Independent Vector Analysis Exploiting Pre-Learned Banks of Relative Transfer Functions for Assumed Target's Positions

Jaroslav Čmejla, Tomáš Kounovský, Jiří Málek and Zbyněk Koldovský

Acoustic Signal Analysis and Processing Group,
Faculty of Mechatronics, Informatics and Interdisciplinary Studies,
Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic
jaroslav.cmejla@tul.cz

Abstract. On-line frequency-domain blind separation of audio sources performed through Independent Vector Analysis (IVA) suffers from the problem of determining the order of the separated outputs. In this work, we apply a supervised IVA based on pilot components obtained using a bank of Relative Transfer Functions (RTF). The bank is assumed to be available for potential positions of a target speaker within a confined area. In every frame, the most suitable RTF is selected from the bank based on a criterion. The pilot components are obtained as pre-separated target and interference, respectively, through the Minimum-Power Distortionless Beamforming and Null Beamforming. The supervised IVA is tested in a real-world scenario with various levels of up-to-dateness of the bank. We show that the global permutation problem is resolved even when the bank contains only pure delay filters. The Signal-to-Interference Ratio in separated signals is mostly better than that achieved by the pre-separation, unless the bank contains very precise RTFs.

Keywords: independent vector analysis, relative transfer function, source separation, speech enhancement

1 Introduction

Frequency-Domain Independent Component Analysis (FD-ICA) [18] provides an effective tool for audio signal separation and enhancement. It is an unsupervised method where each frequency band is treated separately as an instantaneous mixture. This causes the permutation problem as the order of separated frequency components is random [17]. Rather than solving the separation in two

This paper was supported by The Czech Science Foundation through Project No. 17-00902S and partly supported by the Student Grant Scheme 2018 project of the Technical University in Liberec and by the United States Department of the Navy, Office of Naval Research Global, through Project No. N62909-18-1-2040.

steps, i.e., by applying FD-ICA and some depermutation method afterwards, a fast and effective solution is provided through Independent Vector Analysis (IVA). Here, all frequencies are separated jointly; the separated sources should be independent while frequency components corresponding to the same source are forced to be as dependent as possible [5, 8, 15].

The random global order of separated sources is the remaining problem of IVA. Classical solutions impose a constraint on the de-mixing filters obtained through IVA. For example, the filters are constrained to remain close to the pure delay filters where the delays correspond to the expected Directions of Arrival (DOA) of the sources [4]. Unconstrained modifications of IVA have also been proposed exploiting prior or side information. For example, a priori knowledge of temporal power variations of sources is used in [16]. In [9], prior knowledge of target positions was used to initialize the IVA, resulting in faster convergence and known permutation of the targets in the de-mixed signals. This partly solves the global permutation problem, but only when the sources remain in static locations.

A general formulation, referred to as Supervised IVA (S-IVA), has been recently proposed in [13] where higher-order dependencies between so-called pilot components and the separated signals are used. For example, the outputs of a voice activity detector (VAD) and of a video speech detector (VSD) were used as the pilots in [14] to distinguish speakers. In this paper, we propose a cheaper solution relying purely on audio. It is assumed that the position of a targeted speaker is confined to a limited area and that a bank of Relative Transfer Functions (RTFB) for some possible positions of the speaker is available. This bank can be directly used for separation as in [6]. However, it is more realistic to assume that the bank is not that up-to-date due to various changes (variations of acoustic conditions, rotations of the target speaker, new locations, etc.) so that its direct application yields a limited separation accuracy. Therefore, we propose to use the bank for pre-separating the target from the interference and to use these outputs as pilots in the supervised IVA.

The paper is organized as follows. In Section 2, S-IVA and a corresponding algorithm are briefly described. In Section 3, the concept of the RTFB and its deployment for obtaining pilot components for S-IVA are proposed. Section 4 is devoted to experiments with real-world on-line separation where S-IVA is compared with the original IVA and with the beamforming-based separation relying purely on the RTFB. Section 5 concludes the paper.

