INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Speech Separation Using Independent Vector Analysis with an Amplitude Variable Gaussian Mixture Model

Zhaoyi Gu[1,2], *Jing Lu*[1], *Kai Chen*[1]

[1]Key Laboratory of Modern Acoustics, Institute of Acoustics, Nanjing University, Nanjing 210093, China

[2]Nanjing Institute of Advanced Artificial Intelligence, Nanjing 210014, China

guzhaoyi@smail.nju.edu.cn, lujing@nju.edu.cn, chenkai@nju.edu.cn

Abstract

Independent vector analysis (IVA) utilizing Gaussian mixture model (GMM) as source priors has been demonstrated as an effective algorithm for joint blind source separation (JBSS). However, an extra pre-training process is required to provide initial parameter values for successful speech separation. In this paper, we introduce a time-varying parameter in the GMM to adapt to the temporal power fluctuation embedded in the nonstationary speech signal so as to avoid the pre-training process. The expectation-maximization (EM) process updating both the demixing matrix and the signal model is altered correspondingly. Experimental results confirm the efficacy of the proposed method under random initialization and further show its advantage in terms of a competitive separation accuracy and a faster convergence speed.

Index Terms: joint blind source separation, independent vector analysis, Gaussian mixture model

1. Introduction

Blind source separation (BSS) is an important technique aiming at retrieving individual source signals from their mixtures without any information about the mixing system or the original sources [1]. Independent vector analysis (IVA) [2, 3], proposed as a multivariate extension of the well-studied independent component analysis (ICA) algorithm [4, 5], is regarded as a state-of-the-art formulation of joint blind source separation (JBSS) [6, 7] and receives a lot of interest in applications such as biological signal analyzing [8] and speech separation [9]. By imposing a multivariate prior distribution on each source, IVA exploits second and/or higher order dependencies across the datasets in source component vectors (SCV). Therefore, the permutation problem inherent to ICA can be solved within the separation process as IVA minimizes the inter-vector dependencies while maximizing the intra-vector mutual information. Occasionally, block permutation problems occur when the algorithm converges to a local optimum, and extra steps are required to improve the separation performance [10].

Conventional IVA models the source with a spherically symmetric Laplace (SSL) distribution [9, 11], as sources like speeches are generally spherical invariant [12]. Efficient and concise algorithms for SSL–IVA have been introduced in [13–15]. Nevertheless, SSL distribution has its limitations [19, 25]. First, it is symmetric over all datasets in the SCV. Second, distributions for different types of sources are identical. Third, it is isotropic and therefore possesses no second-order correlation. Fortunately, IVA shows great flexibility in choosing source priors, and numerous alternative models such as the harmonic model [16], the chain-like cliques [17, 18], the multivariate Gaussian distributions [19, 20], the Student's *t*-distribution [21], the Kotz distribution family [22], and the recently proposed M-PESM distribution family [23] have been introduced to exploit various spectral dependencies. In [24], relationship between the nonnegative matrix factorization (NMF) and IVA is revealed, and a new IVA framework using the NMF as the source model is presented.

Gaussian mixture model (GMM) is a commonly utilized generative model that can theoretically match any continuous distribution [26]. This has been exploited in an inspiring IVA method as proposed in [25], which introduces the GMM as the source prior. However, despite the flexibility of the GMM distribution, a pre-training process needs to be implemented in advance using target source information, which is unavailable in most practical situations. In this paper, we propose an amplitude variable Gaussian mixture model to remove the pre-training process. The expectation-maximization (EM) algorithm is applied to update both the demixing matrix and the signal model. The separation performance of the proposed scheme is compared with three other typical IVA using different signal models in a practical scenario.

