Sparse Independent Vector Analysis Based on Mel Filter

Takahiro Ushijima
Graduate School of Engineering
Oita University, Japan

Yuuki Tachioka
Denso IT Laboratory, Japan

Shingo Uenohara, Ken'ichi Furuya
Faculty of Science and Technology
Oita University, Japan

***Abstract*—To make independent vector analysis (IVA) robust for whitening, sparse IVA clips spectrum in high frequency bands, because whitening generates artificial noise in high frequency bands. In this paper, to avoid clipping of source spectrum by sparse IVA, we propose an application of Mel filter to the observed spectrum before clipping in order to emphasize spectrum in low frequency bands, because source spectrum in low frequency bands is often more important and sparse than that in high frequency bands. The effectiveness of the proposed method is confirmed by sound source separation experiments.**

***Index Terms*—Proximal splitting algorithm, Sparse IVA, Mel filter, Whitening**

I. Introduction

Blind source separation (BSS) is a method for extracting source signals from observed mixed signals without prior information such as microphone and source location. One of the most effective BSS methods is independent component analysis [1] and its extensions, and IVA [2]. Before BSS, whitening is widely used to improve its performance. However, whitening adds artificial noise to observed spectrum, which distorts the sparseness of source spectrum. To recover this sparseness of source signals, a sparse IVA clips spectrum based on the measurement of sparsity [3], [4].

When we deal with sources whose power is biased to a low-frequency band such as speech or music, the optimal threshold of clipping is different for lower and higher frequency bands but the conventional sparse IVA uniformly clips spectrum in the entire frequency bands. To set optimal threshold for each frequency bin, we propose to clip spectrum on the Mel scale in order to maintain spectrum in the low frequency bands, which are dominated by sources, while suppressing spectrum in the high frequency bands, which are dominated by additional noise. .

II. BSS based on proximal splitting algorithm

A. Objective function

An observed spectrum $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijM})^\top$ is obtained by the short-time Fourier transform of the observed signals from each microphone channel $m = 1, \ldots, M$, where $i = 1, \ldots, I$ represents a frequency bin, $j = 1, \ldots, J$ represents a time frame, and $\top$ denotes the transpose. A source spectrum is $\boldsymbol{s}_{ij} = (s_{ij1}, s_{ij2}, \ldots, s_{ijN})^\top$, where $n = 1, \ldots, N$ is the source ID. Source spectrum can be related to the observed spectrum as $\boldsymbol{x}_{ij} = A_i \boldsymbol{s}_{ij}$, where $A_i \in \mathbb{C}^{M \times N}$ is a mixing matrix. Under the determined condition ($M = N$), its inverse, A_i^{-1}, can be used for a demixing matrix $W_i \in \mathbb{C}^{N \times M}$.

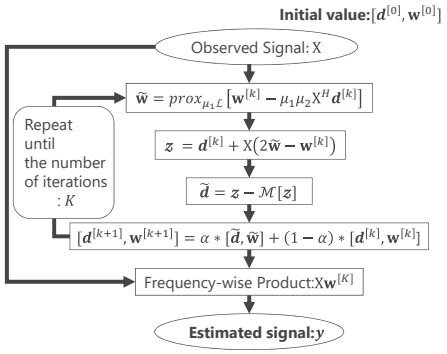

Fig. 1. BSS method based on a primal-dual (proximal splitting) algorithm

Using this demixing matrix, the separated spectrum can be obtained by $\boldsymbol{y}_{ij} = W_i \boldsymbol{x}_{ij}$ where $\boldsymbol{y}_{ij} = (y_{ij1}, y_{ij2}, \ldots, y_{ijN})^\top$ represents a separated signal. Under the assumption of source independence, the above demixing matrix can be estimated if the probability distribution of the source follows certain statistical distributions. The separated signals may be estimated by solving the following minimization problem.

$$\underset{\{W_i\}_{i=1}^I}{\text{Minimize}} \quad \sum_{i=1}^{I} \mathcal{P}\left(W_i \boldsymbol{x}_{ij}\right) - \sum_{i=1}^{I} \log\left|\det\left(W_i\right)\right|. \qquad (1)$$

The first term represents a real-valued penalty function that corresponds to the deviation from the assumed source model and the second term is introduced in order to normalize the scale of a demixing matrix.

