A COMPUTATIONALLY CHEAPER METHOD FOR BLIND SPEECH SEPARATION BASED ON AUXIVA AND INCOMPLETE DEMIXING TRANSFORM

Jakub Janský[1], *Zbyněk Koldovský*[1], *Nobutaka Ono*[2],

[1]Faculty of Mechatronics, Informatics, and Interdisciplinary Studies, Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic
[2]National Institute of Informatics (NII), SOKENDAI (The Graduate University for Advanced Studies), 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan,

ABSTRACT

This paper proposes a modification of an auxiliary-function based algorithm for Independent Vector Analysis (AuxIVA) by applying it to a constrained set of frequency bins where the activity of (speech) signals is high. The de-mixing transform obtained by this approach is incomplete. Its completion is done through the solution of a convex optimization problem known as LASSO. The experiments with blind separation of speech signals show that the proposed method can be twice faster and sometimes even slightly more accurate in terms of signal-to-noise ratio than the original algorithm applied to all frequencies.

1. INTRODUCTION

In Blind Source Separation (BSS), the goal is to separate source signals from their mixture using minimum knowledge. The source signals can be, for example, speech, as assumed in this paper. Independent component analysis (ICA) is a popular method to solve the BSS problem [1, 2] that relies purely on the fact that speeches of two (or more) persons are, as signals, stochastically independent. ICA can separate independent signals up to their original scales and order.

Real-world mixtures of speech signals are convolutive due to reflections and reverberation. It is therefore practical to process the signals in the frequency domain, where the time-domain convolutive mixture is transformed into a set of instantaneous mixtures, one mixture per frequency bin. Then, ICA is applied independently in each frequency bin, which gives rise to the permutation problem: The separated frequency components of original signals have random order. To separate the signals in the time-domain, the permutation ambiguity has to be resolved [3].

A more recent methodology that avoids the permutation problem is Independent Vector Analysis (IVA), proposed in [4, 5]. In IVA, the frequency domain representation of an independent signal is modeled as a vector of stochastic variables that are mutually dependent. Then, similarly to ICA, IVA performs the separation through maximizing the independence between the separated signals (vectors), but it simultaneously aims to preserve (maximize) the dependence between frequency components of each separated signal.

A popular IVA method is the Natural Gradient (NG) algorithm [7, 5], especially, for its computational simplicity, the ability to work in an adaptive regime, and an appealing accuracy when initialized in a close vicinity of the ideal solution. By contrast, its convergence speed and sensitivity to the initialization stand for its drawbacks.

In [8, 9], a new IVA algorithm utilizing update rules based on an auxiliary function was derived. The approach is less sensitive to the initialization than NG and avoids the crucial problem to tune a step-length parameter. The algorithm will be referred to as AuxIVA.

In this paper, we propose a new approach utilizing AuxIVA for separating speech signals. As speech is sparse in the frequency domain, AuxIVA is applied only on a subset of active frequencies in the input, which brings computational savings (proportionally to the number of the selected frequencies). The constrained algorithm yields an incomplete demixing transform. It is subsequently extrapolated through finding its sparsest representation in the time-domain as proposed recently in [10]. The latter step is done by using a fast convex programming algorithm solving LASSO [11].

In experiments, AuxIVA applied in the whole frequency range is compared with the proposed approach. The results show that the latter can be twice faster than AuxIVA without loosing much in terms of the separation accuracy. In some cases where the separation accuracy achieved by AuxIVA is poor the proposed approach improves the separation, depending on the selection of the constrained set of frequencies.

The paper is organized as follows. Section 2 describes the BSS problem and the original AuxIVA algorithm. Section 3 describes the proposed approach, and Section 4 presents the experimental evaluations. Section 5 concludes the paper.

This work was partly supported by The Czech Science Foundation through Project No. 14-11898S, by California Community Foundation through Project No. DA-15-114599, and by Grant-in-Aid for Scientific Research (A) (Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 16H01735).

978-1-5090-2007-2/16/$31.00

2. BLIND SEPARATION USING AUXIVA

2.1. Problem Description

We will focus on the scenario with two speakers and two microphones, although the ideas can be generalized to more sources and microphones. A stereo recording of two speakers can be, in the short-time Discrete Fourier Transform (DFT) domain, described as

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

where $k = 1,\dots,K$ and $\ell = 1,\dots,\tau$ denote, respectively, the frequency and the time frame index; $\mathbf{X}(k,\ell) = [X_1(k,\ell), X_2(k,\ell)]^T$ represents recorded signals on microphones; $\mathbf{S}(k,\ell) = [S_1(k,\ell), S_2(k,\ell)]^T$ are the original speech signals, and $\{\mathbf{H}(k)\}_{ij} = H_{ij}(k)$, $i,j \in \{1,2\}$, are the acoustic transfer functions between the speakers and microphones.