2. The proposed method

2.1. Brief review of IVA

The signal model of determined situation is assumed, where an array of M microphones is utilized to capture the speech signals from M sources. After transforming the signals into short-time Fourier transform (STFT) domain and ignoring the noise, the mixing system can be represented using an instantaneous model as follows

$$\mathbf{x}_{ft} = \mathbf{A}_f \mathbf{s}_{ft}, 1 \le f \le F, 1 \le t \le T, \tag{1}$$

where $\mathbf{x}_{ft} = [x_{ft}^{[1]}, x_{ft}^{[2]}, \ldots, x_{ft}^{[M]}]^{\mathrm{T}}$, $\mathbf{s}_{ft} = [s_{ft}^{[1]}, s_{ft}^{[2]}, \ldots, s_{ft}^{[M]}]^{\mathrm{T}}$ are the multichannel observed and source signals respectively, and $\mathbf{A}_f$ is an $M \times M$ mixing matrix. The subscripts f and t are used to denote the indices of frequency bins and time frames respectively and the corresponding F and T denote the total frequency and frame number. The superscript $[i]$ denotes the index of the source or microphone and $[\cdot]^{\mathrm{T}}$ is a notation for non-conjugate transposition. When $\mathbf{A}_f$ is invertible, the estimated signals $\mathbf{y}_{ft}$ can be obtained using a demixing matrix $\mathbf{W}_f = \mathbf{A}_f^{-1}$ by

$$\mathbf{y}_{ft} = \mathbf{W}_f \mathbf{x}_{ft}. \tag{2}$$

http://dx.doi.org/10.21437/Interspeech.2019-2076

The core concept of IVA is to maximize the independence between the estimated SCVs. Using Kullback-Leibler (KL) divergence, this problem can be reified as minimizing the following cost function

$$\begin{aligned}\mathcal{J}(\mathbf{W},\boldsymbol{\theta}) &= \sum_i H\left(\mathbf{y}_t^{[i]}\right) - H\left(\mathbf{y}_t^{[1]},\mathbf{y}_t^{[2]},\cdots,\mathbf{y}_t^{[M]}\right) \\ &\simeq -\frac{1}{T}\sum_t\sum_i \log\left(p_{\mathbf{s}_t^{[i]}}\left(\mathbf{y}_t^{[i]}\right)\right) - \sum_f \log\left|\det \mathbf{W}_f\right|,\end{aligned} \tag{3}$$

where $\mathbf{y}_t^{[i]} = [y_{1t}^{[i]}, y_{2t}^{[i]}, \ldots, y_{Ft}^{[i]}]^{\mathrm{T}}$ and $\mathbf{s}_t^{[i]} = [s_{1t}^{[i]}, s_{2t}^{[i]}, \ldots, s_{Ft}^{[i]}]^{\mathrm{T}}$ are the estimated and original SCV respectively, $\boldsymbol{\theta}$ contains all model parameters, $H(\cdot)$ is the Shannon entropy function, $p_z(\cdot)$ is the probability density function (PDF) of random variable z, and $\det(\cdot)$ denotes the determinant of a given matrix. IVA retains frequency dependencies by modeling all frequency components using one multivariate PDF and the KL divergence of the estimated sources will be minimized to zero if and only if the outputs $\mathbf{y}_t^{[i]}$ ($i = 1, 2, \ldots, M$) are mutually independent [9].

2.2. Amplitude variable Gaussian mixture model

Suppose that the frequency structure of a given signal can be jointly represented by D different spectral patterns, then the interdependency among frequency bins can be represented using a multivariate complex Gaussian mixture model as follows

$$p\left(\mathbf{s}_t^{[i]}\right) = \sum_{d^{[i]}=1}^{D^{[i]}} p\left(d^{[i]}\right)\prod_f \mathcal{N}_c\left(s_{ft}^{[i]}\middle|0, v_{fd}^{[i]}h_t^{[i]}\right), \tag{4}$$

where $d^{[i]}$ and $D^{[i]}$ are the index and total number of the spectral patterns, $\mathcal{N}_c(\cdot)$ is the notation for circularly-symmetric Gaussian distribution, and $p(d^{[i]})$ represents the prior probability of state $d^{[i]}$. In this model, the precision of $s_{ft}^{[i]}$ under the $d^{[i]}$th state is characterized by the product of two parameters, $v_{fd}^{[i]}$ and $h_t^{[i]}$, where the former captures the shared patterns over different time frames and the latter serves as an amplitude adjusting factor compensating for the amplitude difference between the estimated $v_{fd}^{[i]}$ and the instantaneous output signal power $|y_{ft}^{[i]}|^2$.

