BLIND SOURCE SEPARATION USING INCOMPLETE DE-MIXING TRANSFORM WITH A PRECISE APPROACH FOR SELECTING CONSTRAINED SETS OF FREQUENCIES

Jakub Janský and Zbyněk Koldovský

Acoustic Signal Analysis and Processing Group,
Faculty of Mechatronics, Informatics and Interdisciplinary Studies,
Technical University of Liberec, Liberec, Czech Republic

ABSTRACT

This paper presents a modification of the Natural Gradient algorithm for Independent Vector Analysis estimating incomplete de-mixing transform and performing its completion. Incomplete de-mixing transform is obtained when it is estimated only on subsets of most active frequencies of the sources. The transform is then completed using methods for sparse reconstruction. In previous works, the incomplete subset of frequencies was the same for all separated signals. In this paper, we propose a new approach in which the subsets are source-dependent. Experiments conducted on the CHiME-4 dataset show that the proposed approach improves the separation performance.

Index Terms— Blind Source Separation, Independent Vector Analysis, Natural Gradient, Sparse Representations, Incomplete De-Mixing Transform

1. INTRODUCTION

Separation of source signals from a mixture acquired by a set of microphones is an important topic in audio signal processing. The problem becomes challenging when minimum knowledge about mixture and original signals is given. This problem is referred to as *blind source separation* (BSS). A popular method for solving the BSS problem is *Independent Component Analysis* (ICA) [1, 2], where the separation of audio mixtures proceeds by applying ICA in the frequency domain [3]. ICA is based on the assumption of independence between the sources. Conventional ICA method separate individually each frequency bin, which gives rise to the permutation problem: The frequency components of the sources have a random order. A de-permutation stage has to be done after ICA to achieve the final separation [4].

In recent years, a multivariate version of ICA has become popular [5, 6] which is referred as *Independent Vector Analysis* (IVA). Similarly to ICA, IVA assumes that the sources in the mixture are independent, however, in IVA, each source signal is modelled as a stochastic vector of mutually dependent variables. Hence, IVA does not suffer from the permutation problem.

In IVA, the BSS problem is solved by looking for a *de-mixing transform*, which corresponds to a set of square de-mixing matrices, one per each frequency. A popular algorithm for IVA due to its simplicity is *Natural gradient* (NG) [6], whose one iteration has a low computational complexity. However, conventional NG deals with the problem of setting an appropriate learning rate (step-length) parameter. Large step brings a fast speed of convergence at the cost for lower stability, and vice versa. A method for the step-length selection keeping ensuring stability and fast convergence was proposed in [7]. Methods that do not depend on the learning rate parameter were derived, e.g., in [8].

Recently, we have proposed to compute the de-mixing transform only within a subset of frequency bins, so-called Incomplete De-mixing Transform (IDT), where the input signals show the most significant activity [9, 10]. The other frequencies are not separated through IVA. Instead, they are separated using a completed IDT. These approaches bring computational savings proportional to the percentage of the chosen number of frequency bins in the subset. In addition, it was shown in [10] that the approach can improve the separation performance, e.g., when 35-45% of most active frequencies are used in 2×2 speech mixtures.

In this paper, we propose new approaches for selecting the subsets of frequencies for IVA. In previous works, a percentage of most active frequencies on input channels was used, which does not reflect two important facts. First, the source signals normally do not occupy the same frequency bins. Therefore, some frequencies of a given source can be overshadowed by dominant frequencies of the other sources. To overcome this problem, we propose choosing the frequencies for each source separately. Unfortunately, we do not know original signals. To obtain a frequency characteristic of source signals we propose to use pre-separated signals from the previous iteration. Second, the optimum percentage of selected frequencies need not be the same for all sources, because the sources can be of different nature. Therefore, we

This work was supported by The Czech Science Foundation through Project No. 17-00902S and by the Student Grant Scheme (SGS) at Technical University of Liberec.

 978-1-5386-8151-0/18/$31.00

propose adaptive percentage selecting for each source.

This paper is organized as follows. In Section 2, the BSS problem is formulated, and the solution through IVA by a modified algorithm similar to sNG is described. Section 3 introduces the concept of using incomplete de-mixing transform and its reconstruction. Section 4 describes the proposed solutions in selecting frequency bins for the incomplete separation. Section 5 is devoted to experimental evaluation, and Section 6 concludes the paper.

