Blind Speech Separation in the Frequency Domain Using Assignment Problem Approach

Vuong Hoang Nam, Nguyen Quoc Trung, Tran Hoai Linh
Hanoi University of Science and Technology
Email: namvh-fet@mail.hut.edu.vn

Abstract: **In this paper, we propose a method for blind speech separation of convolutive mixtures in the frequency domain. The main difficulty in a frequency approach is the so-called permutation problem. In the proposed method, we use the Assignment Problem (AP) approach to solve permutation ambiguity in the frequency domain. In our work, we apply three different algorithms including Hungarian, Jonker-Volgenant and Bertsekas's Auction to solve the AP to find the optimal solution. Computer simulation experiments are presented to illustrate the proposed method.**

Keywords: Blind Source Separation, FastICA algorithm, Convolutive Mixtures, Assignment Problem.

I. INTRODUCTION

Blind Source Separation (BSS) is a technique which estimates original source signals using only sensor observations, in which the channel transfer function is unknown. If source signals are non-Gaussian and mutually independent, we can apply the technique of Independent Component Analysis (ICA) to solve a BSS problem [1]. The problem of BSS and ICA has received wide attention in many fields such as speech enhancement, data mining, geophysical data processing, biomedical signal analysis and wireless communications. Let us formulate the BSS model of convolutive mixtures. Suppose N orginal sources are blindly mixed and observed at N sensors. In our case and in most ICA applications, the channel transfer function is usually modelled by a causal and FIR filter and we have the relations between the observations and the sources in the time domain as follow:

$$x_i(n)=\sum_{j=1}^{N}\sum_{k=0}^{\infty}h_{ij}(k)s_j(n-k),\forall i=\overline{1,N} \quad (1)$$

where $x_i(n)$ is the observation at the ith sensor, $s_j(n)$ is the jth source, $h_{ij}(k)$ is the channel transfer function between the jth source and the ith sensor. In our model, the noise is assumed to be negligible as well as the sources are stationary.

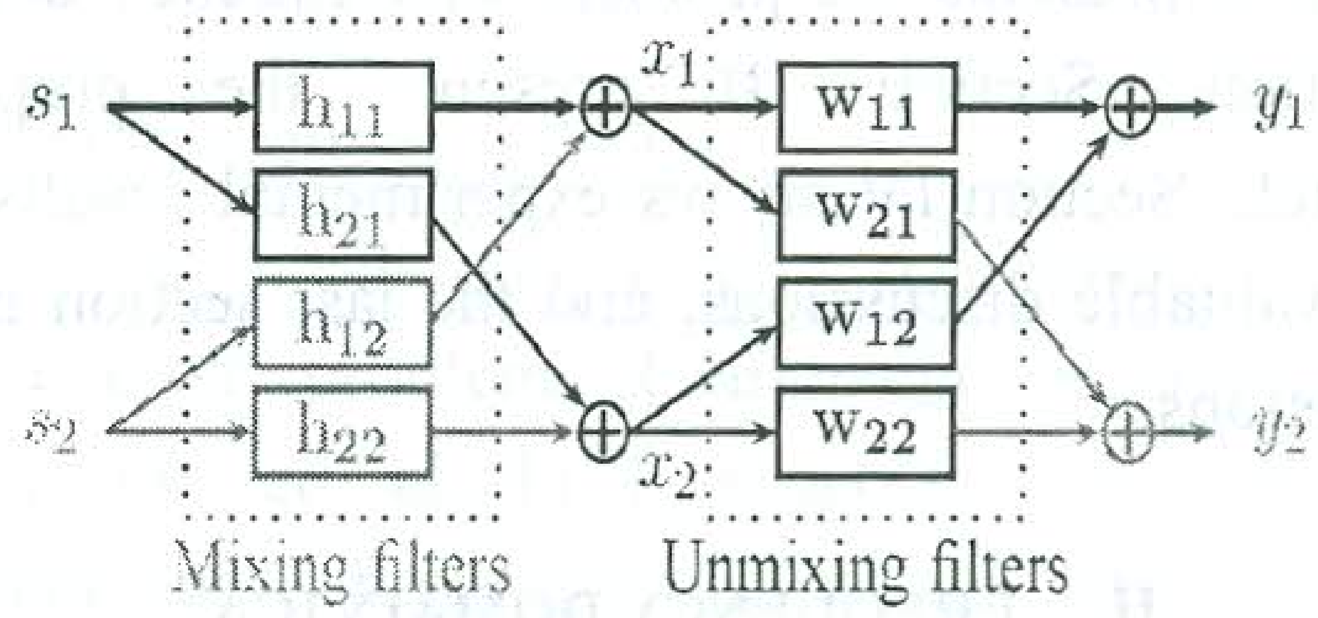

Figure 1. Illutration of TD-ICA with N=2.

