An Introduction to Energy-Based Blind Separating Algorithm for Speech Signals

Mahdi Mahdikhani and Mohammad Hossein Kahaei

We introduce the Energy-Based Blind Separating (EBS) algorithm for extremely fast separation of mixed speech signals without loss of quality, which is performed in two stages: iterative-form separation and closed-form separation. This algorithm significantly improves the separation speed simply due to incorporating only some specific frequency bins into computations. Simulation results show that, on average, the proposed algorithm is 43 times faster than the independent component analysis (ICA) for speech signals, while preserving the separation quality. Also, it outperforms the fast independent component analysis (FastICA), the joint approximate diagonalization of eigenmatrices (JADE), and the second-order blind identification (SOBI) algorithm in terms of separation quality.

Keywords: Blind speech separation, ICA, frequency bin.

I. Introduction

Blind source separation (BSS) is used for separating mixed signals received by sensors with a little knowledge about the mixing model and properties of source signals [1]. A number of BSS algorithms are designed using an independent component analysis (ICA) algorithm [2]. The conventional ICA is a maximum likelihood algorithm based on the minimization of mutual information approach [3] and uses the natural gradient method [4]. In [5], we presented the Variable Situated Matrix (VSM) technique to speed up ICA without loss of quality, which led to the VICA (VSM+ICA) algorithm. In VICA, in all learning steps of each frequency bin, the separation quality of the separating matrix is compared with the situated matrix and the best one is selected as an initial separating matrix in the next learning step. On average, VICA is about six times faster than ICA, while preserving the separation quality.

In this work, we introduce the Energy-Based Blind Separating (EBS) algorithm for very fast separation of mixed speech signals. EBS is introduced because the energy of mixed speech signals is concentrated in a few frequency bins, which means a better separation quality can be obtained because these more energetic bins have a higher signal-to-noise ratio (SNR) than that of the less energetic ones. In EBS, the separation procedure is first performed iteratively using VICA for some individual frequency bins, and the attenuation and delay matrices are estimated. Next, these matrices are applied to other frequency bins to separate the signals in closed form. Here, we only consider the speech signals that are mixed according to the anechoic mixture model. Also, we assume that the number of sources and number of sensors are equal.

II. Mixture Model

We intend to separate the mixed speech signals received from N sources by a uniform linear array (ULA) with M sensors located linearly with an equal space of δ. As a normal condition of ICA, we assume the determined case $M = N$, and the over-determined case in which $M > N$ is easily reduced to a determined case by selecting N sensors or applying more sophisticated preprocessing, such as the principal component analysis (PCA) algorithm. The received signal by the m-th sensor is defined as

$$x_m(t) = \sum_{n=1}^{N} a_{mn} s_n(t - d_{mn}),\ 1 \le m \le M, \qquad (1)$$

where a_{mn} and d_{mn} respectively show the attenuation and time delay of the n-th source with respect to the m-th sensor.

Manuscript received Mar. 17, 2013; revised July 16, 2013; accepted Dec. 23, 2013.

Mahdi Mahdikhani (phone: +98 21 7322566, m_mehdikhani2000@yahoo.com) and Mohammad Hossein Kahaei (kahaei@iust.ac.ir) are with the Department of Communication Engineering, Iran University of Science & Technology, Tehran, Iran.

$\mathbf{A}=[a_{mn}]$ and $\mathbf{D}=[d_{mn}]$, in which $1 \le m,n \le N$, denote the attenuation and delay matrices, respectively [6]. In the frequency domain, the sources and sensors signals are related to each other by mixing matrix $\mathbf{H}_r=[h_r^{mn}]$, defined as $h_r^{mn}=a_{mn}\exp(-j2\pi r\Delta f d_{mn})$ in the r-th frequency bin, where Δf is the frequency resolution and $1 \le r \le R$, for which R is the number of frequency bins.

III. Introducing EBS Algorithm

In Fig. 1, the average spectral energy distribution (SED) of the mixed speech signals is shown for N=2, N=3, and N=4. The sampling rate and the number of discrete Fourier transform (DFT) points are 16 kHz and 512, respectively. A great amount of energy of mixed speech signals is concentrated in a few frequency bins, and we expect that a better separation quality can be obtained because these more energetic bins have a higher SNR than that of less energetic ones.