2. INDEPENDENT VECTOR ANALYSIS

The mixing model, in a Short-time Fourier domain, has the form

$$\mathbf{X}(\omega,\ell) = \mathbf{H}(\omega)\mathbf{S}(\omega,\ell), \tag{1}$$

where $\mathbf{X}(\omega,\ell) = [X_1(\omega,\ell),\dots,X_N(\omega,\ell)]^T$ and $\mathbf{S}(\omega,\ell) = [S_1(\omega,\ell),\dots,S_M(\omega,\ell)]^T$ denote, respectively, the recorded signals on microphones and the source signals; $\omega \in \{1,\dots,K\}$ is the frequency index; ℓ is the frame index; K is the length of FFT. $\mathbf{H}$ (without the argument) represents the mixing transform that involves the matrix matrices

$$\mathbf{H}(\omega) = \begin{bmatrix} H_{1,1}(\omega) & \dots & H_{1,M}(\omega) \\ \vdots & \ddots & \vdots \\ H_{N,1}(\omega) & \dots & H_{N,M}(\omega) \end{bmatrix}. \tag{2}$$

In this paper, we assume that the number of sources M is the same as that of microphones N, i.e., $M = N$. We are seeking for a de-mixing transform represented by $\mathbf{W}$ and consisting of de-mixing matrices $\mathbf{W}(\omega)$, $\omega \in \{1,\dots,K\}$, which satisfies

$$\mathbf{Y}(\omega,\ell) = \mathbf{W}(\omega)\mathbf{X}(\omega,\ell) = \mathbf{P}(\omega)\mathbf{D}(\omega)\mathbf{S}(\omega,\ell), \tag{3}$$

where $\mathbf{Y}(\omega,\ell)$ corresponds to the separated sources, which should be equals to the original sources up to their order and scales. $\mathbf{P}$ and $\mathbf{D}$ represent, respectively, a permutation and a scaling transform. To find the de-mixing transform, *Independent Vector Analysis* (IVA) can be used as follows.

In IVA, the mutual independence of the sources is assumed. The independence can be measured by the Kullback-Leibler divergence [11]. This approach leads to the minimization of the objective function

$$J(\mathbf{W}) = \sum_{i=1}^{N} E_\ell\left[-\log r(\mathbf{Y}_i(\ell))\right] - \sum_{\omega=1}^{K} \log|\det \mathbf{W}(\omega)|, \tag{4}$$

where $E[\cdot]$ denotes the expectation operator, and $r(\cdot)$ denotes the multivariate probability density function (PDF) of the ith separated source $\mathbf{Y}_i(\ell) = [Y_i(1,\ell),\dots,Y_i(K,\ell)]^T$.

To optimize the objective function (4), the update rules of the Natural Gradient algorithm are

$$\mathbf{W}(\omega) = \mathbf{W}(\omega) + \mu\Delta\mathbf{W}(\omega), \tag{5}$$

where

$$\begin{aligned} \Delta\mathbf{W}(\omega) &= \{\mathbf{I} + E_\ell\left[\boldsymbol{\varphi}_\omega(\mathbf{Y}(\ell))\mathbf{Y}(\omega,\ell)^H\right]\}\mathbf{W}(\omega) \\ \boldsymbol{\varphi}_\omega(\mathbf{Y}(\ell)) &= [\varphi_{1\omega}(\mathbf{Y}(\ell)),\dots,\varphi_{N\omega}(\mathbf{Y}(\ell))]^T \\ \varphi_{j\omega}(\mathbf{Y}(\ell)) &= \tfrac{\partial}{\partial Y_j(\omega,\ell)}\log r(\mathbf{Y}_j(\ell)), \end{aligned} \tag{6}$$

where $\Delta\mathbf{W}(\omega)$ is update of $\mathbf{W}(\omega)$, and μ is a positive step-length parameter.

To improve the speed of convergence, we propose the following rescaling of the update rule for the de-mixing matrices as

$$\mathbf{W}(\omega) = \mathbf{W}(\omega) + \frac{\mu}{\|\Delta\mathbf{W}(\omega)\|_2}\Delta\mathbf{W}(\omega). \tag{7}$$

This approach is a slightly modified variant of *Scaled Natural Gradient* (sNG) from [7]. In our experiments with incomplete de-mixing transform in Section 5, we observed a better convergence stability of the proposed modification then of the sNG. Simultaneously, the convergence speed of sNG appeared to be preserved.