2 Blind Separation Using Supervised IVA

2.1 Problem definition

In this paper, we will constrain to situations where two source signals are recorded by two microphones. Let S_n^k and X_m^k be the Short-Term Fourier Transform (STFT) coefficients of the nth source and the mth microphone, respectively, where k is the frequency bin index. The source and the mixture vector will be denoted, respectively, as $\mathbf{S}^k = [S_1^k, S_2^k]^T$ and $\mathbf{X}^k = [X_1^k, X_2^k]^T$. The mixing model within the kth frequency bin reads

$$\mathbf{X}^k = \mathbf{H}^k \mathbf{S}^k + \mathbf{V}^k, \tag{1}$$

where $\mathbf{H}^k$ is the mixing matrix. The objective of IVA is to jointly estimate the set of de-mixing matrices $\{\mathbf{W}^k\}_{k=1,\ldots,K}$; K is the number of frequency bins; see, e.g., [5]. The vector of the nth separated source will be denoted by $\mathbf{Y}_n = [Y_n^1, \ldots, Y_n^K]$ where

$$Y_n^k = \sum_{m=1}^{2} W_{nm}^k X_m^k, \quad k = 1, \ldots, K. \tag{2}$$

Each separated source corresponds to one of the two original sources up to the scaling ambiguity, which we subsequently resolve using Minimal Distortion Principle [11].

2.2 Supervised IVA using Natural Gradient

The supervised IVA (S-IVA) is based on a joint statistical model of the frequency components corresponding to a source and of additional pilot components, because all these components are assumed to be dependent [13]. For simplicity, we assume only one pilot component, for the nth source, denoted by P_n. As in [8], the multivariate super-Gaussian distribution is used for modeling the joint pdf of $\mathbf{Y}_n$ and of P_n, that is,

$$f(\mathbf{Y}_n, P_n) \propto \exp\left(-\sqrt{\sum_{k=1}^{K} |Y_n^k|^2 + |P_n|^2}\right). \tag{3}$$

The log-likelihood function for the joint estimation of $\{\mathbf{W}^k\}_k$ is given by

$$\mathcal{L}(\{\mathbf{W}^k\}_k) = \sum_{k=1}^{K} \log|\det \mathbf{W}^k| + \sum_{n=1}^{N} \mathrm{E}[\log f(\mathbf{Y}_n, P_n)]. \tag{4}$$

which is maximized using the natural gradient-based learning rules

$$\begin{aligned} \mathit{\Delta} W_{nm}^k &= (I_{nm} - \mathrm{E}[\phi^k(\mathbf{Y}_n, P_n)(Y_m^k)^*]) W_{nm}^k, \\ \mathbf{W}_{\text{new}}^k &= \mathbf{W}_{\text{old}}^k + \eta \mathit{\Delta} \mathbf{W}^k, \end{aligned} \tag{5}$$

where η is the step length, I_{nm} is the nmth element of the identity matrix, and $\phi^k = -\partial/\partial Y_m^k(\log f)$, $k = 1, \ldots, K$, are the score functions related to (3). In practice, we use ad hoc modifications of the score functions given by

$$\phi^k(\widetilde{\mathbf{Y}}_n) = \frac{Y_n^k}{\sqrt{(1-\beta_n)\sum_{k=1}^{K} |Y_n^k|^2 + \beta_n |P_n|^2}}. \tag{6}$$

where the hyper-parameter $\beta_n \in (0, 1)$ controls the influence of P_n. In (5), the expectation value is either approximated by the average taken over frames or by the instant value in case of on-line processing.

3 Utilization of the Bank of RTFs

3.1 Bank of Relative Transfer Functions

Given a pair of microphones (we will denote them L and R for left and right), the mixing model (1) can be re-written with respect to one particular (target) source, from here denoted by S, and with respect to the left microphone as

$$\begin{aligned} X_L^k &= S^k + V_L^k \\ X_R^k &= G^k S^k + V_R^k. \end{aligned} \tag{7}$$

S^k denotes the spatial image of the target source on the left microphone, and V_L^k and V_R^k involve the contributions of the other source (in practice also of noise). G^k is the Relative Transfer Function (RTF) related to the microphone pair and to the target source.

Although several methods exist that can estimate G^k from noisy mixtures [2], they can hardly achieve the accuracy of noise-free estimates. These can be obtained when a sufficiently long noise and interference-free interval of recording is available. However, the RTF estimate remains accurate only for the given position of the source. In order to cover the area of the most probable target source occurrence, a bank of RTFs (RTFB) was assumed to be available in [7] such that the RTFs in the bank correspond to several potential target's positions within the confined area. It was assumed that such a bank was prepared in advance during noise-free periods. Then, it can be used in dynamical noisy situations when the target performs movements within the assumed area.