This fact motivates us to develop EBS, in which, unlike the conventional frequency-domain separating algorithms, frequency bins are divided into two sets with different separating procedures: selected frequency bins (SFB) and remaining frequency bins (RFB). In each bin of an SFB set, mixed speech signals are separated iteratively using VICA [7] and corresponding attenuation and delay matrices are estimated. The final $\hat{\mathbf{A}}=[\hat{a}_{mn}]$ and $\hat{\mathbf{D}}=[\hat{d}_{mn}]$ are then estimated by weighted averaging over all attenuation and delay matrices of all bins of the SFB set, respectively. Accordingly, this stage is referred to as iterative-form separation (IS). EBS has two stages: the IS stage and the closed-form separation (CS) stage. Having estimated $\hat{\mathbf{A}}$ and $\hat{\mathbf{D}}$, the separating matrix of each frequency bin of the RFB set is formed to carry out CS. By applying CS instead of IS, a very fast separation is achieved.

In the following, we mathematically express EBS for a considered SFB set consisting of p percent of the energy of the given mixed signals. According to the cumulative SED (CSED) of the given signals, this SFB set includes 1st to (K_p)th bins, as shown in Fig. 2.

1. IS Stage

The procedure of the IS stage is itemized as follows.

Step 1. Assuming $0 < p < 100$, find the corresponding K_p according to the CSED of the given mixed signals and consider 1st to (K_p)th bins as an SFB set. Obviously, the other bins belong to the RFB set.

Step 2. Calculate the source angle $\breve{\boldsymbol{\theta}}=[\breve{\theta}_1 \ \ldots \ \breve{\theta}_N]$ using the DESPRIT algorithm [8]. The $\breve{\boldsymbol{\theta}}$ is constant and independent of the frequency bin.

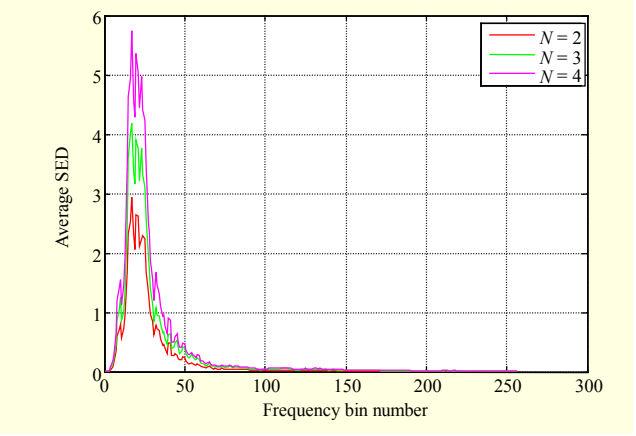

Fig. 1. Average SED of mixed speech signals vs. frequency bin number for N=2, N=3, and N=4.

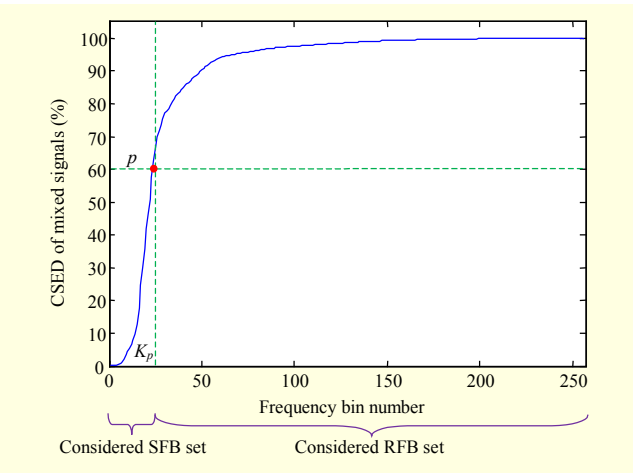

Fig. 2. Considered SFB set consists of p percent of energy of given mixed signals, including 1st to (K_p)th frequency bins, based on CSED.

Step 3. Apply VICA to each bin of the SFB set to iteratively calculate separating matrix $\mathbf{W}_r$, $1 \le r \le K_p$ [5] and then separate the mixed signals for these K_p bins.

Step 4. Obtain mixing matrices $\hat{\mathbf{H}}_r=(\mathbf{W}_r)^{-1}$, $1 \le r \le K_p$.