3. INCOMPLETE DE-MIXING TRANSFORM AND ITS SPARSE RECONSTRUCTION

When the de-mixing transform is computed only for a certain subset of frequencies $U = \{\omega_1,\dots,\omega_{|U|}\} \subset \{1,\dots,K\}$, we refer to an *incomplete de-mixing transform* (IDT) $\mathbf{W}_U$; $|U|$ denotes the number of elements in U. This approach can bring computational savings, especially, when the source signals are sparse in the frequency domain, such as speech signals. The approach might also improve the convergence stability of IVA algorithms since frequencies where is little activity of the signals to be separated do not participate in the blind estimation process.

Once $\mathbf{W}_U$ is estimated through IVA, it should be completed to separate the whole frequency band. Let $\mathrm{w}^{i,j}(\omega)$ denote the ijth element of $\mathbf{W}(\omega)$, and let $\mathbf{w}^{i,j} = [w^{i,j}(1),\dots,w^{i,j}(K)]$, which corresponds to the *transfer function* (TF) between the jth input and ith output channel. For $k \in U$, $w^{i,j}(k)$ is obtained through the IVA, but the other elements of $\mathbf{w}^{i,j}$ are not known. Based on the ideas from [9], we propose to compute the unknown values of $\mathbf{w}^{i,j}$ through solving the basis pursuit problem defined as

$$\begin{aligned} \hat{\mathbf{w}}_{i,j} = \quad & \underset{\mathbf{g}}{\operatorname{argmin}} \|\mathbf{F}^{-1}\mathbf{g}\|_1 \quad \text{subject to} \\ & \mathbf{g}_U = \mathbf{w}_U^{i,j}, \end{aligned} \tag{8}$$

where $\mathbf{w}_U^{i,j}$ denotes the sub-vector of $\mathbf{w}^{i,j}$ whose elements belong to U; $\mathbf{F}$ is the $K \times K$ matrix of the Discrete Fourier Transform; $\hat{\mathbf{w}}_{i,j}$ denotes the reconstructed TF. To solve this optimization problem, we propose to employ the fast proximal algorithm from [12].

4. SELECTION OF SET U

The selection of a proper subset of frequencies is crucial for the presented method. Too small set U would lead to an inaccurate separation. On the other hand, a bigger set does not employ the full potential of the presented approach.

4.1. Different set U for each source

In [10], a fixed percentage of most active frequencies on input channels, measured by Power Spectral Density (PSD), is used. This approach does not reflect the different activity of the source signals on the frequency range. In cases, when one source signal is dominant in a part of the mixture, the active frequencies of the other sources may not be taken into account. This leads to the effective separation only within frequencies where the dominant source signal is active.

To solve this problem, information about active frequencies of each source is needed. During the separation process, this information can be obtained from the PSD of the pre-separated sources in the previous iteration. Therefore, we propose to utilize the most active frequencies of the separated sources from the previous iteration, which allows us to select an individual set U_j for the jth source.

Now, the NG update rule has to be modified, because the jth row of the de-mixing matrix $\mathbf{W}(\omega)$, corresponding to the jth separated source, is updated depending on U_j. The modified update reads

$$\begin{aligned} \mathbf{W}(\omega) &= \mathbf{W}(\omega) + \mu \Delta\mathbf{W}(\omega), \\ \Delta\mathbf{W}(\omega) &= \Delta\tilde{\mathbf{W}}(\omega)\mathbf{W}(\omega), \end{aligned} \tag{9}$$

where

$$\Delta\tilde{\mathbf{W}}_j(\omega) = \begin{cases} \mathbf{e}_j + E_\ell\left[\varphi_{j\omega}(\mathbf{Y}(\ell))\mathbf{Y}(\omega,\ell)^H\right] & \omega \in U_j, \\ \mathbf{e}_j & \omega \notin U_j, \end{cases} \tag{10}$$

where $\Delta\tilde{\mathbf{W}}_j(\omega)$ denotes the jth row of $\Delta\tilde{\mathbf{W}}(\omega)$, $\varphi_{j\omega}(\mathbf{Y}(\ell)) = \frac{\partial}{\partial Y_j(\omega,\ell)} \log r_{Y_j}(\mathbf{Y}_j(\ell))$, and $\mathbf{e}_j$ is the unit vector whose jth element is equal to one.

4.2. Variable percentage of frequencies for each source

Since U_j can be different for each source, it is also possible to select a different number of frequencies for each source. Therefore, the frequency bins where the PSD shows a higher activity than a threshold ν_j are selected. Specifically,

$$U_j(\omega) = \begin{cases} \omega & \mathrm{PSD}(\mathbf{Y}_j(\omega)) \geq \nu_j \\ \emptyset & \mathrm{PSD}(\mathbf{Y}_j(\omega)) < \nu_j \end{cases} \tag{11}$$

where $PSD(\cdot)$ is the normalized Power Spectral Density. Our heuristic selection of ν_j is given by

$$\nu_j = \max\{\max\{\mathrm{PSD}(\mathbf{Y}_j)\} - \lambda, \epsilon\}. \tag{12}$$

Parameter λ responds as cut off parameter derived from frequency bin with highest power in source. Parameter ϵ to the smallest acceptable power of frequency bin.