A transform $\mathbf{W}(k)$, $k = 1,\dots,K$, which fulfills

$$\mathbf{Y}(k,\ell) = \mathbf{W}(k)\mathbf{X}(k,\ell) = \mathbf{P}(k)\mathbf{D}(k)\mathbf{S}(k,\ell), \tag{2}$$

where $\mathbf{Y}(k,\ell) = [Y_1(k,\ell), Y_2(k,\ell)]^T$ are the separated sources up to their original order and scales, that is, $\mathbf{P}(k)$ is a permutation matrix and $\mathbf{D}(k)$ is a diagonal matrix, will be referred to as de-mixing. The goal of IVA is to find a de-mixing transform such that $\mathbf{P}(k)$ are the same for each k, which solves the permutation problem. The scaling ambiguity due to $\mathbf{D}(k)$ is solved by projecting the separated sources back to the microphones, so the final separated signals correspond to the spatial images of the original sources at the microphones [6].

2.2. Independent Vector Analysis

In IVA, a joint probabilistic model of the mixtures (1), that is, for the set of models for $k = 1,\dots,K$, is assumed. Finding of the de-mixing transform is performed through minimizing the Kullback-Leibler divergence between the joint probability function and the product of probability densities of vectors where each vector contains the frequency components of the corresponding source. The components within a vector are allowed to be mutually dependent [5].

For two sources, the objective function can have the form

$$J(\mathbf{W}) = \sum_{i=1}^{2} \frac{1}{\tau} \sum_{\ell=1}^{\tau} [G(\mathbf{Y}_i(\ell))] - \sum_{k=1}^{K} \log|\det \mathbf{W}(k)|, \tag{3}$$

where $\mathbf{W}$ without the argument k stands for the whole de-mixing transform, and $\mathbf{Y}_i(\ell) = [Y_i(1,\ell),\dots,Y_i(K,\ell)]^T$ denotes a vector of the frequency components of the ith source within the ℓth frame. $G(\cdot)$ is a scalar *contrast function*. In theory, the optimum choice in the maximum likelihood sense is $G(\mathbf{Y}_i(\ell)) = -\log p(\mathbf{Y}_i(\ell))$ where $p(\cdot)$ is the multivariate probability density function of the frequency components. The pdf is not known in the blind scenario, so $G(\cdot)$ is often chosen in a heuristic way.

2.3. Auxiliary-function-based IVA

AuxIVA [9] is based on an optimization approach that assumes that $G(\mathbf{Y}_i(\ell))$ is only a function of $\|\mathbf{Y}_i(\ell)\|_2$ and that $G'(r)/r$ is positive and continuous everywhere and is monotonically decreasing in the wider sense for $r \geq 0$.

Auxiliary variables are defined through

$$\mathbf{V}_i(k) = \frac{1}{\tau} \sum_{\ell=1}^{\tau} \frac{G'(r_{i,\ell})}{r_{i,\ell}} \mathbf{X}(k,\ell)\mathbf{X}(k,\ell)^h \tag{4}$$

where

$$r_{i,\ell} = \|\mathbf{Y}_i(\ell)\|_2 = \sqrt{\sum_{k=1}^{K} |\mathbf{w}_i^h(k)\mathbf{X}(k,\ell)|_2^2}, \tag{5}$$

$\cdot^h$ denotes the conjugate transpose, and $\mathbf{w}_i(k)^h$ is the ith row of $\mathbf{W}(k)$. Finding of the minimum of the objective (3) proceeds by minimizing an auxiliary function

$$\begin{aligned} Q_k(\mathbf{W}(k), \mathbf{V}(k)) = \; & \tfrac{1}{2} \sum_{i=1}^{2} \mathbf{w}_i(k)^h \mathbf{V}_i(k) \mathbf{w}_i(k) \\ & - \log|\det \mathbf{W}(k)| + R \end{aligned} \tag{6}$$

in $\mathbf{W}$, and by recalculating $\mathbf{V}$, alternatively. R denotes a positive constant independent of $\mathbf{V}$ and $\mathbf{W}$.

Finally, the algorithm proceeds by repeating the steps until convergence. In the first step, the auxiliary variables are updated according to (4) and (5). Second, the de-mixing transform is updated as

$$\begin{aligned} \mathbf{w}_i(k) &\leftarrow (\mathbf{W}(k)\mathbf{V}_i(k))^{-1}\mathbf{e}_i, \\ \mathbf{w}_i(k) &\leftarrow \mathbf{w}_i(k)/\sqrt{\mathbf{w}_i(k)^h \mathbf{V}_i(k) \mathbf{w}_i(k)}, \end{aligned} \tag{7}$$

where $\mathbf{e}_i$ denotes the ith column of the identity matrix.

