PERMUTATION ALIGNMENT OF FREQUENCY-DOMAIN ICA BY THE MAXIMIZATION OF INTRA-SOURCE ENVELOPE CORRELATIONS

J. Nikunen, T. Virtanen, P. Pertilä

Tampere University of Technology
Korkeakoulunkatu 1, 33720 Tampere, Finland

M. Vilermo

Nokia Research Center
Visiokatu 1, 33720 Tampere, Finland

ABSTRACT

This paper presents a novel method for solving the permutation ambiguity of frequency-domain independent component analysis based on source signal envelope correlation maximization. The proposed method is developed for blind source separation with high sampling frequency and significant spatial aliasing. We propose a method that analyzes the source envelope using a rank-one singular value decomposition (SVD) applied to an initial source magnitude spectrogram obtained by a time difference of arrival (TDoA) based permutation alignment method. The permutation for frequencies with incoherent TDoA are corrected by maximizing the cross-correlation of the SVD analyzed source activation vector and each independent component magnitude envelope. We evaluate the separation quality using real high sampling frequency speech captures and the proposed method is found to improve the separation over the baseline algorithm.

Index Terms— Blind Source Separation, Independent Component Analysis

1. INTRODUCTION

The blind source separation (BSS) of simultaneously emitting sound sources, generally known as the cocktail party problem, has been intensively studied over the years, but is however still categorized as an unsolved problem. In the course of this paper we pursue blind separation of high sampling frequency speech using independent component analysis (ICA) applied in frequency domain leading into frequency-wise permutation ambiguity. The permutation alignment have been previously solved for example based on mixing filter frequency response smoothness [1], temporal structure of the source signals [2], and time-difference of arrival (TDoA) and direction of arrival (DoA) [3, 4] interpretation of ICA mixing parameters. The latter can be considered as generally robust with no assumptions on the source characteristics, however their performance starts to degrade in reverberant conditions and with captures involving lot of spatial aliasing frequencies.

In this paper we propose a novel method for ICA permutation alignment that resolves the component ordering via maximization of intra-source envelope correlations. TDoA based algorithm [4] is used for obtaining an initial solution for the ICA parameter alignment which is to be improved by the proposed method. The proposed algorithm is applied for frequencies where the source TDoA is incoherent due spatial aliasing and reverberation making the source magnitude envelopes more accurate method for permutation alignment. The separation quality of the proposed method is evaluated using high sampling frequency speech captures and the results show an increase in separation quality measured using quantities proposed in [5].

The rest of the paper is organized as follows, in Section 2 we review the frequency domain ICA and the permutation alignment algorithms used in prior art. The proposed method is presented in Section 3. In Section 3.1 we shortly present the TDoA based permutation algorithm [4] used for obtaining an initial permutation solution. The proposed singular value decomposition (SVD) based source envelope analysis and the permutation alignment by maximization of intra-source envelopes is presented in Section 3.2. The source separation quality of speech samples is presented in Section 4.

2. BLIND SOURCE SEPARATION AND INDEPENDENT COMPONENT ANALYSIS

The array capture can be considered by the following convolutive mixture model in the time-domain

$$x_m(t) = \sum_{j=1}^{J} \sum_{\tau} h_{mj}(\tau) s_j(t-\tau) \quad (1)$$

where $x_m(t)$ is the mixture of $j = 1...J$ source signals capture by sensor $m = 1...M$ and sampled in time instances t. The spatial response from the source j to the sensor m is denoted by $h_{mj}(\tau)$ and the source signals are given as $s_j(t)$. Convolutive model (1) is usually approximated by instantaneous mixing in frequency domain as

$$\mathbf{x}(f,n) = \sum_{j=1}^{J} \mathbf{h}_j(f) s_j(f,n) \quad (2)$$

where $\mathbf{x}(f,n) = [x_1, \ldots, x_M]^T$ is the short-time Fourier transform (STFT) of the array capture $x_m(t)$, $f = 1...F$ is

the frequency index and $n = 1...N$ is the frame index. The impulse response $h_{mj}(\tau)$ is replaced with the frequency response denoted by $\mathbf{h}_j(f) = [h_{1j}, \ldots, h_{Mj}]^T$ and the STFTs of source signals are denoted by $s_j(f, n)$.