B. Sparse IVA [4]

Whitening distorts a sparse structure of source spectrum. To recover this, a sparse IVA based on a proximal splitting algorithm has been proposed [3], [4]. Fig. 1 shows the flow of the BSS method based on a proximal splitting algorithm. The mask $\mathcal{M}[\boldsymbol{z}]$ in Fig. 1 is a proximity operator $\mathrm{prox}_{\lambda_1\|\cdot\|_{2,1}+\lambda_2\|\cdot\|_1}[\boldsymbol{z}]$. In addition, two weights Θ_η and $\Xi_\kappa[\cdot]$ are introduced to calculate the sparsity of separated signals and reduce the bias. The mask of the sparse IVA is as follows:

$$\left(\mathcal{M}[\boldsymbol{z}]\right)_{ijn}^{\boldsymbol{x},\eta,\lambda_{1,2},\kappa} = \zeta_{ijn}^{\boldsymbol{z},\lambda_2,\kappa} \times \Xi_\kappa\left[\left(1 - \left(\lambda_1 \Big/ \left(\Sigma_{i=1}^{I}(\Theta_\eta[\boldsymbol{x}])_i \left|\zeta_{ijn}^{\boldsymbol{z},\lambda_2,\kappa} z_{ijn}\right|^2\right)^{1/2}\right)\right)_+\right]. \qquad (2)$$

$\zeta_{ijn}^{\boldsymbol{z},\lambda_2,\kappa} = \Xi_\kappa[(1 - \lambda_2/|z_{ijn}|)_+]$ corresponds to the proximity operator of L_1 norm as the firm threshold, where λ_1, λ_2 are nonnegative threshold valuesand $(\cdot)_+$ is $\max(0, \cdot)$. The argument of $\Xi_\kappa[\cdot]$ corresponds to the proximity operator of $L_{2,1}$ norm, $\kappa \geq 1$ is a magnification factor for a debiasing

978-1-7281-9802-6/20/$31.00 ©2020 IEEE

TABLE I
MUSIC SIGNALS FOR EXPERIMENTS

Signal	Source 1	Source 2
Mixture A	Female($-50°$)	Female($45°$)
Mixture B	Female($-10°$)	Female($15°$)
bearlin1	Piano($0°$)	Vocal($-80°$)
bearlin2	Vocal($-80°$)	Ambient($60°$)
bearlin3	Piano($0°$)	Ambient($60°$)
ultimate1	Guitar($0°$)	Synth($-80°$)
ultimate2	Synth($-80°$)	Drum($60°$)
ultimate3	Guitar($0°$)	Drum($60°$)

operator $\Xi_\kappa[\cdot] = (\kappa z_{ijn}/\max_{ijn}\{z_{ijn}\})_-$, where $(\cdot)_-$ is $\min(1,\cdot)$. Θ_η in Eq. (3) is a frequency-wise weight where $\eta \geqq 0$ is the clipping parameter. When the sparsity in a frequency bin i is greater, the weight $(\Theta_\eta)_i$ is larger. This weight has a noise suppression effect by clipping.

$$(\Theta_\eta[\boldsymbol{x}])_i = (\Upsilon_\eta\left[\left(\Sigma_{m=1}^M\Sigma_{j=1}^J|x_{ijm}|^2\right)^{\frac{1}{2}}/\Sigma_{m=1}^M\Sigma_{j=1}^J|x_{ijm}|\right])_i, \quad (3)$$

where $\Upsilon_\eta[\cdot]$ denotes the clipped L_1 normalization with a threshold η as $\Upsilon_\eta[\boldsymbol{\xi}] = \boldsymbol{\xi}_\eta/\left(\|\boldsymbol{\xi}_\eta\|_1/I\right)$, $\boldsymbol{\xi}_\eta = (\boldsymbol{\xi}-\eta)_+$, where $\boldsymbol{\xi}$ is the argument of Υ_η and $\boldsymbol{\xi}_\eta$ is the clipped $\boldsymbol{\xi}$. This subtraction masks the frequency band where noise is dominant and target signals do not exist.

III. PROPOSED METHOD

A. Overview

Speech tends to have high power density in low-frequency bands. Therefore, we aim to keep the peaks of speech spectrum in the low frequency bands and suppress noise added by whitening in the high frequency bands. For that purpose, we introduce a clipping for Mel spectrogram.

B. Sparse IVA on the Mel scale

We applied Mel filter to a spectrum, as follows:

$$Mel[\boldsymbol{x}]_{cjm} = \sum_{i=1}^{I} H_c(i) * |x_{ijm}|. \quad (4)$$

where $H_c(i)$ is the c-th channel of the Mel filter. Then, instead of an observed signal $\boldsymbol{x}$ for the sparsity of the argument of $\Upsilon_\eta[\cdot]$ in Eq. (3), we use $Mel[\boldsymbol{x}]$,

$$\xi_c^{Mel} = \left(\Sigma_{m=1}^M\Sigma_{j=1}^J Mel[\boldsymbol{x}]_{cjm}^2\right)^{\frac{1}{2}}/\Sigma_{m=1}^M\Sigma_{j=1}^J Mel[\boldsymbol{x}]_{cjm}. \quad (5)$$

Then, an weight $\Theta_\eta[\boldsymbol{x}]]$, which is used a sparse IVA in Eq. (3), is calculated with ξ_i^{Mel} as follows:

$$(\Theta_\eta[Mel[\boldsymbol{x}]])_i = \Upsilon_\eta\left[\sum_{c=1}^{C} H_c(i)^{-1} * \xi_c^{Mel}\right]. \quad (6)$$

To match the number of dimensions between frequency-bins I and channels of the Mel scale C, ξ_c^{Mel} is converted by using inverse Mel filter again.

