiguity.

We propose to use speaker specific features to qualify this relative dominancy for these moderately separated outputs of BSS algorithms. Our preliminary motivation for such approach was the fact that BSS is usually a pre-processing step for following actual speech processors such as speaker identification and speaker verification systems where these features are necessary to be extracted.

Separated signals from each BSS are segmented into frames of length 30 msec. For each frame, the feature vectors are extracted and their Euclidean distance from each other is calculated. In Fig. 2, feature vectors are shown as f_i and the Euclidean distance between feature vectors f_i and f_j is shown as d_{ij}. Consequently, two decision variables are defined as:

$$dec_1 = d_{13} + d_{24} \quad (7)$$
$$dec_2 = d_{13} + d_{24} \quad (8)$$

For each new frame, we calculate the above decision variables and decide the permutation of that frame as:

$$\begin{aligned} &if\ dec_1 > dec_2\ then \\ &\qquad y_1 \equiv y_3\ , y_2 \equiv y_4 \\ &else \\ &\qquad y_1 \equiv y_4\ , y_2 \equiv y_3 \end{aligned} \quad (9)$$

Finally, we use the majority vote to decide about the permutation of the whole separated signals; each decision resembles a vote for one of possible permutations. After calculating these votes for all of the available frames, we choose the permutation which attains most of the votes.

The above discussion about solving permutation ambiguity was developed for 2×2 scenario. The same procedure might be easily generalized for 3×3 and 4×4 scenarios. We provide the simulation results for 3×3 scenario in section IV. Note that as the dimensionality of the problem increases the number of possible permutations are also increased. For example the two possible permutations for 2×2 scenario is increasing to 6 possible permutations for 3×3 scenario. Consequently we should define 6 decision variables like the ones in (7) and (8).

III. SPEAKER SPECIFIC FEATURES OF SPEECH SIGNALS

In this section, we provide a brief review about the nature of investigated speech features. Most of these features are based on the *source-filter* model of speech production. The *system* comprises of the vocal tract and lip radiations and

depends on the physical attributes of the speakers while the *source* represents the pulses produced by the air flow through the vocal cords and includes such information as the fundamental frequency which is mostly influenced by the contents of the speech signal rather than physical characteristics of speakers. The speech production process is modelled as filtering of the source spectrum (glottal pulses) by the system (vocal tract) [12]. Depending on the nature of uttered sound, source might be composed of periodic pulses (for voiced sounds) or white noise (for unvoiced sounds).

A. Source based features

The source information of a speaker depends on factors such as the shape and timing of the glottal pulses, whether or not the vocal folds close completely and, the trade-off between the glottal source and supraglottal source during voiced obstruent sounds. Based on these factors, one can describe the way a speaker sounds in terms of the voice quality of the speaker. These speaker-specific characteristics determine (a) the high frequency roll off of the speech spectrum, (b) the relative amplitudes of the very low-frequency harmonics, and (c) the harmonic and in-harmonic structure of the speech waveform respectively[5].

The most well known source based feature is the fundamental frequency of speech signals. Although this feature is useful for gender classification but it doesn't provide enough distinctions between the members of the same gender. Although, this distinction would be met if we track the fundamental frequency changes over time instead of observing its statistical average over all of the frames. But still, since only voiced sounds are periodic, the voiced/unvoiced decision has to be made. Also, the methods of fundamental frequency estimation are mostly prone to the presence of co channel interference of simultaneous speakers which is the case for outputs off the BSS algorithm.

For these reasons we used the A-Periodic Periodic detector (APP) introduced in [6] in order to exploit source properties of speech signals. The *summary measure* defined in [6] provides a quantified index about the amount of periodic and a-periodic energy of speech frames. It doesn't require deciding whether a given frame is voiced or unvoiced since it summarizes the amount of periodicity, *i.e.* the largest dip strength of Average Magnitude Difference Function (AMDF), of the envelopes of several sub-bands of speech signal (generated by auditory filter bank analysis) into *summary measure* index. For a strongly periodic frame, the *summary measure* will results in clusters across fundamental frequency and its integer multiplies; whereas for a strongly a-periodic (unvoiced) frame, the summary measure will result in dips that are randomly scattered over the range of the possible lag values with no prominent clusters. Fig. 3, presents *summary measure* for a strongly periodic and a strongly a-periodic frame. The use of auditory filter bank causes the *summary measure* to be an appropriate index which successfully reduces the effect of the non-dominant interfere in the final feature vector. However, since we expect simultaneous speakers to utter different sentences, we are sure that their corresponding *summary measure* would be absolutely different.