The ICA applied to the frequency domain model (2) has been successfully used for determined BSS [1, 2, 3, 4] where $M \geq J$. ICA is applied separately for each frequency bin f to obtain $J \times M$ unmixing matrix $\mathbf{W}$ as in

$$\mathbf{y}(f, n) = \mathbf{W}(f)\mathbf{x}(f, n). \tag{3}$$

where $\mathbf{y}(f, n) = [y_1, \ldots, y_J]^T$ corresponds to the sources $s_j(f, n)$ with an arbitrary permutation of sources indices at each frequency f. Further we assume that the unmixing matrix is invertible and define $\mathbf{A}(f) = \mathbf{W}(f)^{-1}$, thus we can write the ICA model as,

$$\mathbf{x}(f, n) = \mathbf{A}(f)\mathbf{y}(f, n) \tag{4}$$

If $J < M$, the mixing matrix is obtained via Moore–Penrose pseudoinverse $\mathbf{A}(f) = \mathbf{W}(f)^+$. $\mathbf{A}$ is constructed of column vectors $[\mathbf{a}_1, ..., \mathbf{a}_J]$ and each vector denotes the response of single source j to the each capturing sensor $m = 1, \ldots, M$.

In the earliest frequency-domain ICA based BSS methods [1] the permutation alignment was solved by assuming a smooth frequency response of the mixing filters $\mathbf{h}_j(f)$. Later in [2] the temporal structure of the separated signals $\mathbf{y}(f, n)$ was considered and the permutation was solved by maximizing cross-correlation of magnitudes of neighboring frequencies. TDoA and DoA interpretation of component bases $\mathbf{a}_j(f)$ has been proposed in [3] and in [6] the TDoA approach was combined with the magnitude envelope correlation maximization. More recently a method only relying on anechoic source signal propagation model estimation was proposed in [4], which will be used as a baseline in this paper.

There also exists ICA-based methods that unify the source dependencies across frequencies, independent vector analysis [7] and recursively regularized ICA across frequencies [8]. In this paper we will concentrate only to the frequency bin-wise ICA model (3) and improving the permutation alignment in case of high sampling frequency captures and severe spatial aliasing over the baseline [4]. Other related work combining TDoA with envelope correlation maximization include for example [6, 9].

3. PROPOSED METHOD

The proposed method for ICA permutation alignment combines a TDoA based algorithm [4] with a novel source envelope analysis by rank-one SVD and source temporal activity cross-correlation maximization across frequencies. With the proposed algorithm we aim for improving performance of TDoA based algorithms with high sampling frequency captures by using source magnitude envelope information in the permutation alignment.

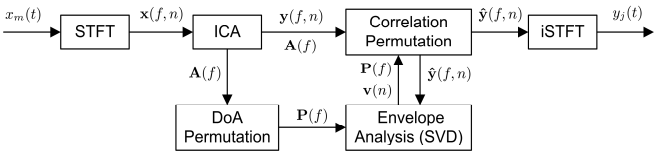

Fig. 1. Block diagram of the frequency domain ICA and the proposed permutation alignment by source envelope cross-correlation.

The block diagram of the proposed method is illustrated in Figure 1. First the input signal x_m is STFT analyzed to get $\mathbf{x}(f, n)$. The ICA is applied for each frequency f separately to obtain mixing matrix $\mathbf{A}(f)$ and the source signals $\mathbf{y}(f, n)$. The mixing matrix entries $\mathbf{a}_j(f)$ are clustered using a TDoA permutation alignment algorithm to get an initial permutation matrices $\mathbf{P}(f)$. The source signals $\mathbf{y}(f, n)$ are aligned using the initial permutations to obtain $\hat{\mathbf{y}}(f, n)$ and the source envelopes are analyzed using rank-one SVD. The obtained source envelope $\mathbf{v}(n) = [v_1, \ldots, v_J]^T$ is used for finding the permutation that maximizes the cross-correlation with $|\hat{\mathbf{y}}(f, n)|$ at each frequency f. The SVD envelope analysis and cross-correlation matching is repeated until no changes are made for $\hat{\mathbf{y}}(f, n)$ The time domain source signals are obtained via inverse STFT.