Specifically, in every processing frame, null beamforming using all RTFs can be performed. The RTF corresponding to the null beamformer yielding output with the lowest L_p norm is then selected as the most fitting solution [10]. Since we assume that both target and interference are speech signals, the Null Beamformer using the correct RTF should notice an increased sparsity on its output. Therefore, the value of p is chosen to be $p \leq 1$. Several other methods for selecting the best RTF from the RTFB have also been proposed; see, e.g., [6, 12].

3.2 Pre-separation Using Beamformers

Let us assume for now that G^k is known. We now describe simple approaches for obtaining separated signals of the target and interference. To obtain the target, we can apply a minimum power distortionless beamformer (MPDR) whose output is given by

$$\hat{S}^k = \left(\frac{(\mathbf{C}_x^k)^{-1}\mathbf{u}^k}{(\mathbf{u}^k)^H (\mathbf{C}_x^k)^{-1}\mathbf{u}^k} \right)^H \mathbf{X}^k, \tag{8}$$

where $\mathbf{u}^k = (1, G^k)^T$, and $\mathbf{C}_x^k$ is the covariance matrix of $\mathbf{X}^k$; the superscript ^{H} denotes the conjugate transpose. In the on-line processing regime, $\mathbf{C}_x^k$ has to be estimated in a recursive way as

$$\mathbf{C}_x^{k,\ell} = \lambda \mathbf{C}_x^{k,\ell-1} + (1-\lambda)\mathbf{X}^k(\mathbf{X}^k)^H, \tag{9}$$

where ℓ stands for the frame index.

Next, a signal containing only the interference can be obtained through blocking the target signal (null beamforming). Specifically, the reference signal is obtained as

$$Z^k = G^k X_L^k - X_R^k = G^k V_L^k - V_R^k, \tag{10}$$

which involves only V_L^k and V_R^k.

3.3 Pilot Component Definition

The performance of the beamforming approaches highly depends on the accuracy of the RTFs in the RTFB. To achieve optimum separation, the RTFs must be up-to-date with respect to changes of the acoustic environment, the RTFB should cover the entire area of possible target's positions, and the time domain length of the RTFs must be sufficiently long with respect to reverberation. Since these requirements are hardly met in practice, it is better to take into account a limited performance of the beamforming methods.

It is more realistic to assume that the separated signals $\hat{S}^k$ and Z^k are only dominated by the target and interference, respectively. Then, we propose to exploit these signals as pilots within S-IVA, which might finally achieve better separation. Thus, the pilot components are defined as

$$P_1 = \sum_{k=1}^{K} |\hat{S}^k|, \quad \text{and} \quad P_2 = \sum_{k=1}^{K} |Z^k|, \tag{11}$$

for the target and the interference output, respectively.

4 Experiments

In this section, we present results of experiments whose goal is to demonstrate the influence of the accuracy of the RTFB on the solution of the global permutation by S-IVA, and to compare the separation accuracies achieved though S-IVA and the beamforming methods from Section 3.2.

4.1 Scenario

The experimental setup is illustrated in Fig. 1. Two speakers (simulated by loudspeakers) recorded by two microphones with mutual distance of 18 cm are considered in a room with reverberation time $T_{60} = 700$ ms. The target source is located within a 15×15 cm area that is located approx. 1 m in front of the microphones. The area is covered by a regular grid of 16 positions with inter-grid distance of 5 cm, for which the RTFB is prepared using noise-free training recordings played from these exact positions. For the experiment, a testing recording is obtained when the target loudspeaker is randomly moved within the area.