Step 5. Estimate angle vectors $\hat{\boldsymbol{\theta}}_r=[\hat{\theta}_1^r \ \ldots \ \hat{\theta}_N^r]$, $1 \le r \le K_p$ according to [7].

$$\hat{\theta}_n^r=\cos^{-1}\left(\frac{angle(\hat{h}_r^{2n}/\hat{h}_r^{1n})}{2\pi r\Delta f v_c^{-1}\delta}\right), \quad 1 \le r \le K_p, 1 \le n \le N, \qquad (2)$$

where $\hat{h}_r^{mn}$ is the (m,n)th component of $\hat{\mathbf{H}}_r$, v_c is the velocity of sound, and $angle(.)$ is the angle operator. Unlike $\breve{\boldsymbol{\theta}}$, $\hat{\boldsymbol{\theta}}_r$ is dependent on the frequency bin.

Step 6. Compute weight factors $\kappa(\breve{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}_r)$, defined as

$$\kappa(\breve{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}_r)=\begin{cases}1 & |\breve{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_r| \le \sigma \\ 0 & \text{other}\end{cases}, \quad 1 \le r \le K_p, \qquad (3)$$

where σ is a constant threshold. Here, we consider $\sigma=10^\circ$.

http://dx.doi.org/10.4218/etrij.14.0213.0130

Step 7. Estimate the components of attenuation and delay matrices $\hat{a}_{mn}^{r} = |\hat{h}_{r}^{mn}|$ and $\hat{d}_{mn}^{r} = (-1/2\pi r\Delta f)\angle(\hat{h}_{r}^{mn})$ in each bin of the SFB set for $1 \le m,n \le M$, in which $|.|$ and $\angle(.)$ denote amplitude and phase operators, respectively.

Step 8. Compute the final $\hat{a}_{mn}$ and $\hat{d}_{mn}$ by weighted averaging $\hat{a}_{mn}^{r}$ and $\hat{d}_{mn}^{r}$ over K_p bins of the SFB set according to

$$\hat{a}_{mn} = \sum_{r=1}^{K_p} \kappa(\breve{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_r)\hat{a}_{mn}^{r} \Big/ \sum_{r=1}^{K_p} \kappa(\breve{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_r), \quad (4)$$

$$\hat{d}_{mn} = \sum_{r=1}^{K_p} \kappa(\breve{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_r)\hat{d}_{mn}^{r} \Big/ \sum_{r=1}^{K_p} \kappa(\breve{\boldsymbol{\theta}}, \hat{\boldsymbol{\theta}}_r). \quad (5)$$

Step 9. Form the final estimation of the attenuation and delay matrices $\hat{\mathbf{A}} = [\hat{a}_{mn}]$ and $\hat{\mathbf{D}} = [\hat{d}_{mn}]$.

2. CS Stage

After estimating $\hat{\mathbf{A}}$ and $\hat{\mathbf{D}}$, the RFB set including $R - K_p$ bins from $K_p + 1$ to R are separated by the CS stage as described below.

Step 1. Form the mixing matrix $\hat{\mathbf{H}}_r$ whose components are $\hat{h}_{r}^{mn} = \hat{a}_{mn}\exp\{-j2\pi r\Delta f\hat{d}_{mn}\}$, $K_p+1 \le r \le R$.

Step 2. Obtain separating matrix $\mathbf{W}_r = (\hat{\mathbf{H}}_r)^{-1}$, $K_p+1 \le r \le R$ and separate the signals of the remaining $R - K$ bins of the RFB set.

Implementation of this stage is very fast because IS is replaced by CS. After separating the signals from IS and CS stages and solving the permutation problem based on the direction of arrival approach [9], we apply the inverse short-time Fourier transform to estimate all separated signals in the time domain.

IV. Simulation Results

In all the experiments, three-second-long speech signals are selected from a standard database. The distance between two adjacent sensors is 2 cm, and sources are located 1.5 m away from the 1st sensor. These results are averaged over 50,000 independent trials for each experiment. The signal-to-distortion ratio (SDR) and perceptual estimation of speech quality (PESQ) are calculated to measure the separation quality [9]. EBS can work with any SFB set (any p), but, in practice, we only use specific sets of bins defined as the best SFB (bSFB) sets for which EBS yields the best performance. Accordingly, the simulation results are presented in two phases: the training phase and the test phase.