3.1. Permutation Alignment by Signal Propagation Model

The initial alignment of separated components is obtained by algorithm presented in [4], which is shortly reviewed in this section. The algorithm provides initial magnitude spectrogram matrices $|\hat{\mathbf{y}}(f, n)|$ in order to be SVD analyzed and corrected by the proposed algorithm presented in Section 3.2.

The parameters $\mathbf{a}_j(f)$ are phase and amplitude normalized with respect to a chosen reference sensor by subtracting the reference sensor phase and dividing by its norm, result is denoted by $\tilde{\mathbf{a}}_j(f)$. Normalization gives the relative TDoA of the mixing parameters in terms of phase difference with respect to the reference sensor. The source propagation model is defined as

$$\hat{h}_{mj}(f) = \lambda_{mj} \exp(-i2\pi f \tau_{mj}) \tag{5}$$

which approximates the mixing filter frequency response $h_{mj}(f)$ by having a fixed time delay τ_{mj} and attenuation λ_{mj} from source j to each capturing sensor m over all frequencies. The propagation model (5) translates into a fixed spatial position in means of TDoA in anechoic conditions, which further can be viewed as DoA estimate of the source.

The permutations are solved by minimizing the cost function

$$D = \sum_{j=1}^{J} \sum_{f=1}^{F} ||\tilde{\mathbf{a}}_{P_f(j)}(f) - \hat{\mathbf{h}}_j(f)||^2 \tag{6}$$

where the permutation of $\tilde{\mathbf{a}}_j(f)$ for each frequency f is given by $P_f(j)$ and the propagation model (5) is given in vector

form $\hat{\mathbf{h}}_j(f)$. The permutation alignment and the propagation model estimation is solved simultaneously and the correct permutations depend on the accuracy of the estimated propagation model. With no further algorithm details we assume to obtain the permutation matrix $\mathbf{P}(f)$ for changing the rows of $\mathbf{y}(f,n)$ and estimated propagation model $\hat{\mathbf{h}}_j(f)$ that minimizes the cost function (6). The details of the algorithm can be found from [4].

3.2. Source Envelope Analysis and Cross-correlation Maximization of Magnitude Envelopes

We start the derivation of the proposed algorithm by considering which of the frequency indices after the permutation alignment given in Section 3.1 have high confidence of being correct. These frequency indices are used as a reference for analyzing the source envelopes using a rank-one SVD. The proposed algorithm is applied for correcting the permutation of the rest of the frequency indices.

The confidence of correct permutation at each frequency after TDoA permutation can be extracted by evaluating the following distance measure,

$$D(f) = \sum_{j=1}^{J} ||\tilde{\mathbf{a}}_{P_f(j)}(f) - \hat{\mathbf{h}}_j(f)||^2 \tag{7}$$

and sorting $D(f)$ in ascending order. Choosing the k_R first frequencies, denoted by set $\mathcal{F}_R$, will serve as a reference having the lowest distance to the estimated propagation model $\hat{\mathbf{h}}$. The frequency indices to be corrected by the proposed method are chosen by taking indices $k_Q, \ldots, F$ from the sorted $D(f)$, denoted by set $\mathcal{F}_Q$. These have the most incoherent TDoA and amplitude difference with respect to the estimated propagation model. Note that $\mathcal{F}_R$ and $\mathcal{F}_Q$ can have overlapping frequencies if $k_Q < k_R$.

The confidence measure (7) assumes that the estimation of the propagation model $\hat{\mathbf{h}}_j(f)$ has converged close to the actual spatial position in terms of TDoA and that the anechoic source propagation assumption holds for the observed data. It is shown by an example in Section 4 that the lowest frequencies fit to the model (5) more accurately whereas the ICs at higher frequencies suffer from the spatial aliasing and phase modification by reverberation making the permutation uncertain according to (7).