IV. EXPERIMENTS

A. Experimental conditions

In this section, we confirm the performance improvement of a sparse IVA that clips spectrum on the Mel scale through BSS experiments for speech signals and music signals composed of two sound sources with an observation of two microphones. We prepared Mixture A, and Mixture B composed of two female sources of dev1 in the UND task from the SiSEC database [5]. The music signals were created by dry instrumental sources: the Bearlin-Roads and Ultimate-Nz-Tour from the SiSEC database convolved with impulse responses (E2A) of the RWCP [7] database. The sampling rate was 16kHz, the frame size was 2048, and the shift size was 1024. The combination of sources and source direction are shown in Table I. We set an identity matrix to the initial demixing matrices W_i and zero vector to dual-value $\boldsymbol{d}$. The experimental parameters were $\mu_1 = \mu_2 = 1$, $\lambda_1 = 2$, $\lambda_2 = 0.01$, $\kappa = 1.1$, $\alpha = 0.5$. We set the clipping parameter with parameter $\delta = 0.5$ in Eq. (6) as follows: $\eta = \hat{\xi}_k, k = \lfloor \delta I \rfloor$,where $\hat{\xi}_k$ is ξ_i sorted in the ascending order and $\lfloor\cdot\rfloor$ is a flooring function. The performance was evaluated in terms of SDR [dB] [6].

TABLE II
EXPERIMENTAL RESULT EVALUATED IN TERMS OF SDRS [dB]

	Conventional	Avg(Conv)	Proposed	Aveg(Prop)
MixtureA	10.8	8.7	**11.0**	**9.3**
MixtureB	6.7		**7.7**	
bearlin1	**7.4**	6.3	7.2	6.3
bearlin2	**8.2**		8.0	
bearlin3	3.4		**3.7**	
ultimate1	4.5	5.3	**7.1**	**6.9**
ultimate2	6.2		**6.5**	
ultimate3	5.1		**7.1**	

B. Results and discussion

Table II shows the SDR, which is the average of two separated signals and the total average of speech and musical pieces. The proposed method improved the average SDR for speech signals by 0.6dB. Regarding music signals, although the average SDR of the proposed method to Bearlin-Roads was the same to that of the conventional method, the average SDR to Ultimate-Nz-Tour was improved by 1.6dB. In particular, for Ultimate1 and Ultimate 3, SDRs were improved by 2dB because the source composed of guitar was characteristically biased to low-frequency bands.

V. CONCLUSION

In this paper, to improve the BSS performance of sparse IVA, we proposed a sparse IVA that uses clipping for Mel spectrogram. This proposed method can suppress the noise added by whitening mainly in the higher frequency bands and maintain the source spectrum in the lower frequency bands. Experiments indicated that the proposed method was effective both for speech signals and for music signals whose power is biased to low frequency bands.

REFERENCES

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

[2] N. Ono, "Stable and fast update rules for independent vector analysis based on auxiliary function technique," in Proc. WASPAA, pp. 189-192, 2011.

[3] K. Yatabe, D. Kitamura, "Determined blind source separation via proximal splitting algorithm," in Proc. ICASSP, pp. 776-780, 2018.

[4] K. Yatabe, D. Kitamura, "Time-frequency-masking-based determined BSS with application to sparse IVA," in Proc. ICASSP, pp. 715-719, 2019.

[5] S. Araki, F. Nesta, E. Vincent, Z. Koldovsky, G. Nolte, A. Ziehe, A. Benichoux, "The 2011 signal separation evaluation campaign (SiSEC2011): -audio source separation -," in Proc. ICLVA, pp. 414-422, 2012.

[6] E. Vincent, H. Sawada, P. Bofill, S. Makino, J. P. Rosca, "First stereo audio source separation evaluation campaign: Data algorithms and results," in Proc. Int. Conf. Ind. Compon. Anal. Blind Source Separation (ICA'07), pp. 552-559, 2007.

[7] RWCP, "Sound scene database in real acoustic environment (RWCP-SSD)" Speech Resources Consortium, [http://research.nii.ac.jp/src/RWCP-SSD .html], accessed 2020/06/07.

978-1-7281-9802-6/20/$31.00 ©2020 IEEE