1. Training Phase

In the training phase, the bSFB sets for different N are obtained based on comparing the results of several experiments. To elaborate, the procedure of finding the bSFB set is explained in the following for $N=3$.

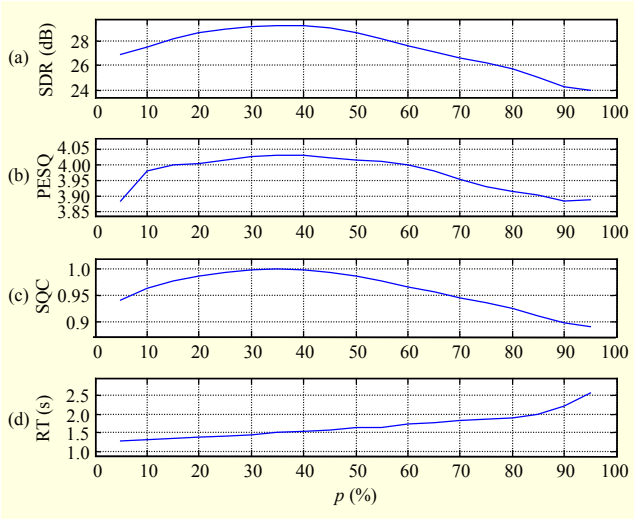

Fig. 3. Average (a) SDR, (b) PESQ, (c) SQC, and (d) RT of EBS algorithm for different SFB sets (different p) for N=3.

Step 1. From training speech signals, $N=3$ signals are randomly selected. Then, N mixed signals are synthetically made.

Step 2. The SFB set consisting of $0 < p \le 100$ percent of the energy of the given mixed signals is considered. Based on the CSED, this SFB set includes the 1st to (K_p)th frequency bins.

Step 3. EBS is applied to considered N mixed signals (from Step 1) and the SFB set (from Step 2) to separate the signals.

Step 4. The SDR and PESQ of the separated signals are calculated. The elapsed time for separating the signals is considered to be the running time (RT).

Step 5. Steps 1 through 4 are repeated for several experiments with different mixtures and values of p (different SFB sets). Accordingly, the average values of SDR, PESQ, and RT can be plotted versus p, as shown in Figs. 3(a), 3(b), and 3(d), respectively. Also, we incorporate the SDR and PESQ as a single quality criteria (SQC), defined as

$$\mathrm{SQC}(p) = 0.5 \times [|\frac{\mathrm{PESQ}(p)}{PESQ_{\max}}| + |\frac{\mathrm{SDR}(p)}{SDR_{\max}}|], \quad (6)$$

where $SDR_{\max}$ and $PESQ_{\max}$ show the maximum value of the SDRs and PESQs, respectively. Because the SFB set with $p=35\%$ presents the best separation quality, we consider it the bSFB set for $N=3$. We similarly carry out Steps 1 through 5 to find the bSFB sets for $N=2$ and $N=4$, in Table 1.

2. Test Phase

In the test phase, we can practically apply EBS to any mixed

speech signals by considering the bSFB sets found in the training phase and introduced in Table 1, as shown in Fig. 4. Here, the bSFB sets are not recalculated.

Figure 5 shows a comparison of the SDR, PESQ, and RT of the EBS, ICA, FastICA, JADE, and SOBI algorithms for $N=2$, $N=3$, and $N=4$. The corresponding average values of the SDR, PESQ, and RT are shown in Table 2. Notably, EBS is almost 43 times faster than the conventional ICA, while preserving the separation quality. Also, the performance of EBS is much better than FastICA, JADE, and SOBI in terms of quality and speed.

Table 1. bSFB sets for different number of sources.

	N=2	N=3	N=4
Value of p for bSFB sets	20	35	55

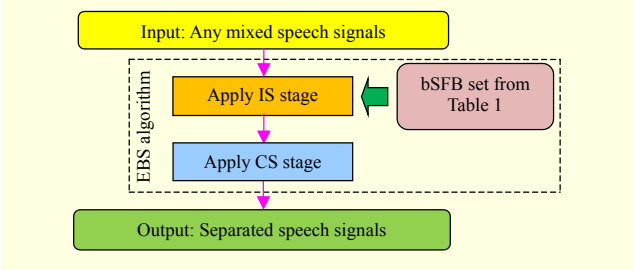

Fig. 4. Applying EBS to any mixed speech signals.