The permutation matrix $\mathbf{P}(f)$ obtained from the TDoA based alignment is used to change the ordering of rows of vector $\mathbf{y}(f,n)$ to correspond to a single source signal defined as $\hat{\mathbf{y}}(f,n) = \mathbf{P}(f)\mathbf{y}(f,n)$. The magnitude spectrogram matrix of the sources after initial permutation is denoted as $[\hat{\mathbf{Y}}_{(j)}]_{f,n} = |\hat{y}_j(f,n)|$.

The permutation correction algorithm is described as follows. For each source $j = 1...J$ we apply the SVD to the magnitude spectrogram of the sources given as,

$$\hat{\mathbf{Y}}_{(j)} = \mathbf{U}_{(j)}\mathbf{\Sigma}_{(j)}\mathbf{V}^*_{(j)}, \quad f \in \mathcal{F}_R \tag{8}$$

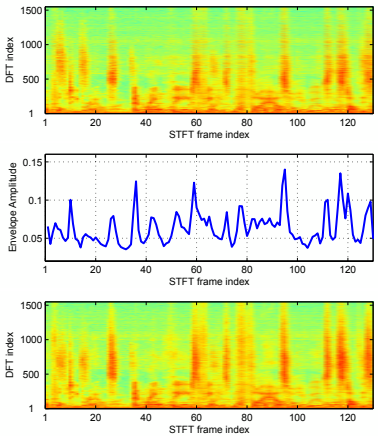

Fig. 2. An example of the SVD analyzed source envelope in the middle, the source magnitude spectrogram after baseline [4] on the top and the source magnitude spectrogram after proposed permutation alignment in the bottom.

where subindices (j) denote the matrix indexing corresponding to each source and each $\hat{\mathbf{Y}}_{(j)}$ is of size $k_R \times N$. To obtain a rank-one approximation of the source magnitude spectrogram we take the singular vectors $\mathbf{U}_{(j)i,:}$ and $\mathbf{V}_{(j)i,:}$ corresponding to the largest singular value $\mathbf{\Sigma}_{(j)i,i}$. The singular vector $\mathbf{U}_{(j)i,:}$ contains the average source spectrum and the corresponding temporal activity is given by $\mathbf{V}_{(j)i,:}$ which we propose to use as the reference source envelope.

The analyzed source envelope for each STFT frame n is hereafter denoted by $\mathbf{v}(n) = [v_1, \ldots, v_J]^T = \mathbf{V}_{(j)i,n}$. The SVD analyzed envelope is assumed to capture quintessential temporal activity features of the source and thus can be used as a reference for aligning permutation for frequencies $f \in \mathcal{F}_Q$ by maximizing the cross-correlation of source magnitudes and $\mathbf{v}(n)$. An example of the SVD analyzed envelope and the source magnitude spectrogram before and after the proposed permutation alignment is illustrated in Figure 2.

The permutation optimization with the obtained source envelopes $\mathbf{v}(n)$ can be defined as

$$\mathbf{P}(f) \leftarrow \text{argmax}_{\mathbf{P}(f)} \sum_{n=1}^{N} \mathbf{v}(n)^T \mathbf{P}(f)\hat{\mathbf{y}}(f,n), \quad \forall f \in \mathcal{F}_Q \tag{9}$$

which equals finding a new permutation matrix $\mathbf{P}(f)$ which maximizes the cross-correlation of $\mathbf{v}(n)$ and source magnitude envelopes $\mathbf{P}(f)\hat{\mathbf{y}}(f,n)$ within the frequency set $f \in \mathcal{F}_Q$. In practice the maximization is implemented by searching through all combinations of $\mathbf{P}(f) : \{1, \ldots J\} \rightarrow \{1, \ldots J\}$ and choosing the one producing largest cross-correlation, this is computationally feasible for low number of sources. As a result we obtain a new permutation matrix $\mathbf{P}(f)$ which is used for aligning the permutations as

$$\hat{\mathbf{y}}(f,n) \leftarrow \mathbf{P}(f)\hat{\mathbf{y}}(f,n) \tag{10}$$

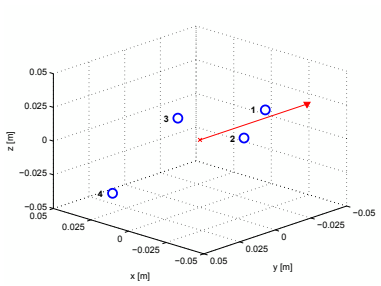

Fig. 3. The capturing array used in the simulations. Microphones are denoted by circles and the zero angle references axis by an arrow.