B. system based features

we have used several source based features such as formant frequencies, MFCCs and PLPCCs. The frequencies of the formants during sonorant sounds provide information about the length and shape of the vocal tract. Formants are the peaks of the spectral envelope of speech frames, which is calculated as frequency response of Linear Prediction Coefficients of order 12-14. Generally, F1 and F2 vary considerably due to the vowel being articulated, whereas F3 and F4 change very little [5].

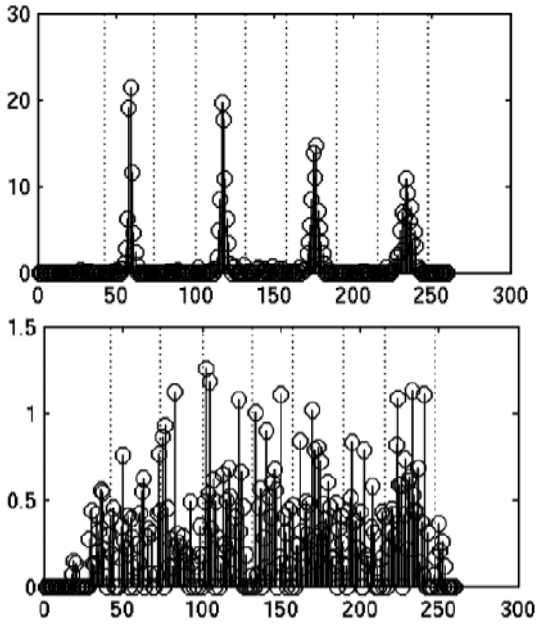

Fig. 2. The resultant *summary measure* for (up) a strongly periodic frame where it consists of strong clusters around fundamental frequency and its integer multiplies and (down) an strongly a-periodic frame where weak dips are randomly scattered over the range of all possible lag values [6].

Another feature set that represents the filter characteristics of the source-filter model is the MFCC feature set. The mel frequency scale is used in order to mimic the cochlear filtering processes in the ear which places more emphasis on certain frequencies. This emphasis is done by frequency wrapping of logarithmic spectrum to the mel scale. The reference point of the mel scale is at a tone of 1000Hz. Hereafter, the mel intervals become logarithmically distributed. The mel scale was experimentally derived by measuring the difference between a linear frequency scale and the perceived pitch that human listeners registered during a series of tests [7]. The MFCCs implicitly code the vocal tract information along with some information about the source and they are the most popular feature set in speaker identification systems [5].

An additional feature set that implements an approximation to the human auditory system is the Perceptual Linear Prediction Cepsteral coefficients [7]. PLP

analysis is a combination between spectral analysis and linear prediction analysis and gives rise to modified autocorrelation coefficients, that correspond to the LPC analysis coefficients. PLP analysis consists of a preprocessing stage that not only warps the speech segments power spectrum to the Bark scale, but also applies other auditory approximations to obtain a more precise modeling of the processes in the ear such as equal loudness curve and power intensity normalization. In addition, the warping implemented here is done prior to the derivation of the AR coefficients and thus the input to the linear prediction analysis is speech that is already modified so that it contains perceptually significant information. This approximation to the biological processes that are executed in the human ear and the consequent smoothing of the spectrum is proven to be helpful in the effective discernment between different speakers. In fact, PLP is reported to be more robust against noisy observations [7].

IV. Simulation Results

In this section we compare the feasibility of using each feature set for resolving permutation ambiguity, on a large database of different speech sources. We use ROOMSIM [10] toolbox in order to simulate the configuration of 2 sources and two sets of microphone arrays as shown in Fig. 3. ROOMSIM provides us with the impulse responses between each source and microphone pairs. First, We apply the BSS algorithm of [8] to the observed mixtures of two speech signals in order to obtain the separation system impulse responses; then, we apply this exact separation system to all the other mixtures generated by different pairs of speech signals which are randomly selected from the database. The reason for using these fixed impulse responses for all of the source pairs is to test them under equal conditions. Table I, summarizes the achieved SIR gains of obtained separating system.