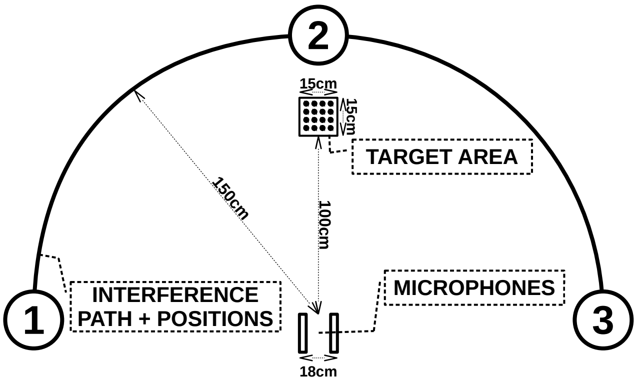

Fig. 1. The illustration of the experimental setup

The interference is represented by another loudspeaker which is moving between 0°through 180°around the microphones at the distance of 1.5 m. The interference moves through major positions, denoted in Fig. 1, in the order 1, 2, 3, 2, 1, which is repeated several times. The interference source stopped at each major position for about 20 seconds.

The utterances played by the loudspeakers were taken from the TIMIT dataset [3]; the length of the entire testing recording is about 8 minutes; the sample rate is 16 kHz. Target and interference were recorded separately and mixed afterwards at the initial global Signal-to-Interference Ratio (SIR) of 0 dB.

The on-line S-IVA algorithm was used to separate the sources in the STFT domain with the frame length of 4096 samples and 75% frame overlap. The separated signals were reconstructed in the time domain by the overlap-add method. To improve the convergence of S-IVA, the scaled natural gradient modification of (5) described in [1] was used.

For evaluation, the improvement of SIR (iSIR) is computed on each frame and averaged over microphones. This gives us an improvement in SIR for all separated sources. Averages of these results are used as a measure of separation quality.

4.2 Results

The following notation is used for all figures. MPDR denotes the separation provided by the combination of the MPDR and Null beamforming. DOA setting indicates that the RTFs are pure delay filters. S-IVA followed by the specification of the hyper-parameters, β_n, denotes the proportion of piloting by the outputs of the beamformers (11). "S-IVA oracle" corresponds to the S-IVA piloted by original (oracle, separated) signals. IVA indicates the original unsupervised IVA algorithm.

Fig. 2 shows the per-frame performance of the above-mentioned methods. It contains results for two different settings of the MPDR: RTFs are set to be

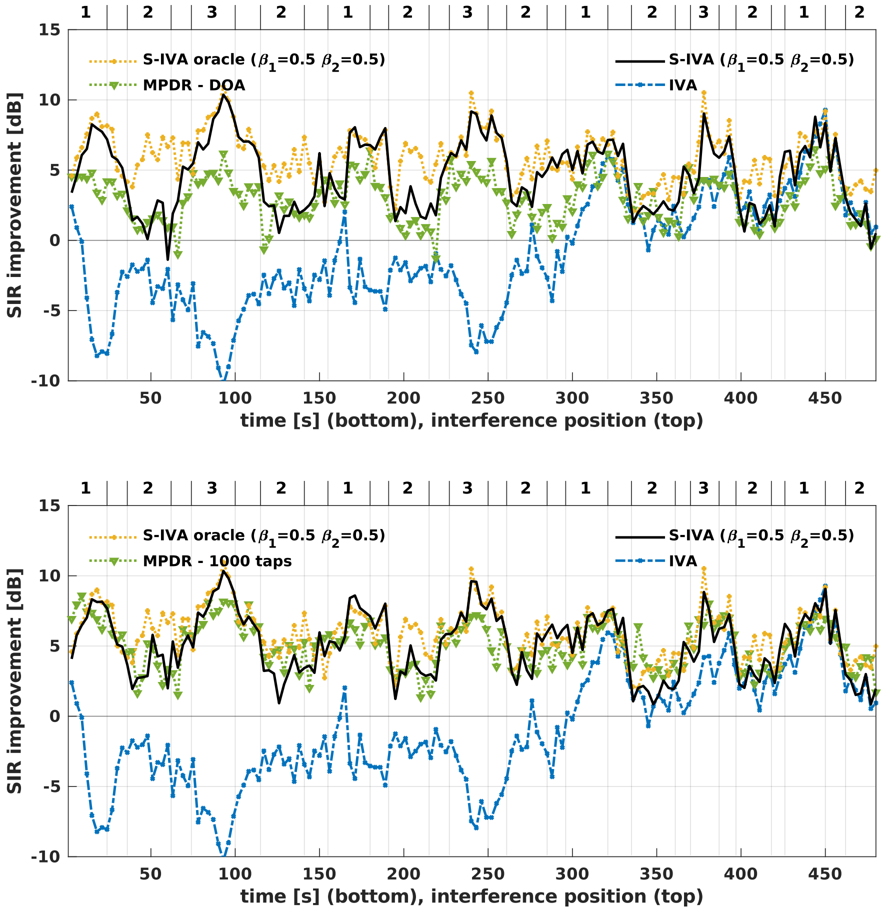

Fig. 2. Results in terms of iSIR as functions of time when the RTFB consists of pure delay filters (DOA) and 1000 taps long RTFs (in the time domain), respectively.