The experiments with the algorithm has shown that choosing $k_Q < k_R$ produces generally best results regarding the separation quality. In this case the evaluation of permutation optimization given in Equation (9) may also change permutation for frequency indices $f \in F_R$ which further affects on SVD analysis (8). The proposed algorithm is implemented by iteratively evaluating Equations (8) - (10) until no permutation changes are made in (10). With a suitable choice of k_R and k_Q the algorithm usually converges in less than 10 iterations. The choice of k_R and k_Q is discussed in more details in Section 4.

4. EVALUATION OF SEPARATION QUALITY

In this section we evaluate the separation performance of the proposed algorithm against the TDoA based algorithm proposed in [4]. The evaluation consist of real audio captures recorded in following conditions: sampling frequency was 48kHz, the room dimensions were 4.53 $\times$ 3.96 $\times$ 2.59 m and the reverberation time (T60) was approximately 0.26s.

The capturing array consists of four DPA 4060-BM prepolarized omnidirectional miniature condensator microphones.The array dimensions are given in Table 1 and the array geometry with reference axis is illustrated in Figure 3. The spatial aliasing frequency for the given array is 1563 Hz which corresponds to STFT frequency bin $f = 133$.

The test samples used included three male and one female speakers from Librivox audiobook database which were played with Genelec 1029A speakers. The utterance length is 10 seconds. Each speaker was captured separately and signals were combined into mixtures of three simultaneous speakers. The angle of the speakers with respect the reference axis of the microphone array are given in Table 2.

4.1. Implementation Considerations

For the ICA parameter estimation we used the complex-valued version of JADE algorithm [10]. Other parameters were chosen as follows: STFT window length $= 4096$ with 50% window overlap, number of target sources $= 3$. Two

Mic	x (mm)	y (mm)	z (mm)
1	0	-46	6
2	-22	-8	6
3	22	-8	6
4	0	61	-18

Table 1. Geometry of the array used for evaluation. Illustrated in Figure 3.

Identification	Angle
Speaker 1	180°
Speaker 2	90°
Speaker 3	45°
Speaker 4	0°

Table 2. Speaker positions with respect to array zero angle axis.

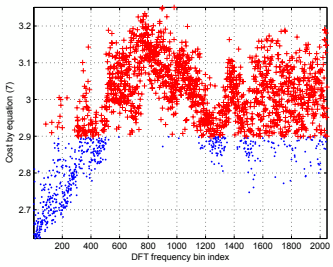

Fig. 4. Cost function (7) for an individual test sample. The reference frequencies $f \in F_R$ are denoted by dots and rest of the cost function entries are denoted by plus-marks.

datasets were used, dataset one consisting of speakers 1, 2 and 4 and dataset two consisting of speakers 2, 3 and 4. The total number of 10-second utterances in both datasets is five. It is shown in Section 4.2 that no separate training stage or development set is needed for the choice of k_R and k_Q due the separation quality not being affected by a wide range of choosing k_R and k_Q. The values for the separation evaluation were chosen as $k_R = 600$ and $k_Q = 300$ producing a good average performance.

An example of the TDoA coherence cost defined by Equation (7) is illustrated in Figure 4. The reference bins chosen are denoted by dots and the rest of the frequencies are denoted by plus-marks. It is clear from the shape of the cost function that the lowest frequencies have the most coherent TDoA regarding the estimated propagation model and are chosen mostly for the reference frequency group $f \in F_R$. Also some higher frequencies fit the model well and serve as a reference.