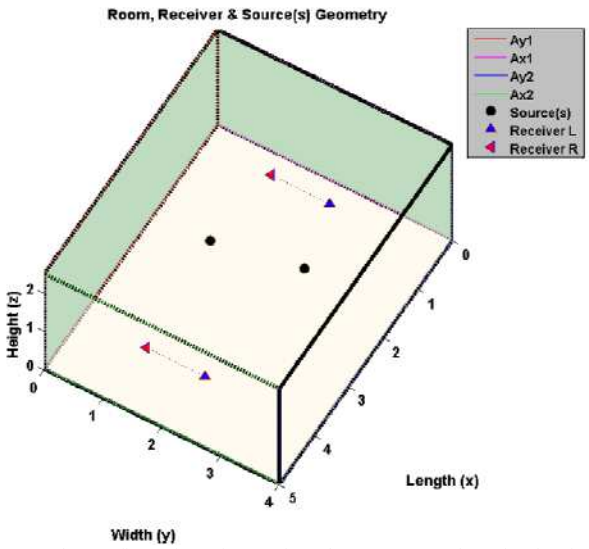

Fig. 3. The configuration of two microphone arrays and a pair of active speakers simulated using RoomSim [10].

TABLE I
THE SIR GAINS FOR THE USED SEPARATING SYSTEM

Array No. 1	SIR CH1	10.04 dB
	SIR CH2	8.83 dB
Array No. 2	SIR CH1	7.41 dB
	SIR CH2	13.12 dB

A comparative table of obtained SIR for different BSS algorithms is presented in [11] which implies that the reported SIR gains in table I are reasonable values for testing the ability of feature sets to resolve permutation ambiguity. Our database of speech signals consists of 15 male and 15 female voices of length 3 seconds.

The frame length is chosen to be 30 milliseconds which is equal to the stationary period of speech signal and is an appropriate choice for periodic/a-periodic analysis of frames. We use *summary measure,* 13^{th} order MFC and PLP coefficients and the first four formants: F1, F2, F3 and F4 as feature sets. Also the cross correlation decision making procedure introduced in [4] is used to compare its results with feature based ambiguity resolver for 250 different signal pairs which are randomly chosen from the database.

Table II summarizes the averages precent of correct decisions for each one of these feature sets. Also in order to evaluate the stability of each feature set, the standard deviation of correct decisions is reported as well along with the total number of complete failures; by failure we mean the vote below 50% which implies the wrong choice of permutations.

TABLE II
THE RESULTED AVERAGE DECISION STATISTICS.

Feature set	Correct votes	Standard deviation	Number of failures
Cross correlation [4]	59%	10.4%	24
MFCC	80.6%	6.2 %	0
PLPCC			
Formants	62.41%	6.25%	4
Summary measure	87.2%	4.47%	0

Table II, reveals that the most robust feature set is the *summary measure*. This is because this measure explicitly tracks dynamics of speech signal which are mostly caused by source properties of speech production model. Also, this measure successfully reduces the effect of mask (interferer) signal by using auditory filter bank analysis.

The same statement could be made for MFCC and PLPCC feature sets, since they also employ human like pre-processing. Although, they contain the information about both source and the vocal tract properties and this might be the reason for their suboptimal results compared to *summary measure* which solely represents source dynamics. In fact, the vocal tract filter properties (the spectral envelope of frames) are very likely to be overridden by frequency response of too many FIR filters (of mixing and unmixing systems) that are on the way of BSS

outputs. So we don't expect system based features to be as distinctive as they are for pure speech signals.

Specially, the obtained results of formant feature set proves this rationalization. Also, The cross correlation decision making obtained the worst results and we can say that it's an completely inappropriate criteria for permutation ambiguity resolving. Practically, we observed that this criteria only works well when the mixed speeches are as non overlapping as possible in time domain. When high intensity portions of two speech signals occur simultaneously, this measure is very likely to fail.

We must also mention the computational complexity of computing each feature set. Table III reports the average runtime for accomplishment of one decision (for one 30 ms frame). It reveals that, despite the very good performance of *summary measure* it's runtime is several times of the other feature sets. If the runtime is an important design parameter, then one could prefer to use PLPCC or MFCC feature sets which are almost as distinctive as *summary measure* for our problem in hand.

TABLE III
AVERAGE RUNTIME FOR ACCOMPLISHMENT OF ONE DECISION.