There are two major approaches of solving the convolutive mixtures using ICA, which all have some advantages and disadvantages [2]. The first is Time Domain (TD) ICA, which is directly applied to the convolutive mixtures (Fig.1). We can separate the mixtures by estimating a set of unmixing FIR filters as expressed by the following equation:

$$y_i(n)=\sum_{j=1}^{N}\sum_{k}w_{ij}(k)x_j(n-k),\forall i=\overline{1,N} \quad (2)$$

The FIR filter which is used in separation in (2) can be either causal or noncausal depending on the method. In causal case k runs from 0 to $+\infty$, and in non-causal case k runs from $-\infty$ to $+\infty$. This method achieves a good result once upon the algorithm converges, but it is computationally expensive because of dealing with convolution operations.

The second method is Frequency Domain (FD) ICA, where the convolutive mixtures are first converted to frequency domain, then ICA is applied to each frequency bin, which is seen now as instantaneous mixtures, since convolution in time domain is equal to multiplication in frequency domain. This method is quite simple, but the problem of permutation and scaling is a big challenge since different frequency bands may have different permutation and scaling.

In this paper, Assignment Problem (AP) approach solving the permutation ambiguity problem in the frequency domain is exploited. Also, some efficient algorithms have been proposed. The paper is organized as follows. After introduction in Section I, Section II presents the problem of frequency domain separation. Section III presents the proposed approach. Section IV shows experimental results and some valuable discussions, and the last section is the conclusions.

II. FREQUENCY DOMAIN ICA

A. Frequency Domain ICA

Smaragdis [3] has proposed the Short Time Fourier Transform (STFT) method applying for the convolutive mixtures. Using STFT, the time-domain observed signal are transformed into frequency-domain signals. Let $x(n)$ be a digital signal and $x(m,\tau)=\phi(m)x(m+\tau R)$ the windowed and time-shift version of $x(n)$. The STFT $X(\omega_k,\tau)$ is given by:

$$X(\omega_k,\tau)=\sum_{m=0}^{K-1}x(m,\tau)e^{-j\omega_k m} \tag{3}$$

where $\omega_k=\dfrac{2\pi k}{K}$ is the discrete normalized frequency, with bin index $k=0,..,K-1$, frame index $\tau=0,..,L-1$; $\phi(m)$ is a Hanning (in our case) window of length K and R is the frame shifting interval for this transform. Then, the convolutive BSS problem in the time domain is converted to multiple instantaneous problems with complex valued data in frequency bins. For a fixed frequency ω_k, the instantaneous linear mixture can be expressed by the equation:

$$X_i(\omega_k,\tau)=\sum_{j=1}^{N}H_{ij}(\omega_k)S_j(\omega_k,\tau) \tag{4}$$

where $X_i(\omega_k),S_j(\omega_k),H_{ij}(\omega_k)$ are the STFT of $x_i(n),s_j(n),h_{ij}(n)$ at the frequency bin k-th, respectively. Then the BSS model is converted into the frequency domain:

$$X(\omega_k,\tau)=H(\omega_k)S(\omega_k,\tau) \tag{5}$$

where $H(\omega_k)$ is NxN the mixing matrix in the frequency ω_k, $X=[X_1,..,X_N]^T$ and $S=[S_1,..,S_N]^T$ are time-frequency representations of the observed and source signals, respectively. And the estimated signals turned into:

$$Y(\omega_k,\tau)=W(\omega_k)X(\omega_k,\tau) \tag{6}$$

where $Y=[Y_1,..,Y_N]^T$ and $W(\omega_k)$ is the NxN un-mixing matrix in the frequency ω_k.

FD-ICA in this part is summarized in the block diagram shown in Fig.2. The convolved signal is first passed through the Short Time Fourier Transform (STFT), thereby converting it into frequency domain. After this step, we estimate the unmixing matrix for every frequency bin using an instantaneous ICA algorithm in the complex domain (see 2.2) to get the separated signals in frequency bins; we thereafter deal with the indeterminancy of permutation and scaling by the proposed method in the next section. Afterwards is the Inverse Short Time Fourier Transform (ISTFT) which converts the separated frequency bins into separated time domain signals.