The proposed model is an improved version of the GMM discussed in [25] by including the time-varying term $h_t^{[i]}$, and we name the model described by (4) as the amplitude variable Gaussian mixture model (AV–GMM). Note that although distributions under different states are factorized, the joint distribution is not, so that the nonlinear inter-vector dependencies are well preserved in the ultimate mixture model. In the AV–GMM, the time-varying parameter $h_t^{[i]}$ is added to model energy fluctuation in the time domain, which is an inherent feature in nonstationary signals.

2.3. Objective function

Using the introduced cost function in (3), the specific objective function to be maximized can be derived as

$$\begin{aligned}\mathcal{L}_{\boldsymbol{\theta}} = &\sum_t \log\left\{\sum_{\mathbf{d}} p(\mathbf{d})\prod_f \mathcal{N}_c\left(\mathbf{y}_{ft}\middle|0,\boldsymbol{\Phi}_{f\mathbf{d}}\right)\right\} \\ &+ T\sum_f \log\left|\det \mathbf{W}_f\right|, \quad s.t.\ \sum_{\mathbf{d}} p(\mathbf{d}) = 1.\end{aligned} \tag{5}$$

where $\mathbf{d}$ is an M-dimensional vector whose ith element corresponds to the state index $d^{[i]}$, $p(\mathbf{d})$ is the joint probability representing $\prod_i p(d^{[i]})$, and $\boldsymbol{\Phi}_{f\mathbf{d}}$ is a diagonal precision matrix of which the ith diagonal element is $v_{fd}^{[i]}\, h_t^{[i]}$ because of the independence assumption.

It can be found that (5) is equivalent to the likelihood function of the observed signal [9]. In this paper, we consider the 2 × 2 determined case, where two sources and two microphones are presented.

2.4. Pre-whitening and post-processing steps

In order to facilitate the optimization, the output signals $\mathbf{y}_{ft}$ are kept spatially white since uncorrelation is a natural corollary of independence. By enforcing a pre-whitening procedure on the observed signals $\mathbf{x}_{ft}$ and restricting the demixing matrix $\mathbf{W}_f$ to be unitary, the elements in $\mathbf{y}_{ft}$ become uncorrelated.

Let $\mathbf{W}_f$ have the form

$$\mathbf{W}_f = \begin{bmatrix} a_f & b_f \\ -b_f^* & a_f^* \end{bmatrix}, \tag{6}$$

where $(\cdot)^*$ denotes complex conjugation, then the number of unknown variables in $\mathbf{W}_f$ can be reduced by half under the constraint $|a_f|^2 + |b_f|^2 = 1$. Let $\mathbf{C}_f$ denote the approximate correlation matrix for the observed signals, i.e.

$$\mathbf{C}_f = \frac{1}{T}\sum_t \mathbf{x}_{ft}\mathbf{x}_{ft}^{\mathrm{H}}, \tag{7}$$

then the pre-whitening process is conducted by

$$\tilde{\mathbf{x}}_{ft} = \mathbf{C}_f^{-0.5}\mathbf{x}_{ft}, \tag{8}$$

where $[\cdot]^{\mathrm{H}}$ is a notation for conjugate transposition and $\tilde{\mathbf{x}}_{ft}$ denotes the whitened signals.

After learning the separation matrix $\mathbf{W}_f$, a post-processing step is required to recover the spectral dynamic from the whitening procedure, as well as to correct the scaling problems inherent to IVA, and the final output $\hat{\mathbf{y}}_{ft}$ is obtained by adopting the minimal distortion principle [27] as

$$\hat{\mathbf{y}}_{ft} = \mathrm{diag}\left(\mathbf{C}_f^{0.5}\mathbf{W}_f^{-1}\right)\tilde{\mathbf{y}}_{ft}, \tag{9}$$

where diag($\mathbf{A}$) is a matrix operator that sets all the off-diagonal elements of $\mathbf{A}$ to zero.