3. PROPOSED METHOD

We propose to apply AuxIVA on a constrained set of the models (1). It means that the de-mixing matrices will be estimated only for certain frequencies, which results in an *incomplete de-mixing transform*. Let the constrained set of frequency bins be $S = \{k_1,\dots,k_{|S|}\} \subset \{1,\dots,K\}$. We propose to select S as the set of the most active frequencies in the signal mixture (either on the left or right microphone or on average taken over both microphones). This choice appears to be economical with speech signals as their short-term or medium-term activity in the frequency domain is sparse.

The modification of AuxIVA for the constrained set is straightforward: The updates (4) and (7) are performed only for $k \in S$, and the sum in (5) is computed only over $k \in S$.

Algorithm 1 Proposed method

Input: mixed signals, $S = \{k_1, \ldots, k_{|S|}\}$, μ, MIter
Output: $\mathbf{W}$

```
initialize: W ← I
for Iter= 1 to MIter do
    for i = 1 to 2 do
        r_{i,ℓ} = sqrt( Σ_{k∈S} |w_i(k)^h X(k,ℓ)|_2^2 )
        for {k_1, ..., k_{|S|}} do
            Compute V_i(k) using (4)
            Compute w_i(k) using (7)
        end
    end
end
for i = 1 to 2 do
    Z_i = [ w_{i,1}(k_1)/w_{i,2}(k_1), ..., w_{i,1}(k_{|S|})/w_{i,2}(k_{|S|}) ]^h
    Ĥ_{RTF,i} = arg min_H μ||F^{-1}H||_1 + ||Z_i − H_S||_2^2
    w_i(k)^h ← [Ĥ_{RTF,i}(k), −1]
end
```

3.1. Reconstruction of incomplete de-mixing transform

The output of the proposed method is an incomplete set of de-mixing matrices $\mathbf{W}(k)$, $k \in S$. To complete, we reformulate the problem as to reconstruct *relative transfer functions* (RTFs) between the microphones related to the sources.

The RTF for the ith source is defined as [10]

$$H_{RTF,i}(k) = \frac{H_{2i}(k)}{H_{1i}(k)}. \tag{8}$$

Provided that $\mathbf{W}(k)$ is de-mixing, the scaling ambiguity allows to multiply it by a non-singular diagonal matrix so that $\mathbf{W}(k)$ gets the form (up to the permutation of its rows)

$$\mathbf{W}(k) = \begin{Bmatrix} H_{RTF,2}(k) & -1 \\ H_{RTF,1}(k) & -1 \end{Bmatrix}. \tag{9}$$

The de-mixing property is preserved.

Now, since $\mathbf{W}(k)$ are known only for $k \in S$, the RTFs obtained through (9), from now represented by vectors $\mathbf{Z}_i$, $i \in \{1, 2\}$, are incomplete as well (iRTF). We adopt the approach from [10] to complete the iRTFs though finding their sparsest representations in the time domain. This is justified by fact that relative impulse responses are fast decaying sequences, so the RTFs are approximately sparse in the time domain.

To find the representations, the formulation through LASSO is used [11]. The ith reconstructed RTF is computed as

$$\hat{\mathbf{H}}_{RTF,i} = \arg\min_{\mathbf{H}} \|\mathbf{Z}_i - \mathbf{H}_S\|_2^2 + \mu\|\mathbf{F}^h\mathbf{H}\|_1, \tag{10}$$

where $\mathbf{H}_S$ is the vector of elements of $\mathbf{H}$ whose indices are in S, $\mathbf{F}$ is the matrix of the Discrete Fourier transform, $\|\cdot\|_1$ represents the ℓ_1 norm and μ is a positive parameter that controls the sparsity of the solution. In this paper, Alternating direction method of multipliers (ADMM) is used to find the solution [12].

By putting the reconstructed RTFs back to (9), the completed de-mixing transform is obtained.

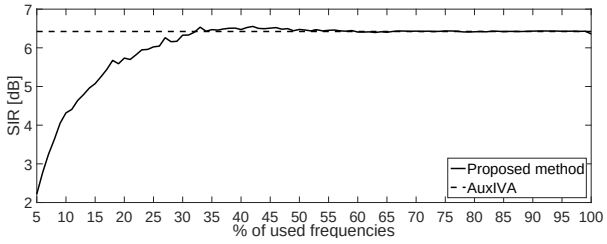

Fig. 1. SIR as a function of the number of selected frequencies averaged over 900 mixtures.