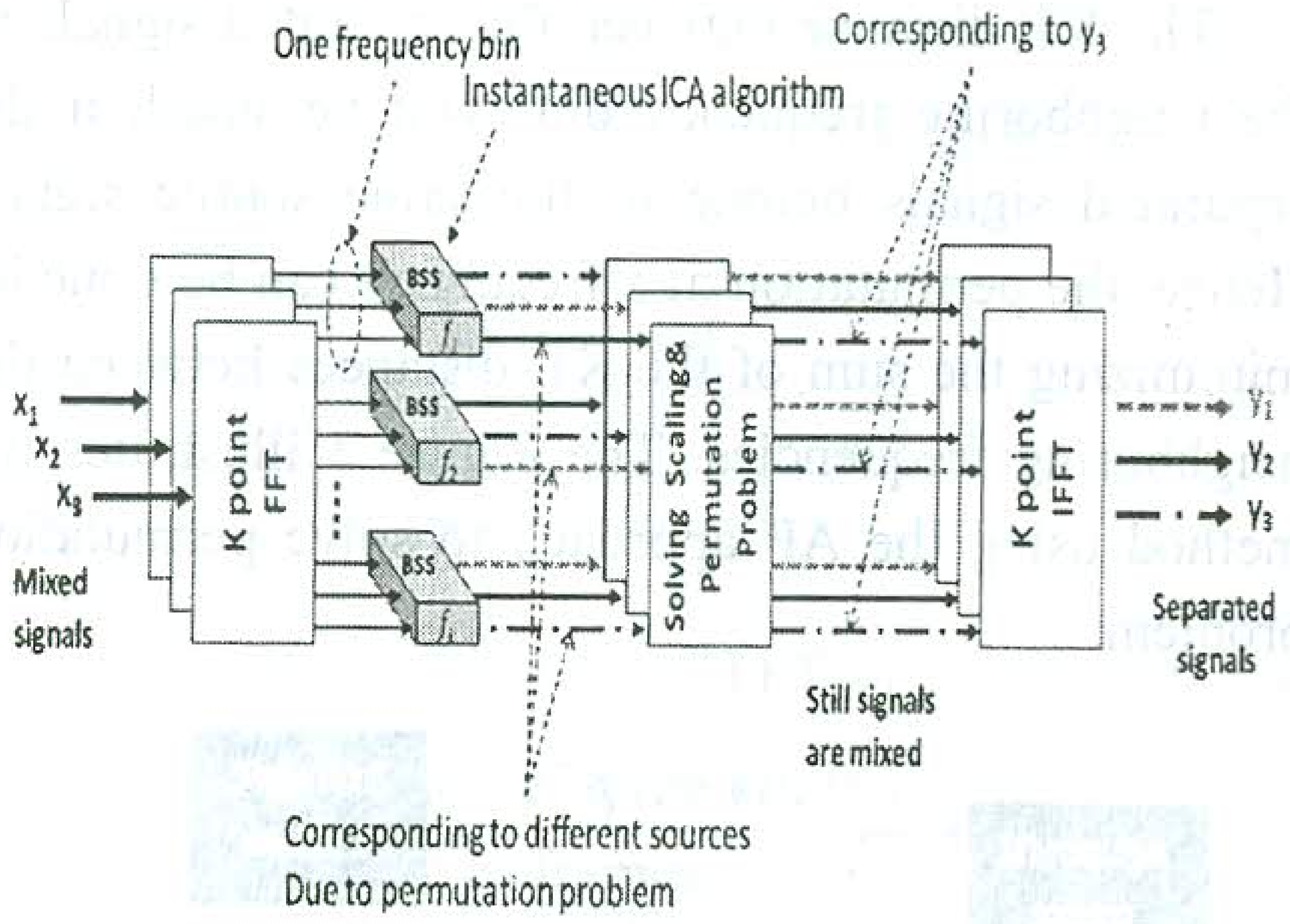

Figure 2. Illutration of FD-ICA.