2.5. Optimization using the EM algorithm

The EM algorithm [28] is an efficient tool to deal with inference problems. Applying Jensen's inequality to (5), and considering the unitary restriction on $\mathbf{W}_f$, we have

$$\mathcal{L}_{\boldsymbol{\theta}} \geq \sum_t\sum_{\mathbf{d}} q_t(\mathbf{d})\log\left\{p(\mathbf{d})\prod_f \mathcal{N}_c\left(\tilde{\mathbf{y}}_{ft}\middle|0,\boldsymbol{\Phi}_{f\mathbf{d}}\right)\right\} \triangleq \mathcal{Q}_{\boldsymbol{\theta}}, \tag{10}$$

where $q_t(\mathbf{d})$ is the posterior probability of state $\mathbf{d}$ given the observed $\mathbf{x}_{ft}$. In the expectation step, it is estimated as

$$q_t(\mathbf{d}) = \frac{p(\mathbf{d})\prod_f \mathcal{N}_c\left(\tilde{\mathbf{y}}_{ft}\middle|0,\boldsymbol{\Phi}_{f\mathbf{d}}\right)}{\sum_{\mathbf{d}'} p(\mathbf{d}')\prod_f \mathcal{N}_c\left(\tilde{\mathbf{y}}_{ft}\middle|0,\boldsymbol{\Phi}_{f\mathbf{d}'}\right)}. \tag{11}$$

In the maximization step, the model parameters are updated by maximizing the right-hand side of (10), i.e. $\mathcal{Q}_{\boldsymbol{\theta}}$.

Rewriting $\mathcal{Q}_{\boldsymbol{\theta}}$ as a function of $\mathbf{W}_f$ and using the Lagrangian multiplier method, we get

$$\mathcal{Q}_{\boldsymbol{\theta}} \overset{c}{=} -\sum_{t,\mathbf{d}} q_t(\mathbf{d})\left(\tilde{\mathbf{x}}_{ft}^{\mathrm{H}}\mathbf{W}_f^{\mathrm{H}}\boldsymbol{\Phi}_{ft\mathbf{d}}\mathbf{W}_f\tilde{\mathbf{x}}_{ft}\right)+\lambda_f\left(\left|a_f\right|^2+\left|b_f\right|^2\right). \quad (12)$$

where λ_f is the Lagrange multiplier. A similar trick used in [25] is applied to simplify the notation. By dividing the precision matrix $\boldsymbol{\Phi}_{ft\mathbf{d}}$ into the following form

$$\boldsymbol{\Phi}_{ft\mathbf{d}} = \begin{bmatrix} v_{fd}^{[1]}h_t^{[1]} - v_{fd}^{[2]}h_t^{[2]} & 0 \\ 0 & 0 \end{bmatrix} + v_{fd}^{[2]}h_t^{[2]}\mathbf{I}\,, \quad (13)$$

and defining $\mathbf{w}_f$ as $[a_f, b_f]^{\mathrm{H}}$, a compact expression of $\mathcal{Q}_{\boldsymbol{\theta}}$ can be written as follows

$$\mathcal{Q}_{\boldsymbol{\theta}} \overset{c}{=} -\sum_{t,\mathbf{d}} q_t(\mathbf{d})\phi_{ft\mathbf{d}}\mathbf{w}_f^{\mathrm{H}}\tilde{\mathbf{x}}_{ft}\tilde{\mathbf{x}}_{ft}^{\mathrm{H}}\mathbf{w}_f + \lambda_f\mathbf{w}_f^{\mathrm{H}}\mathbf{w}_f\,, \quad (14)$$

where $\phi_{ft\mathbf{d}}$ is the estimated precision difference between two sources that equals $v_{fd}^{[1]}h_t^{[1]} - v_{fd}^{[2]}h_t^{[2]}$. Taking the derivative of (14) with respect to $\mathbf{w}_f^*$ and set it to zero, we obtain $\mathbf{M}_f\mathbf{w}_f = \lambda_f\mathbf{w}_f$, where

$$\mathbf{M}_f = \sum_{t,\mathbf{d}} q_t(\mathbf{d})\phi_{ft\mathbf{d}}\tilde{\mathbf{x}}_{ft}\tilde{\mathbf{x}}_{ft}^{\mathrm{H}}\,. \quad (15)$$

Since the optimization goal is to maximize (10), the updated value of $\mathbf{w}_f$ is exactly the eigenvector of $\mathbf{M}_f$ corresponding to the smaller eigenvalue.