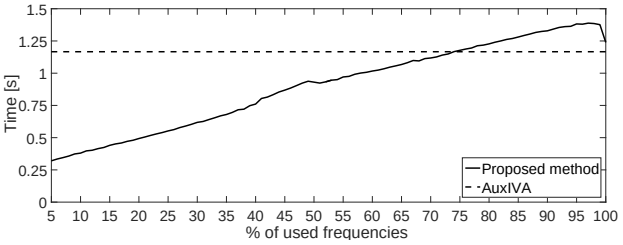

Fig. 2. Average computational time to separate one mixture.

3.2. Method summary

The input parameter is the number of percents (*percentage*) of the most active frequencies, based on which the set $S = \{k_1, \ldots, k_{|S|}\}$ is selected. The Welch's power spectral density estimates applied to both input signals are used for the selection. Then, the constrained AuxIVA is applied, and the obtained incomplete transform is completed using the procedure described in the previous subsection. The pseudocode is given in Algorithm 1.

4. EXPERIMENTAL EVALUATION

We conducted an experimental study with synthesized mixtures of speech. The source signals were taken from the ATR Japanese speech database, set B [15]. Impulse responses recorded in a variable reverberation room (E2A) were taken from the RWCP Sound Scene Database in Real Acoustical Environments [13]. The sources were located on a half circle at the distance of 2 m from microphones. The positions were at angles from 10° through 170° with 20° spacing. Other details of the scenario are given in Table 1. In total, 900 mixtures were generated. In AuxIVA as well as in the proposed method, the scalar contrast function $G(r) = r$ was chosen.

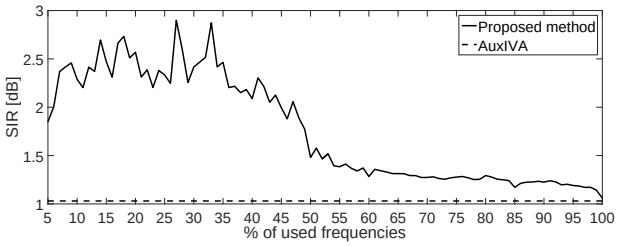

Fig. 3. Average SIR over 148 trials in which AuxIVA achieves SIR less than 2dB.

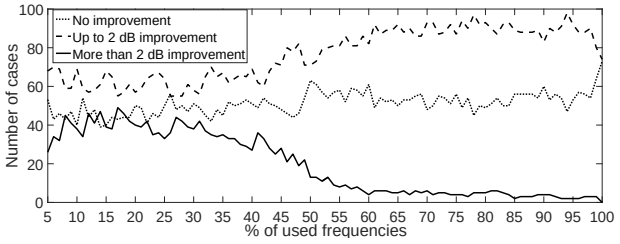

Fig. 4. Number of cases out of 148 where the proposed method improved the separation compared to AuxIVA.

The mixtures of signals were separated and evaluated in terms of the Signal-to-Interference Ratio (SIR) averaged over the separated sources, using the BSS_EVAL toolbox [14]. In each trial, AuxIVA was initialized by the identity matrix within each frequency, and 15 iterations were performed. The experiments were performed in Matlab R2013b on a PC Intel core i5-4590 3.30 GHz.

Table 1. Experimental setup

Sentences on each position	5
Microphone spacing	2.83 cm
Reverberation time	300 ms
Length of signals	10 s
Signal sampling	16 kHz
Frame length	512
Frame shift	128

Fig. 1 shows the SIR averaged over the 900 mixtures as a function of the percentage within the proposed method. When all frequencies are used, the method coincides with the original AuxIVA up to a negligible bias due to LASSO. Its performance is decreasing with the number of frequencies almost monotonically. Nevertheless, the performance loss is negligible until 30% and is even slightly improved compared to that of the original AuxIVA for 35-45%.

The computational times are compared in Fig. 2. Above 70%, the proposed method is slower than AuxIVA due to running the ADDM algorithm. However, it is faster below 70%. Taking into account both comparisons in Figures 1 and 2, the proposed method appears to be most effective for about 40%, because it is twice faster than AuxIVA while its performance

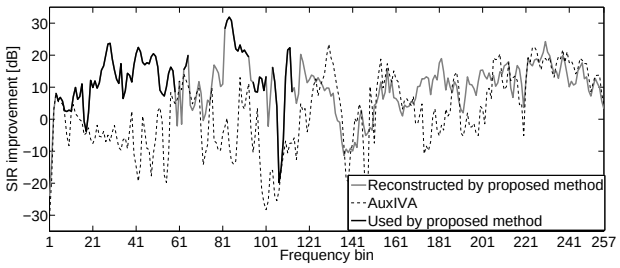

Fig. 5. SIR improvement as a function of the frequency index. The frequencies used by the proposed method (35%) are highlighted by black line.

is comparable or could be even slightly better.