To summarize, Algorithm 1 shows a pseudocode of one iteration of the proposed method. For simplicity, frequency and frame index were omitted.

Algorithm 1: One iteration of Natural gradient with incomplete de-mixing matrix and variable set U for each source

Input: $\mathbf{W}$,$\mathbf{X}$,μ, λ, ϵ
Output: $\mathbf{W}$

$\mathbf{Y} = \mathbf{W}\mathbf{X}$;
$\nu_j = \max\{\max\{\mathrm{PSD}(\mathbf{Y}_j)\} - \lambda, \epsilon\}$;
foreach $\omega = 1 \dots K$ **do**
  foreach $j = 1 \dots M$ **do**
    if $PSD(\mathbf{Y}_j(\omega)) \geq \nu_j$ **then**
      $\Delta\tilde{\mathbf{W}}_j(\omega) = \mathbf{e}_j + \frac{1}{L}\sum_{\ell=1}^{L}\left[\varphi_{j\omega}(\mathbf{Y}(\ell))\mathbf{Y}(\omega,\ell)^H\right]$;
    else
      $\Delta\tilde{\mathbf{W}}_j(\omega) = \mathbf{e}_j$;
    end
  end
  $\Delta\mathbf{W}(\omega) = \Delta\tilde{\mathbf{W}}(\omega)\mathbf{W}(\omega)$;
  $\mathbf{W}_U(\omega) = \mathbf{W}(\omega) + \frac{\mu}{\|\Delta\mathbf{W}(\omega)\|_2}\Delta\mathbf{W}(\omega)$;
end
$\mathbf{W} \leftarrow$ reconstruction$(\mathbf{W}_U)$

5. EXPERIMENTS

To evaluate the performance of proposed approaches, we apply them on CHiME-4 challenge [13] development simulated data set. This dataset contains different speakers in four different noise environments, bus, cafeteria, pedestrian area and street. For each environment, we randomly select 500 mixtures, which means 2000 mixtures in total. The length of each mixture is between 6 s and 10 s. In this paper, human voice source will be referred as first source and environmental noise as the second source. Experiments were evaluated by the BSS_eval toolkit [14] in Matlab using three metrics: *signal-to-interference ratio* (SIR), *signal-to-distortion ratio* (SDR) and *signal–to-artifacts ratio* (SAR). The fourth metric was the time needed for the separation of one mixture. All four metrics were averaged over all mixtures.

We tested four different methods of choosing set of frequency bins U. First method (*Fixed percentage*), utilize flat fixed percentage of frequency bins according previous work [10]. We shows two variation of this method, with reconstruction described in (8) and without reconstruction. Second method (*Variable U*) use approach discussed in subsection 4.1. Third method (*Fixed threshold*) utilize variable percentage of used frequency bins described by equation (11) with

Table 1. Results of the experiments

Method	SIR [dB]		SAR [dB]		SDR [dB]		%		Time [s]
	1st	2nd	1st	2nd	1st	2nd	1st	2nd	
Classical NG	10.75	10.73	10.17	2.49	6.94	0.89	100	100	4.53
Fixed perc. no recon.	10.99	10.67	10.29	2.57	7.13	1.00	70	70	3.52
Fixed percentage	11.00	11.04	10.32	2.56	7.15	1.06	70	70	3.84
Variable U	11.02	11.39	10.49	2.60	7.27	1.27	60	60	3.51
Fixed threshold	11.41	11.98	10.42	**2.80**	7.40	1.56	**37.80**	**72.48**	**3.25**
Variable threshold	**11.42**	**12.34**	**10.47**	2.74	**7.46**	**1.57**	41.45	81.57	3.38

Table 2. Experimental setup

Used microphones	3 and 4
μ	1
Signal sampling	16 kHz
Frame length	2048
Frame shift	1024
# iterations	150