B. The Complex Fast-ICA algorithm

To estimate the unmixing matrix for every frequency bin, we use the fixed-point algorithm, the complex version of FastICA for complex signals, under the instantaneous ICA model [4]. The algorithm uses a deflation scheme to search the extrema of $E\left\{G\left(\left|w^H x\right|^2\right)\right\}$, where G is a contrast function. Different choices of G have been suggested in [4]. A good contrast function is one for which the estimator given by the contrast function is more robust to outliers in the sample values. In our work, we choose $G(t) = log(0.1+t)$ as the contrast function. The robustness of the estimator is captured in the slow growth of G, as its argument increases [4]. The algorithm requires a preliminary sphering of the data. The fixed-point algorithm for each unit is

$$w^+ = E\left\{x\left(w^H x\right)^* g\left(\left|w^H x\right|^2\right)\right\} - $$

$$- E\left\{g\left(\left|w^H x\right|^2\right) + \left|w^H x\right|^2 g'\left(\left|w^H x\right|^2\right)\right\} w \quad (7)$$

$$w^+ = \frac{w^+}{\left\|w^+\right\|}$$

where g, g' are the derivatives of G and g.

III. THE PROPOSED APPROACH

A. Scaling Problem

We use a simple approach to solve the scaling problem. The scaling invariance means that the scaling of every frequency bin can be different, which will of course result in spectral deformation of the original sources. This problem can be remedied by forcing the determinant of the unmixing matrices to unity by:

$$W_\omega^{norm} = W_\omega^{orig} . \left|W_\omega^{orig}\right|^{-\frac{1}{N}} \quad (8)$$

where W_ω^{norm} and W_ω^{orig} are respectively the normalized and original NxN unmixing matrices at the frequency bin ω, respectively.

B. Permutation Problem

The key technique in FD-ICA is how to solve the permutation problem. Ciaramella et al [5] and Tichavsky et al [12] proposed to apply the Assignment Problem approach for solving permutation problem. To solve AP, they used Hungarian algorithm. We note that since the AP is a special case of the Transportation Problem we could use the same algorithms to solve the AP. In our work, we apply three algorithms including Hungarian algorithm (HA) [6], Bertsekas's Auction (BA) [7] and Jonker-Volgenant algorithm (JV) [8], for optimal assignment. The BA has usually been presented for the maximization version of the AP, but for congruence with our work, we use its application to a minimization problem.

The first two algorithms (HA and BA) are primal-dual methods based on maximum flow. The HA of Kuhn [6] actually served as the algorithm from which the general primal-dual algorithmwas derived. The original method has computational complexity $O\left(n^4\right)$, but later versions $O\left(n^3\right)$ were developed. Bertsekas [7] also presents a primal-dual algorithm. The method is Hungarian-type, and the best version even switches to the Hungarian algorithm itself as soon as the orginal method becomes less effective.

The JV is a shortest paths based algorithm (Dijkstra algorithm). It uses flow augmentation along paths in an auxiliary network, that, depending on the current flow, can be constructed from the original one. By using some reduction procedures, theoretically, the JV can reduces expected running time to $O(n^2.\log n)$. The average computation times have been proven be uniformly lower than the other algorithms (HA,BA) for all tesed instances [8].

We considered that the outputs of two adjacent bins as sources in the source set A and destinations in the destination set B in an AP. The task then is to find a perfect matching (one to one) between A and B such that the sum of the weights of the matching is minimized. The weights correspond to the distance between two elements. Let x_{ij} be a variable which is 1 if a_i in A maps to b_j in B and 0 otherwise and d_{ij} is the distance between two elements. The mathematical model for the AP is given by the following:

$$\text{Minimize} \quad f(A,B)=\sum_{i=1}^{N}\sum_{j=1}^{N} d_{ij}x_{ij} \qquad (9)$$

subject to

$$\sum_{j=1}^{N} x_{ij}=1, i=\overline{1,N}; \ \sum_{i=1}^{N} x_{ij}=1, j=\overline{1,N}; \qquad (10)$$
$$x_{ij}\in\{0,1\} \ \forall i,j$$

In the AP, to evaluate the distance between bins the symmetric Kullback-Leibler (KL) divergence is used [11]. The KL divergence is defined between two discrete M-dimensional probability density function $p=[p_1 \dots p_M]$ and $q=[q_1 \dots q_M]$ as:

$$KL(p,q)=\sum_{i=1}^{M} p_i \log\frac{p_i}{q_i} \qquad (11)$$

The KL distance between the separated signals in the neighboring frequency bins will be small, if the separated signals belong to the same source signal. Hence, the permutation at a frequency can be done by minimizing the sum of the KL distances between the neighboring frequencies. The Figure 3 illustrates the method using the AP approach to solve permutation problem.