Feature set	Average runtime (seconds)
Cross correlation [4]	0.0052
MFCC	0.0077
PLPCC	0.0071
Formants	0.03
Summary measure	4.95

Also, another experiment was done for 3×3 scenario. 300 different combinations of 3 speakers was chosen from the database. Under the same conditions as previous experiment, the generalization of the decision logic in section IV was used to solve permutation ambiguity. Table VI summarizes the averages precent of correct decisions for each one of the feature sets.

TABLE VI
THE RESULTED AVERAGE DECISION STATISTICS IN 3×3 SCANARIO.

Feature set	Correct votes	Standard deviation	Number of failures
Cross correlation [4]	49%	7.1%	130
MFCC	72.93%	7.73 %	1
PLPCC	67.52%	6.48%	2
Formants	33.5%	5.4%	295
Summary measure	50.7%	7.16%	130

Note that, since the number of possible permutations are increased from 2 in 2×2 scenario to 6 in 3×3 scanario, we expect the increase of false decisions. As table VI suggests, formants and cross correlations are completely wrong in their decisions because of their less than 50% average vote. Also, despite the very good performance of summary measure in 2×2 scenario, it demonstrates very poor results in 3×3 case. But MFCC and PLPCC coefficients still show acceptable majority votes.

V. CONCLUSION

In order to solve global permutation ambiguity of moderately separated outputs of BSS, we used several speaker specific features of speech We employed some of the most popular feature sets which are used in system identification systems. Among them, 2×2 scenario the *summary measure* over performed other filter based features such as MFCC, PLPCC and formants, but if we take execution runtime into account, we may say that MFCC and PLPCC are the most preferred ones. But in 3×3 scenario, summary measure completely fails in making correct decisions just like the formants and cross correlations. However, PLPC and MFCC features are still reliable for solving the ambiguity problem.

VI. REFERENCES

[1] H. Buchner, R. Aichner, J. Stenglein, H. Teutsch, and W. Kellermann, "Simultaneous localization of multiple sound sources using blind adaptive MIMO filtering," in Proc. ICASSP, Mar. 2005, vol. 3, pp. 97–100.

[2] E. Jan yand J. Flanagan, "Sound Source Localization in Reverberant Environments using an Outlier Elimination Algorithm," 1998.

[3] R. Shimoyama and K. Yamazaki, "Multiple acoustic source localization using ambiguous phase differences under reverberant conditions," Acoust. Sci. & Tech. 25, 6 , july 2004.

[4] A. Lombard, H. Buchner, "Multidimensional Localization of Multiple Sound Sources Using Blind Adaptive MIMO System Identification", 2006 IEEE International Conference on Multi-sensor Fusion and Integration for Intelligent Systems," September 2006, Heidelberg, Germany.

[5] C. Y. Espy-Wilson, S. Manocha and S. Vishnubhotla, "A New Set of Features for Text-Independent Speaker Identification," Institute for Systems Research and Dept. of Electrical & Computer Engineering, 2004.

[6] O. Deshmukh, C.Espy-Wilson, A. Salomon & J. Singh, "Use of Temporal Information: Detection of the Periodicity, Aperiodicity and Pitch in Speech," IEEE Trans. on Speech and Audio Proc., vol.13, pp. 776 – 786, September 2005.

[7] J. Campbell, ``Speaker recognition: A tutorial," Proc. of the IEEE, vol. 85, pp. 1437--1462, Sept 1997.

[8] K. Kokkinakis and A. K. Nandi, "Multichannel blind deconvolution for source separation in convolutive mixtures of speech," IEEE Trans. Audio, Speech, Lang. Process., Vol. 14, No. 1, pp. 200-212, Jan. 2006.

[9] H. Buchner, R. Aichner, and W. Kellermann. A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. IEEE Trans. Speech Audio Processing, 13(1):120–134, Jan. 2005.

[10] D. Campbell, K. Palomäki, and G. Brown, "A MATLAB. Simulation of "Shoebox" Room Acoustics for use in Research and Teaching , June 2005.

[11] M. Syskind Pedersen, J. Larsen, U. Kjems, and L. C. Parra, "A survey of convolutive blind source separation methods," in Springer Handbook on Speech Processing and Speech Communication, 2006.

[12] Maïa E.M. Weddin, "Speaker Identication for Hearing Instruments," Master's Thesis, Denmarks Technical University, March 2005.