$v_{fd}^{[i]}$ and $h_t^{[i]}$ can also be updated by setting the derivatives of $\mathcal{Q}_{\boldsymbol{\theta}}$ to zero. After some straightforward mathematical manipulations, updating rules for $v_{fd}^{[i]}$ and $h_t^{[i]}$ become

$$v_{fd}^{[i]} = \frac{\sum_t q_t(\mathbf{d}_i = d)}{\sum_t q_t(\mathbf{d}_i = d)h_t^{[i]}\left|\tilde{y}_{ft}^{[i]}\right|^2}\,, \quad (16)$$

$$h_t^{[i]} = \frac{F}{\sum_{f,d^{[i]}} q_t(\mathbf{d}_i = d)v_{fd}^{[i]}\left|\tilde{y}_{ft}^{[i]}\right|^2}\,, \quad (17)$$

where $q_t(\mathbf{d}_i = d)$ is the marginal posterior probability of $d^{[i]} = d$ for the tth frame.

Similarly, the updating rules for the prior probability can be derived as

$$p(\mathbf{d}) = \frac{1}{T}\sum_t q_t(\mathbf{d})\,. \quad (18)$$

The EM algorithm iterates alternatively between the expectation and the maximization step until a global or local maximum of the objective function is found.

2.6. Discussion

The Gaussian mixture model utilized in [25] for IVA is time irrelevant. Despite the whitening process mentioned in (8), the energy fluctuation over time is not negligible because (7) is an inaccurate estimation for nonstationary sources. Therefore, apart from capturing spectral patterns in the frequency domain, parameter $v_{fd}^{[i]}$ is also responsible for energy compensation over time, making it hard to find an effective estimation of the frequency structures under random initialization. The proposed model, on the other hand, considers the nonstationary property of speech signals. An unbiased estimation of $v_{fd}^{[i]}$ and a stable estimation of the demixing matrix can be learned effectively without the pre-training process by introducing the time-varying parameter that can address the temporal feature.

3. Experiment

3.1. Configurations

This section presents performance evaluations of the proposed model under a practical scenario where two microphones are used to capture the signals from two main speakers, along with a background interfering speech source of a relatively low energy level. The microphones used for recording are free-field 1/2 inch CHZ-223 microphone with a sensitivity level of 39.0 mV/Pa. The results are evaluated by signal-to-interference ratio (SIR) and signal-to-distortion ratio (SDR) in decibels using the BSS_EVAL toolbox [29]. In all our evaluation sets, both the target and interfering sources are ten-second-long speech signals selected from the TIMIT database [30]. All the speech segments are recorded separately with various incident angles in a room of size of 8 m × 6 m × 3 m and reverberation time of 300 ms. The schematic diagram of the source-microphone layout is given in Figure 1. Totally, 30 mixed signals are generated where the target sources have similar energy levels and the SIR between the target and the interference source varies from 6 dB to 14 dB. Other common configurations for the algorithms are listed in Table I. Exemplary audio samples are available online at "https://github.com/annie-gu/AV-GMM-IVA".