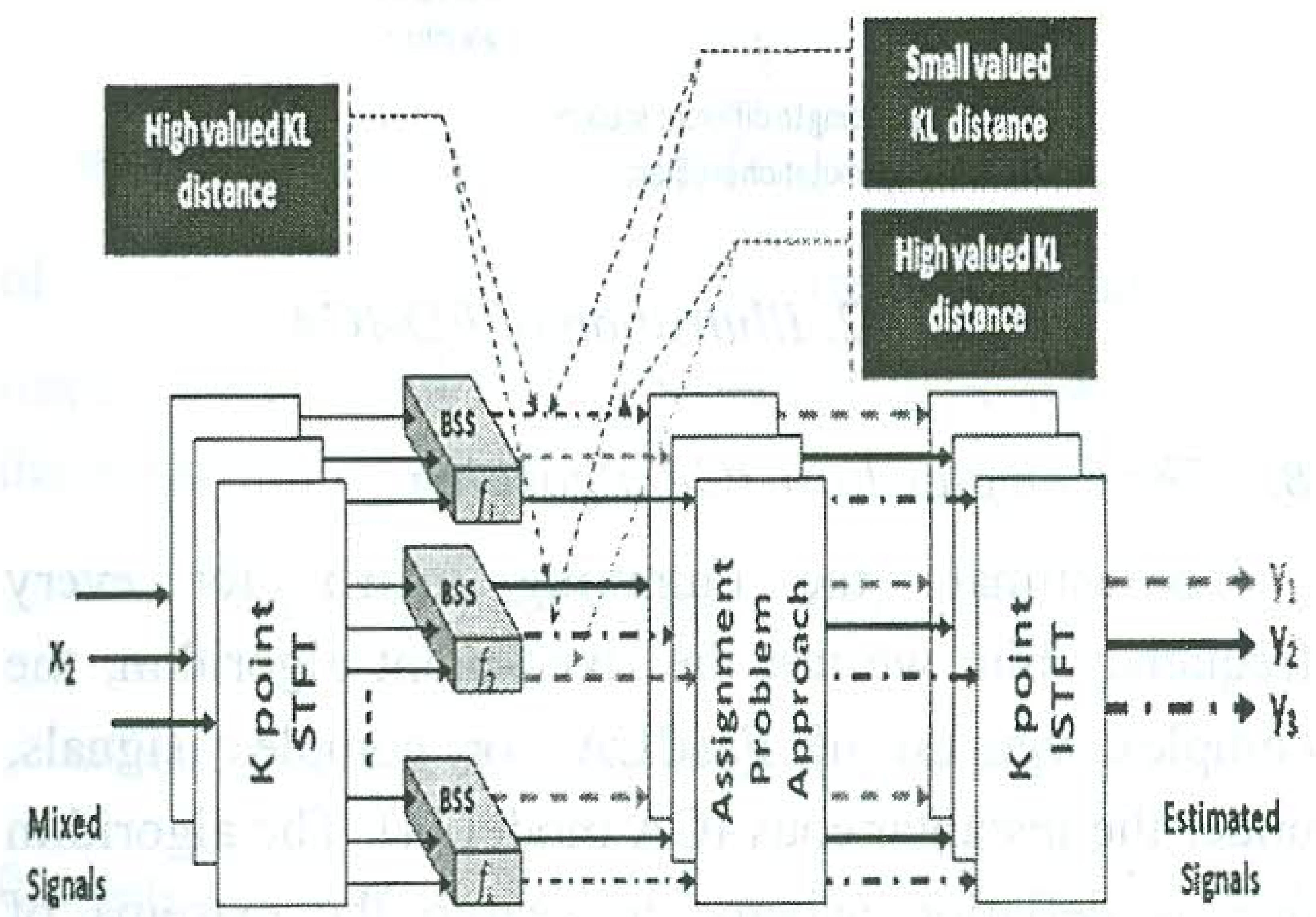

Figure 3. Illustration of the method using the AP approach to solve permutation problem

Figure 4 illustrates the adjacent bands KL divergence method in the case of N=2 and we have to solve the permutation problem in the two adjacent bands (K-1)- and K-th.