4.2. Separation Results

The results from the separation quality evaluation using metrics signal-to-distortion ratio (SDR), image-to-spatial distortion Ratio (ISR), signal-to-interference ratio (SIR) and signal-to-artefact ratio (SAR) proposed in [5] are given in Table 3. The measures are averaged over all sources and all utterances. With the proposed method the SDR separation quality increases by 0.72 dB and 0.48 dB in the datasets one and two, respectively. The source interference (SIR) is improved noticeably, the separated source spatial image accuracy (ISR) improves as well and the separation artifacts are decreased

Dataset	Baseline [4]		Proposed	
	1	2	1	2
SDR (dB)	3.88	0.92	4.60	1.40
ISR (dB)	8.92	5.04	10.19	5.63
SIR (dB)	8.37	1.42	9.64	2.93
SAR (dB)	6.24	3.85	6.82	4.43

Table 3. Separation results for datasets one and two.

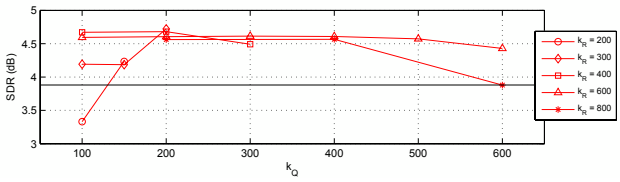

Fig. 5. Separation quality in terms of SDR with different combinations of k_R and k_Q with the dataset one. Solid line denotes the baseline performance.

(SAR).

Each source in dataset one are spatially separated at least by 90° whereas in the dataset two the spatial separation is 45°, which significantly decreases the separation performance. In case of dataset two where the initial separability of the sources is poor the proposed algorithm is still able to improve the average separation of sources, considering the fact that the derivation of the algorithm assumes obtaining a fair initial separation for the envelope analysis.

The effect of the algorithm parameters k_R and k_Q is illustrated in Figure 5 where the SDR separation performance is given with different combinations of k_R and k_Q. The performance of the proposed algorithm is almost equivalent regardless of the choice of the parameters. Only too few reference frequencies $k_R = 200$ degrades the SDR quality below the baseline performance. The results in Figure 5 indicate high robustness towards the choice of the parameters and eliminates the need of a separate training stage.

Temporal activity based permutation alignment algorithms are known to be less efficient with short signals and thus the proposed method was additionally tested with the signals from the test set one split to duration of 2.5 seconds. The average SDR was 2.82 dB and 3.15 dB for the baseline and the proposed algorithm, respectively, indicating improved separation with the proposed method also in such cases.

5. CONCLUSION

In this paper we proposed an algorithm for independent component analysis (ICA) permutation alignment when used for blind source separation (BSS) of simultaneous speakers. The proposed method is based on analysis of source envelopes by rank-one SVD and maximizing the cross-correlations of the analyzed envelope and source magnitude envelopes at each individual frequency. The proposed method is aimed for improving the time difference of arrival (TDoA) based alignment algorithms suffering from spatial aliasing in case of high sampling frequency speech and it was found to improve the separation quality in such conditions.

6. REFERENCES

[1] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1, pp. 21–34, 1998.

[2] J. Anemüller and B. Kollmeier, "Amplitude modulation decorrelation for convolutive blind source separation," in *Proc. of ICA*. Helsinki, Finland, 2000, pp. 215–220.

[3] M.Z. Ikram and D.R. Morgan, "A beamforming approach to permutation alignment for multichannel frequency-domain blind speech separation," in *Proc. of ICASSP*, 2002, pp. 881–884.

[4] 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 ASLP*, vol. 15, no. 5, pp. 1592–1604, 2007.

[5] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[6] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 5, pp. 530–538, 2004.

[7] I. Lee, T. Kim, and T.W. Lee, "Independent vector analysis for convolutive blind speech separation," *Blind speech separation*, pp. 169–192, 2007.

[8] F. Nesta, P. Svaizer, and M. Omologo, "Convolutive bss of short mixtures by ica recursively regularized across frequencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 624–639, 2011.

[9] F. Nesta, T.S. Wada, and B.H. Juang, "Coherent spectral estimation for a robust solution of the permutation problem," in *in Proc. of WASPAA*, 2009, pp. 105–108.

[10] J.F. Cardoso and A. Souloumiac, "Blind beamforming for non-gaussian signals," *IEE Proc-F. Radar and Signal Processing*, vol. 140, no. 6, pp. 362–370, 1993.