Table 1: *Common configurations in algorithms*

Parameters	Value
Sampling frequency	16 kHz
Frame length	128 ms
Frame overlap	75 %
Window type	Hanning

3.2. Experimental results

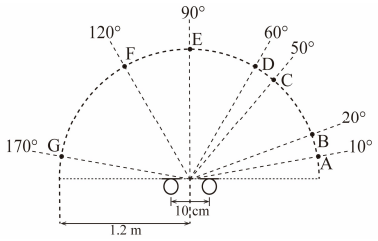

Figure 1: *Diagram of the source-microphone layout.*

We compare the separation performance of IVA algorithms with four different source models, namely, the proposed AV–GMM–IVA, the GMM–IVA proposed in [25], the conventional SSL–IVA with a spherical symmetric Laplace prior [11], and the NMF–IVA [24], a recently proposed flexible framework which combines nonnegative matrix factorization with IVA. Model parameters in the AV–GMM–IVA, the GMM–IVA, and the NMF–IVA are all initialized using random values drawn from a uniform distribution between 0.999 and 1.001. The number of bases for the NMF is set to be 2, and the number of the Gaussian models for the AV–GMM and the GMM are set

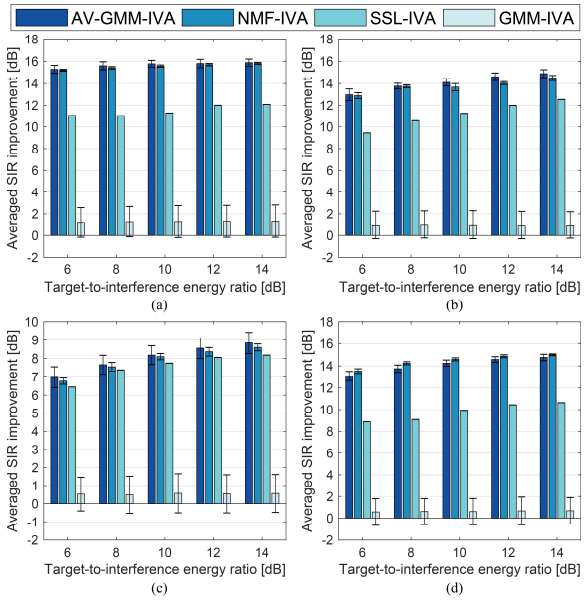

Figure 2: *Averaged SIR improvements of four IVA models under four different source-microphone configurations*

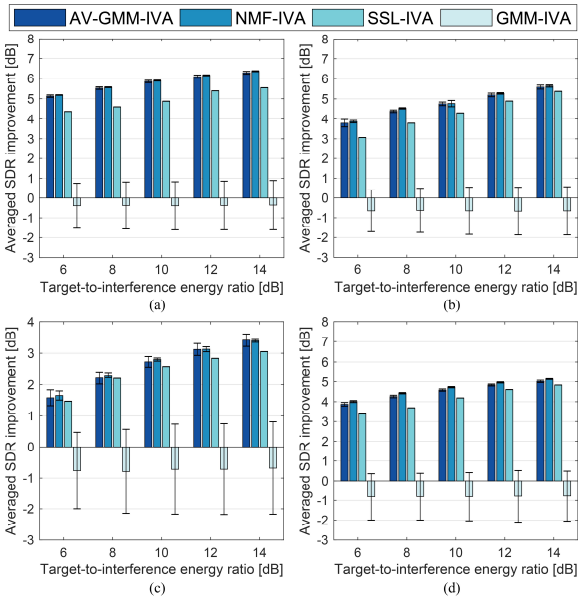

Figure 3: *Averaged SDR improvements of four IVA models under four different source-microphone configurations*

to be 2 and 15 respectively. All the 30 mixtures are tested and for each sample, we conduct 10 trials with different pseudo-random seeds.

Figure 2 and 3 present the average SIR and SDR metrics over both outputs of all samples, and the standard deviations in 10 trials are illustrated by the length of the error bar. We evaluate the separation performance under five different target-to-interference energy ratios and four different location settings. Each subgraph represents the performance under a specific source-microphone configuration as listed in Table 2. All algorithms stop updating when the decrement of the corresponding cost functions between adjacent iterations becomes smaller than 10^{-6}. It can be seen that without a proper pre-training stage, the conventional GMM–IVA can barely separate the sources under random initialization. On the other hand, the AV-GMM-IVA effectively separates the speech and obtains a similar performance compared to the NMF-IVA with respect to SIR and SDR in all configurations, and both of these two algorithms show significant superiority over the conventional SSL–IVA in configurations (a) and (b). It should be noted that unlike the case in the SSL–IVA, the other three algorithms are all influenced by random initializations, and methods based on the GMM are more sensitive to the initial values.