In our method, the Cost matrix in the AP be an *NxN* matrix whose elements is defined as

$$KL_{ij}=KL\left(p_{K-1}^{i},p_{K}^{j}\right) \quad \forall i,j=\overline{1,N} \qquad (12)$$

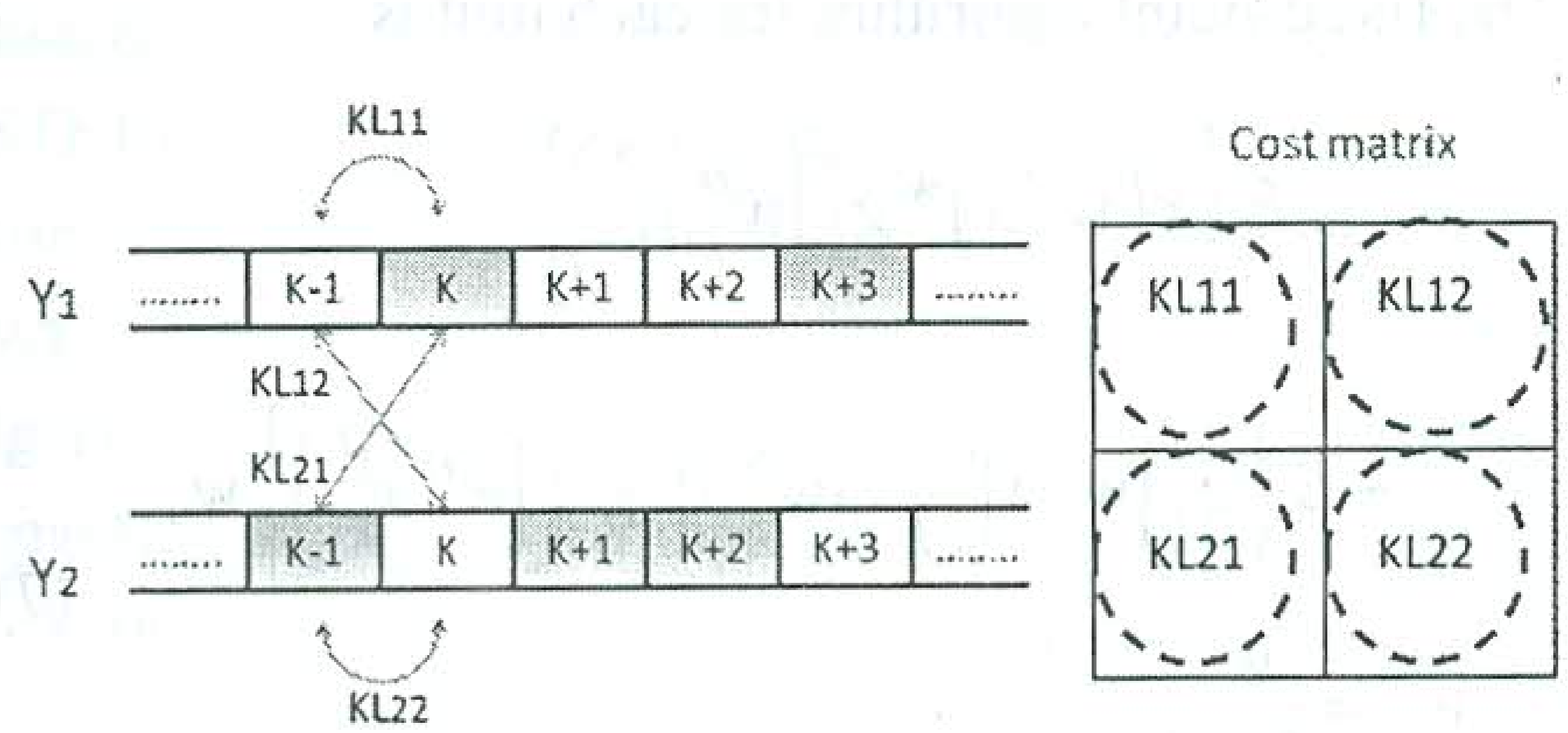

Figure 4. Illutration of the adjacent bands KL divergence method (N=2).

where the l th component of the discrete distribution $p_k^i = \left[p_{k,1}^i \ldots p_{k,M}^i \right]$ was defined as

$$p_{k,l}^i = \frac{\left|X_i(\omega_k, l)\right|^2}{\sum_{s=1}^{M} \left|X_i(\omega_k, s)\right|^2} \tag{13}$$

where $X_i(\omega_k, l)$ is the STFT of the ith extracted source at the lth frame of the kth frequency bin and M is the total number of frames.