delay filters (DOA, top result) and full RTFs having the length of 1000 taps in the time domain (bottom result). The most difficult periods for successful separation are when the interference source stays in position 2, that is, when the angular positions of both sources are the same. It can be observed that for those cases the iSIR of all methods drops down close to 0 dB. In these situations, the original IVA suffers from the global permutation problem, because the order of the separated outputs can be changed with high probability. In our experiment, the IVA performance suffers due to the global permutation (frames with negative iSIR). By contrast, the results show that with S-IVA the problem is solved, even with the DOA pilots. S-IVA piloted by clean signals in average achieves the best results and the result shows limits of the separation provides by S-IVA.

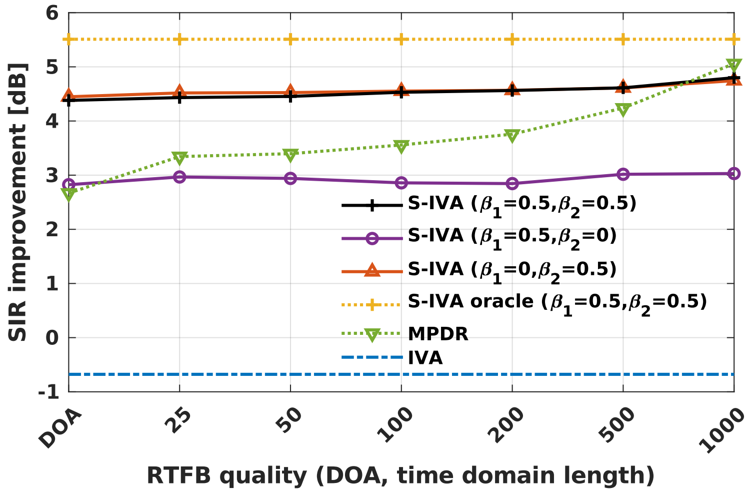

Fig. 3. Average separation performance (iSIR averaged over all frames) for various RTFB settings.

The performance of the beamforming methods is close to that of S-IVA only when the time-domain length of RTFs is 1000 taps or more. So the solution through S-IVA does not seem to bring many advantages compared to the methods from [7, 6, 10] in this configuration. However, S-IVA provides better performance when the RTFs are less accurate.

In the second experiment, we examine different time domain lengths of the RTFs in order to simulate a deteriorated performance of the RTFB. Fig. 3 compares average iSIR (over all frames) for all of the above-mentioned methods as functions of the various lengths (including the DOA setting).

The original IVA yields -1 dB of average iSIR due to the global permutation problem. The performance achieved through beamforming steadily grows with the time domain length of RTFs and outperforms S-IVA when the length exceeds 500 taps.

S-IVA is presented with three settings: First, S-IVA piloted by the output of the MPDR beamformer ($\beta_1 = 0.5, \beta_2 = 0$), which is dominated by the target signal. Second, S-IVA piloted only by the output of the null beamformer ($\beta_1 = 0, \beta_2 = 0.5$) that is dominated by the interference signal. Finally, S-IVA piloted by both pilot components ($\beta_1 = 0.5, \beta_2 = 0.5$). The results show that all variants solve the global permutation problem. Nevertheless, S-IVA piloted only by the MDPR beamformer is significantly worse than the other variants. This can be explained by the fact that the separation of the target from the interference is harder than the separation of interference from the target, because the target source is much closer to the microphones. Consequently, the piloting of the global permutation is more efficient when using the output of the null beamformer. Finally, we should mention the fact that the performance of S-IVA is not much influenced by the length of the RTFs.