Figure 4 depicts the averaged SIR convergence over the 30 mixtures and four source-microphone configurations with an initial target-to-interference energy ratio of 10 dB. Note that when the base and state numbers are small, computational complexities for both NMF–IVA and the proposed method are of O(*FT*). It can be seen that significantly fewer iterations are

Table 2: *Relative source-microphone position in experiments*

Label	source 1	source 2	Interference
a	G	C	F
b	F	C	G
c	D	B	F
d	E	A	C

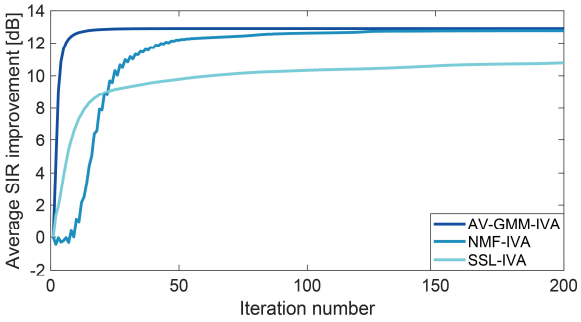

Figure 4: *Averaged SIR convergence of three algorithms.*

needed for the proposed AV-GMM-IVA to reach the steady state. On average, it takes AV–GMM–IVA less than 30 iterations to meet the convergence criterion, while 172 and 286 iterations are needed for NMF–IVA and SSL–IVA to converge respectively.

4. Conclusions

This paper proposes an amplitude variable Gaussian mixture model as an improvement to the time-invariant GMM for joint speech source separation using IVA. A time-varying parameter is utilized to adapt to the temporal signal power fluctuation and remove the pre-training process required by the normal GMM-IVA algorithm. The updating of both the model parameters and the demixing matrix are derived using the EM scheme. Experimental results show that the proposed method has a competitive separation performance and a faster convergence speed compared to conventional IVA algorithms.

5. Acknowledgements

We would like to thank Dr. Jiucang Hao, the first author of Ref. 25, for his kind help in implementing the related algorithm. This work was supported by the National Natural Science Foundation of China (Grant No.11874219)

6. References

[1] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent component analysis and applications.* Academic press, 2010.

[2] T. Kim, T. Eltoft, and T. W. Lee, "Independent vector analysis: An extension of ICA to multivariate components," in *International Conference on Independent Component Analysis and Signal Separation*, *March 5-8, Charleston*, *SC*, *USA*, 2006, pp. 165–172.

[3] T. Kim, I. Lee, and T. W. Lee, "Independent vector analysis: definition and algorithms," in *2006 Fortieth Asilomar Conference on Signals, Systems and Computers, October 29-November 1, Pacific Grove, CA, USA, Proceedings*, 2006, pp. 1393–1396.

[4] P. Comon, "Independent component analysis, a new concept?" *Signal processing,* vol. 36, no. 3, pp. 287–314, 1994.

[5] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks,* vol. 13, no. 4–5, pp. 411–430, 2000.

[6] A. Weiss, S. A. Cheema, M. Haardt, and A. Yeredor, "Performance analysis of the Gaussian quasi-maximum likelihood approach for independent vector analysis," *IEEE Transactions on Signal Processing,* vol. 66, no. 19, pp. 5000–5013, 2018.

[7] J. Čmejla, T. Kounovský, J. Málek, and Z. Koldovský, "Independent vector analysis exploiting pre-learned banks of relative transfer functions for assumed target's positions," in *International Conference on Latent Variable Analysis and Signal Separation, July 2-6, Guildford, UK*, 2018, Springer, pp. 270–279.