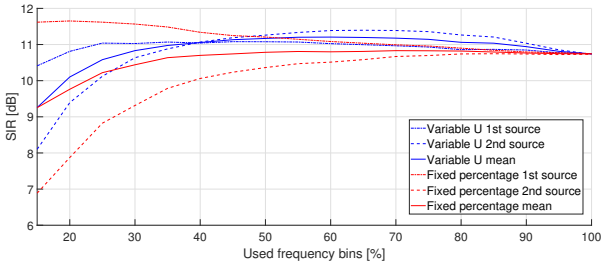

Fig. 1. Comparison of SIR of method *Fixed percentage* and *Variable U*

fixed $\nu_j = 0.00015$. The fourth method (*Variable threshold*) also variable percentage however with variable ν_j described in equation (12) with parameters $\lambda = 35$ dB and $\epsilon = -60$ dB. For comparison we also include classical NG algorithm without reconstruction. Initial value of de-mixing transform was chosen as $\mathbf{W}(\omega) = \mathbf{I}$ for all ω. Setup of other parameters in experiment is in Table 2.

Results of the experiment are shown in Table 1. For method *Fixed percentage* and *Variable U* we run experiments for 15% through 100% of used frequency bins and selects the best result with respect to SIR. Because percentage of used frequencies is different for every mixture for methods *Variable threshold* and *Fixed threshold* the percentage of used frequency bins was averaged over all mixtures. According to results, the best method is *Variable threshold* in almost all BSS metrics, however a bit slower. For methods with fixed percentage, best results are achieved on 75% and 60% of used frequency bins. This result is different from best results (25% - 40%) from our previous work [10]. This is caused by the fact that different sources were used. In previous work, only human voice mixtures were considered. In this paper, there are speech sources and nonstationary noise. The noise is active between 60% - 80% of frequency bins. Methods utilizing variable percentage reflect each source in the mixture in a realistic way. For these methods, a first source, corresponding to the speech signal, was separated with approximately 40% of frequency bins, while the second source, corresponding to noise, require more then 70% of frequency bins.

For further comparison, Fig. 1 shows an average of both sources, first source and a second source SIR of *Fixed percentage* and *Variable U* methods for 15% to 100% of used frequency bins. Old approach *Fixed percentage* achieve same separation performance between 35% and 100%. However, method *Variable U* achieve improvement in this range. For method *Fixed percentage*, SIR of the second source is falling, because a frequency bins important for separation are occupied mainly by 1st source, which has increasing SIR. In contrast, *Variable U* method SIR of the second source is obtaining the best value between 50% - 80% and SIR of the first source is slightly increasing till 25%. For this method, it is an expected behavior, because the frequency bins are not shared.

6. CONCLUSION

This paper presents a solution of the blind audio source separation problem by using incomplete de-mixing transform. We proposed novel approaches for selecting sets of frequency bins for the incomplete de-mixing transform. Selecting different set for each source in mixture with a number of elements turns out to be better with respect to the evaluation metrics. The proposed approach is more suitable for mixtures with different source signal types, for example, human speech and environmental noise. For a future work, we plan to apply the proposed approach to the more sophisticated IVA methods, e.g. AuxIVA.

7. REFERENCES

[1] T. Lee, *Independent Component Analysis: Theory and Applications*, Kluwer Academic Publishers, Norwell, MA, USA, 1998.

[2] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, no. 4-5, pp. 411–430, May 2000.

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

[4] 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," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 530–538, Sept 2004.

[5] A. Hiroe, "Solution of permutation problem in frequency domain ICA, using multivariate probability density functions," in *Independent Component Analysis and Blind Signal Separation*, Berlin, Heidelberg, 2006, pp. 601–608, Springer Berlin Heidelberg.

[6] T. Kim, T. Eltoft, and T. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *Independent Component Analysis and Blind Signal Separation*, Berlin, Heidelberg, 2006, pp. 165–172, Springer Berlin Heidelberg.

[7] S. C. Douglas and M. Gupta, "Scaled natural gradient algorithms for instantaneous and convolutive blind source separation," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, April 2007, vol. 2, pp. II–637–II–640.

[8] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct 2011, pp. 189–192.

[9] Z. Koldovský, J. Málek, and S. Gannot, "Spatial source subtraction based on incomplete measurements of relative transfer function," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 8, pp. 1335–1347, Aug 2015.

[10] J. Janský, Z. Koldovský, and N. Ono, "A computationally cheaper method for blind speech separation based on AuxIVA and incomplete demixing transform," in *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, Sept 2016, pp. 1–5.

[11] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, 2006.

[12] N. Parikh and S. Boyd, *Proximal Algorithms*, Foundations and Trends in Optimization. Now Publishers, 2013.

[13] E. Vincent, S. Watanabe, A.A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.

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