5 Conclusion

In this work, we have proposed a novel variant of the Supervised IVA where the pilot component is obtained as the output of the MPDR or of the Null beamformer steered by a bank of pre-learned RTFs. We have shown by experiments that this variant of S-IVA is more practical than just using MPDR and Null beamforming taking the most appropriate RTF from the bank, because their performance is highly dependent on the quality of the RTFB. By contrast, we have shown that the performance of S-IVA piloted by outputs of the beamformers is robust against poor accuracy of RTFB, while the global permutation problem is efficiently solved.

In future works, we plan to generalize the proposed method for multiple microphones and sources. A straightforward way is to derive appropriate pilot components for all sources. Alternatively, a practical situation is when only some sources should be extracted from the mixture. Pilot components should be used to supervise the extraction of the sources as independent vector components. The goal is ensure that the blind method extracts the desired signal.

We also plan to compare our method with approaches that impose constraints on de-mixing filters to solve the global permutation problem, such as [4].

Acknowledgements

We are due to Dr. Francesco Nesta from Synaptics for his helpful comments and useful suggestions.

References

[1] Douglas SC, Gupta M (2007) Scaled natural gradient algorithms for instantaneous and convolutive blind source separation. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07, vol 2, pp II–637–II–640

[2] Gannot S, Burshtein D, Weinstein E (2001) Signal enhancement using beamforming and nonstationarity with applications to speech. IEEE Transactions on Signal Processing 49(8):1614–1626

[3] Garofolo JS, Lamel LF, Fisher WM, Fiscus JG, Pallett DS, Dahlgren NL (1993) Darpa timit acoustic phonetic continuous speech corpus cdrom

[4] Khan AH, Taseska M, Habets EAP (2015) A Geometrically Constrained Independent Vector Analysis Algorithm for Online Source Extraction, Springer International Publishing, Cham, pp 396–403

[5] Kim T, Attias HT, Lee SY, Lee TW (2007) Blind source separation exploiting higher-order frequency dependencies. IEEE Transactions on Audio, Speech, and Language Processing pp 70–79

[6] Koldovský Z, Málek J, Tichavský P, Nesta F (2013) Semi-blind noise extraction using partially known position of the target source. IEEE Transactions on Audio, Speech, and Language Processing 21(10):2029–2041

[7] Koldovský Z, Tichavský P, Botka D (2013) Noise reduction in dual-microphone mobile phones using a bank of pre-measured target-cancellation filters. In: Proceedings of IEEE International Conference on Audio, Speech and Signal Processing, pp 679–683

[8] Lee I, Kim T, Lee TW (2007) Independent vector analysis for convolutive blind speech separation. In: Blind speech separation, Springer, pp 169–192

[9] Liang Y, Naqvi SM, Chambers JA (2012) Audio video based fast fixed-point independent vector analysis for multisource separation in a room environment. EURASIP Journal on Advances in Signal Processing 2012(1):183

[10] Málek J, Koldovský Z, Gannot S, Tichavský P (2013) Informed generalized sidelobe canceler utilizing sparsity of speech signals. In: Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on, IEEE, pp 1–6

[11] Matsuoka K (2002) Minimal distortion principle for blind source separation. In: Proceedings of the 41st SICE Annual Conference. SICE 2002., vol 4, pp 2138–2143 vol.4

[12] Nesta F, Fakhry M (2013) Unsupervised spatial dictionary learning for sparse underdetermined multichannel source separation. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pp 86–90

[13] Nesta F, Koldovský Z (2017) Supervised independent vector analysis through pilot dependent components. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 536–540

[14] Nesta F, Mosayyebpour S, Koldovský Z, Paleček K (2017) Audio/video supervised independent vector analysis through multimodal pilot dependent components. In: Proceedings of European Signal Processing Conference, pp 1190–1194

[15] Ono N (2011) Stable and fast update rules for independent vector analysis based on auxiliary function technique. In: Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp 189–192

[16] Ono T, Ono N, Sagayama S (2012) User-guided independent vector analysis with source activity tuning. In: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp 2417–2420

[17] Sawada H, Mukai R, Araki S, Makino S (2003) A robust and precise method for solving the permutation problem of frequency-domain blind source separation. In: Proceedings of International Conference on Independent Component Analysis and Signal Separation, pp 505–510

[18] Smaragdis P (1998) Blind separation of convolved mixtures in the frequency domain. Neurocomputing 22:21–34