We analysed 148 cases out of 900 where the SIR achieved by AuxIVA was less then 2 dB. Here, the proposed method achieved higher average SIR than AuxIVA for the full range of the percentage; see Fig. 3. Fig. 4 shows details of the number of cases where the proposed method improved the separation compared to AuxIVA. These results can be indicative of an improved global convergence of the proposed method.

By contrast, only in 63 trials out of 900, the SIR achieved by the proposed methods was lower by more than 3 dB compared to AuxIVA for the percentage greater or equal than 30%.

A particular trial is analyzed in Fig. 5 when the sources were located at 30° and 90°. Here, the proposed method with 35% achieved 8 dB SIR improvement compared to AuxIVA. Fig. 5 compares SIR improvements achieved by the methods related to the first source as functions of frequency. In case of the proposed method, black color signifies the active frequencies in S. This result can be indicative of the fact that AuxIVA can sometimes get stuck in a local minimum, by which the permutation problem is not fully resolved. By contrast, the proposed method improves SIR in most of the selected frequencies.

5. CONCLUSION

We have proposed a method that combines AuxIVA and the concept of incomplete de-mixing transform. The incomplete transform is extrapolated using convex programming. The experiments have shown that the proposed method can be, on average, twice faster and even slightly more accurate in terms of SIR than AuxIVA, provided that an appropriate constrained set of frequencies is chosen.

Future works will be focused on a deeper analysis of the cases where AuxIVA did not improve SIR by more than 2 dBs. Namely, more experiments with parameters such as frame length, initialization and the number of iterations, which can influence the final separation, will be done.

6. REFERENCES

[1] T. W. Lee, *Independent Component Analysis - Theory and Applications*. Boston, MA: Springer US, 1998.

[2] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent component analysis*. New York: J. Wiley, 2001.

[3] H. Sawada, R. Mukai, S. Araki, S. Makino, "A Robust and Precise Method for Solving the Permutation Problem of Frequency-Domain Blind Source Separation," *IEEE Trans. SAP*, vol. 12, no. 5, pp. 530–538, Sept. 2004.

[4] T. Kim, H. T. Attias; T.-W. Lee, "Blind Source Separation Exploiting Higher-Order Frequency Dependencies," *IEEE Trans. ASLP*, vol. 15, no. 1, Jan. 2007.

[5] T. Kim, T. Eltoft, and T.-W. Lee, "Independent Vector Analysis: An Extension of ICA to Multivariate Components", *Proc. ICA*, pp. 165–172, 2006.

[6] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," *Proceedings of 3rd International Conference on Independent Component Analysis and Blind Source Separation (ICA '01)*, pp. 722-727, San Diego, Calif, USA, Dec. 2001.

[7] A. Hiroe, "Solution of Permutation Problem in Frequency Domain ICA Using Multivariate Probability Density Functions", *Proc. ICA*, pp. 601–608, 2006.

[8] N. Ono and S. Miyabe, "Auxiliary-function-based Independent Component Analysis for Super-Gaussian Sources," *Proc. LVA/ICA*, pp. 165–172, 2010.

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

[10] Z. Koldovský, J. Málek, and S. Gannot, "Spatial Source Subtraction Based on Incomplete Measurements of Relative Transfer Function," *IEEE/ACM Trans. on Speech, Audio and Language Processing*, vol. 23, no. 8, pp. 1335–1347, Aug. 2015.

[11] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc. B.,* vol. 58, pp. 267–288, 1996.

[12] N. Parikh and S. Boyd, "Proximal Algorithms," *Foundations and Trends in Optimization*, 1(3):123–231, 2014.

[13] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical Sound Database in Real Environments for Sound Scene Understanding and Hands-Free Speech Recognition," *Proc. LREC*, pp. 965–968, 2000.

[14] E. Vincent, C. Févotte, and R. Gribonval, "Performance Measurement in Blind Audio Source Separation," *IEEE Trans. ASLP*, vol. 14, no. 4, pp. 1462–1469, 2006.

[15] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, K. Shikano, "ATR Japanese Speech Database as a Tool of Speech Recognition and Synthesis," *Speech Communication*, vol. 9, no. 4, pp. 357-363, 1990.