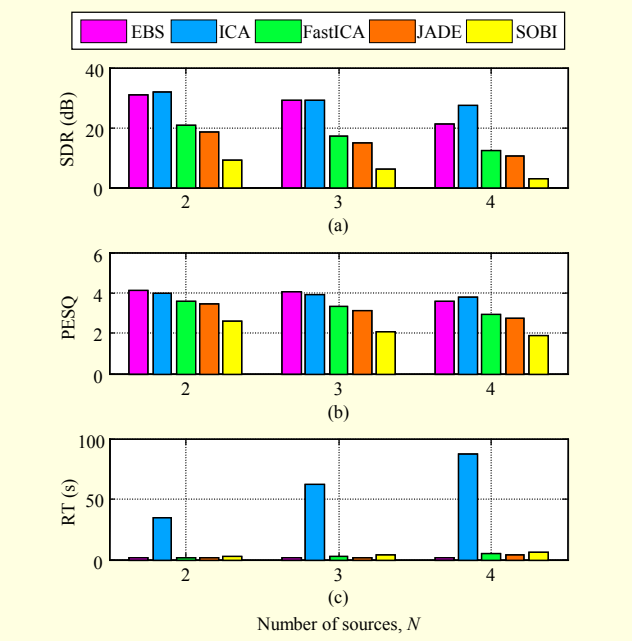

Fig. 5. (a) SDR, (b) PESQ, and (c) RT for EBS, ICA, FastICA, JADE, and SOBI for N=2, N=3, and N=4.

Table 2. Average SDR, PESQ, and RT for different algorithms.

		EBS	ICA	FastICA	JADE	SOBI
Criterion	SDR (dB)	27.04	29.36	16.71	14.52	9.95
	PESQ	3.91	4.03	3.29	3.12	2.19
	RT (s)	1.59	68.37	3.29	2.48	4.41

V. Conclusion

The EBS algorithm was introduced to significantly speed up the separation of mixed speech signals without losing the quality. This was performed by separating mixed signals iteratively in an SFB set and estimating mixing parameters and by incorporating estimated parameters for signal separation via closed-form separation in an RFB set. Also, it was shown that separation of speech signals can be performed in bSFB sets to significantly decrease the number of computations. Using simulation results, it was demonstrated that EBS is approximately 43 times faster than ICA, while the separation quality is preserved. Also, EBS is superior to FastICA, JADE, and SOBI in term of separation quality.

References

[1] M.S. Pedersen, "A Survey of Convolutive Blind Source Separation Methods," *Springer Handbook on Speech Processing Speech Communication.*, J. Benesty, M.M. Sondhi, and Y. Huang, Eds., Berlin: Springer-Verlag, 2007, pp. 1-33.

[2] T.W. Lee, *Independent Component Analysis-Theory and Applications*, Norwell, MA: Kluwer, 1998.

[3] A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, New York: John Wiley & Sons, Inc., 2001.

[4] S. Makino, T.W. Lee, and H. Sawada, *Blind Speech Separation*, Dordrecht, The Netherlands: Springer, July 2007.

[5] M. Mahdikhani and M.H. Kahaei, "Using CSM and VSM Techniques to Speed Up the ICA Algorithm without a Loss of Quality," *Turkish J. Electr. Eng. Comput. Sci.*, 2013, pp. 1930-1943.

[6] M. Mahdikhani and M.H. Kahaei, "Blind Source Separation Using Virtual Sensors," *4th Int. Conf. Dig. Telecommun.*, Colmar, France, July 2009, pp. 107-110.

[7] H. Sawada et al., "A Robust and Precise Method for Solving the Permutation Problem of Frequency-Domain Blind Source Separation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, Sept. 2004, pp. 530-538.

[8] T. Melia and S. Rickard, "Underdetermined Blind Source Separation in Echoic Environments Using DESPRIT," *EURASIP J. Adv. Signal Process.*, May 2006, pp. 1-19.

[9] R. Gribonval et al., "Proposals for Performance Measurement in Source Separation," *4th Int. Symp. ICA BSS*, Nara, Japan, 2003, pp. 763-768.