[8] Y. O. Li, T. Adali, W. Wang, and V. D. Calhoun, "Joint blind source separation by multiset canonical correlation analysis," *IEEE Transactions on Signal Processing,* vol. 57, no. 10, pp. 3918–3929, 2009.

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

[10] Y. Liang, "Enhanced independent vector analysis for audio separation in a room environment," September 2013, pp. 69–90.

[11] T. Kim, H. T. Attias, S. Y. Lee, and T. W. Lee, "Blind source separation exploiting higher-order frequency dependencies," *IEEE Transactions on Audio Speech & Language Processing,* vol. 15, no. 1, pp. 70–79, 2006.

[12] I. Lee and T. W. Lee, "On the assumption of spherical symmetry and sparseness for the frequency-domain speech model," *IEEE Transactions on Audio Speech & Language Processing,* vol. 15, no. 5, pp. 1521–1528, 2007.

[13] I. Lee, T. Kim, and T. W. Lee, "Fast fixed-point independent vector analysis algorithms for convolutive blind source separation," *Signal Processing,* vol. 87, no. 8, pp. 1859–1871, 2007.

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

[15] M. Anderson, "Independent vector analysis: Theory algorithms and applications," May 2013, pp. 104–112.

[16] C. H. Choi, W. Chang, and S. Y. Lee, "Blind source separation of speech and music signals using harmonic frequency dependent independent vector analysis," *Electronics letters,* vol. 48, no. 2, pp. 124–125, 2012.

[17] I. Lee, G. J. Jang, and T. W. Lee, "Independent vector analysis using densities represented by chain-like overlapped cliques in graphical models for separation of convolutedly mixed signals," *Electronics letters,* vol. 45, no. 13, pp. 710–711, 2009.

[18] I. Lee and G. J. Jang, "Independent vector analysis based on overlapped cliques of variable width for frequency-domain blind signal separation," *EURASIP Journal on Advances in Signal Processing,* vol. 2012, no. 1, pp. 113, 2012.

[19] M. Anderson, T. Adali, and X. L. Li, "Joint blind source separation with multivariate Gaussian model: Algorithms and performance analysis," *IEEE Transactions on Signal Processing,* vol. 60, no. 4, pp. 1672–1683, 2012.

[20] M. Anderson, X. L. Li, and T. Adalı, "Complex-valued independent vector analysis: Application to multivariate Gaussian model," *Signal Processing,* vol. 92, no. 8, pp. 1821–1831, 2012.

[21] Y. Liang, G. Chen, S. Naqvi, and J. A. Chambers, "Independent vector analysis with multivariate student's *t*-distribution source prior for speech separation," *Electronics Letters,* vol. 49, no. 16, pp. 1035–1036, 2013.

[22] M. Anderson, G. S. Fu, R. Phlypo, and T. Adali, "Independent vector analysis, the Kotz distribution, and performance bounds," in *ICASSP, May 26-31, Vancouver, Canada*, 2013, pp. 3243–3247.

[23] R. Giri, B. D. Rao, and H. Garudadri, "Reweighted algorithms for independent vector analysis," *IEEE Signal Processing Letters,* vol. 24, no. 4, pp. 362–366, 2017.

[24] D. Kitamura, N. Ono, H. Sawada, H. Kameoka, and H. Saruwatari, "Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech and Language Processing,* vol. 24, no. 9, pp. 1622–1637, 2016.

[25] J. Hao, I. Lee, T. W. Lee, and T. J. Sejnowski, "Independent vector analysis for source separation using a mixture of Gaussians prior," *Neural computation,* vol. 22, no. 6, pp. 1646–1673, 2010.

[26] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006, pp. 110–113.

[27] K. Matsuoka and S.Nakashima, "Minimal distortion principle for blind source separation," in *Proc. Int. Conf. Independent Compon. Anal. Blind Source Separation,* 2001, pp. 722–727.

[28] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*. John Wiley & Sons, 2007.

[29] 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, 2006.

[30] J. S. Garofolo et al., "TIMIT acoustic-phonetic continuous speech corpus," in *Linguistic Data Consortium,* 1993.