IV. EXPERIMENTAL RESULTS

In this section, some computer simulation experiments with artificial and benchmark speech data are presented to illustrate the AP approach. We note that this approach is tested with different AP algorithms. In our experiments, we choose $K = 512, R = 256$ and experiments are performed on a PC with Intel Core 2 Duo CPU @ 2.40 GHz,1GB of RAM.

A. Speech-Speech Separation

In our initial experiment, we use Ikeda benmark data avaiable at [10].We have to separate the sounds of two male speakers recorded with a sampling rate of 16KHz from real convolutive mixtures. The signals are the world "good morning" and a Japanese word "konbaga". In Fig 5.a, we plot the convolved mixtures. In Fig 5.b, Fig 5.c and Fig 5.d, we show the results obtained after separation using HA, BA, JV algorithm, respectively.

B. Speech-Noise Separation

In this experiment, we use Paris Smaragdis benchmark data available at [10]. We want to separate the sound of a male speaker from background noise. The signals are sampled at 44.1KHz. In Fig 6.a, we plot the convolutive mixtures. In Fig 6.b, Fig 6.c and Fig 6.d, we show the results obtained after separation using HA, BA, JV algorithm, respectively.

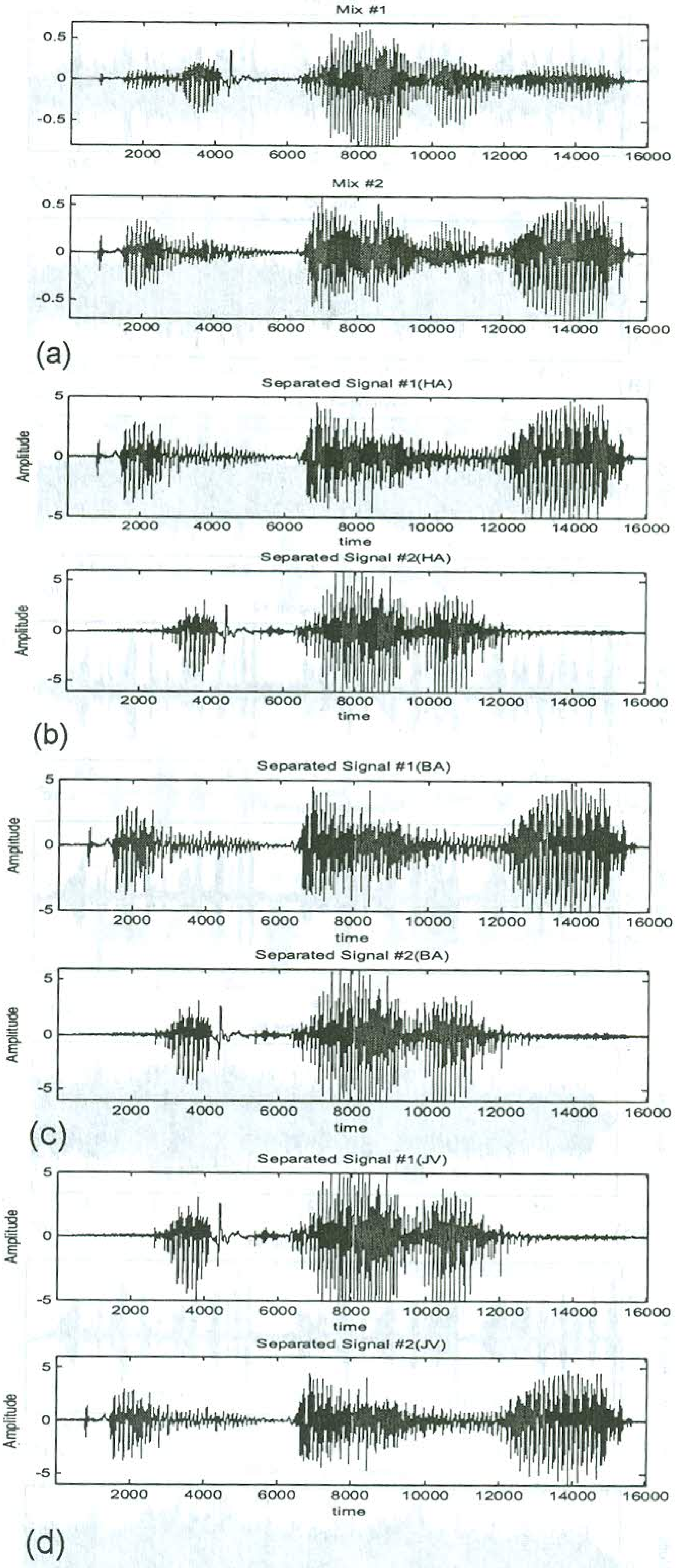

Figure 5. The results obtained from the first experiment.

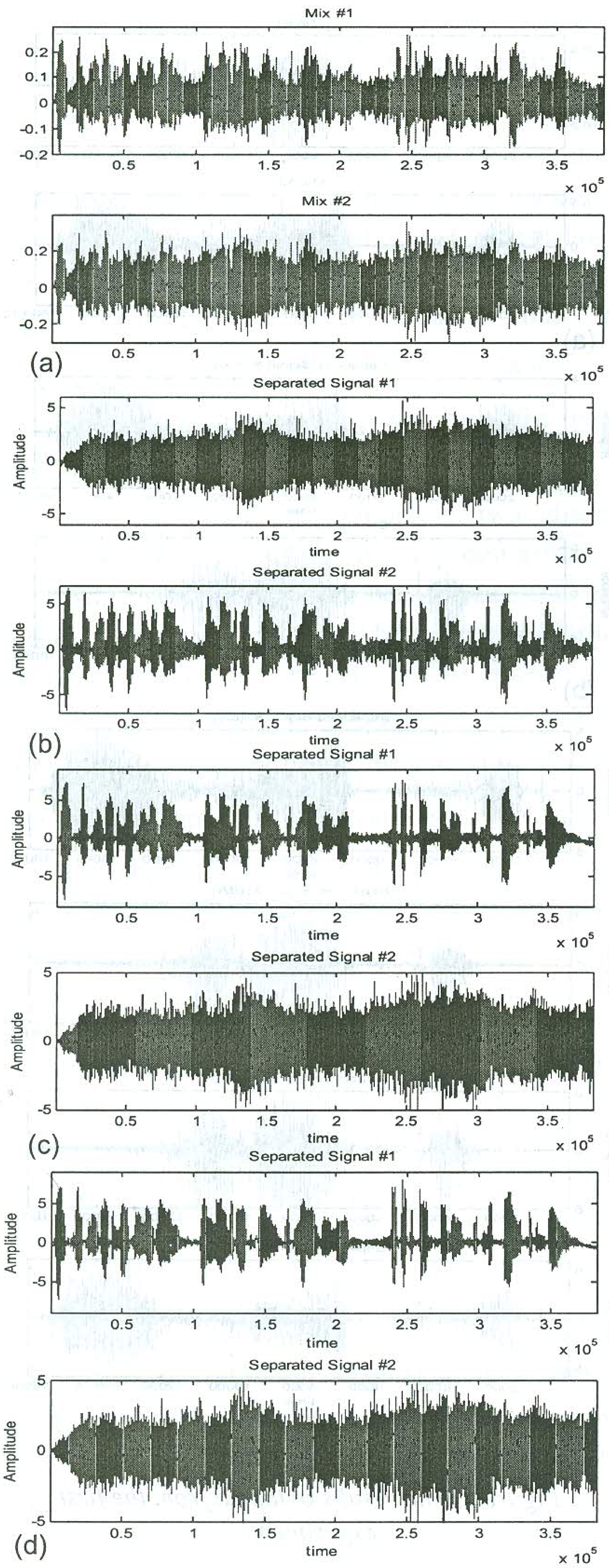

Figure 6. The results obtained in the second experiment

C. Artificial Mixtures Separation

In our last experiment, convolutive mixtures from two Vietnamese speech sources shown in Fig 7.a, and sampled at 16 KHz. We tested our algorithm with real 64th-order mixing filters measured at the ears of a dummy head [9]. We selected the impulse responses associated with source positions defined by 45- and 110-degree angles in relation to the dummy head. The mixing filters are depicted in Fig 7.b. We used these responses to create the mixed signals as follows:

$$
\begin{aligned}
x_1(n) &= h_{11} * s_1(n) + h_{12} * s_2(n) \\
x_2(n) &= h_{21} * s_1(n) + h_{22} * s_2(n)
\end{aligned}
\tag{14}
$$

The two mixed signals are shown in Fig 7.c. In Fig 7.d, Fig 7.e and Fig 7.f, we show the results obtained after separation using HA, BA, JV algorithm, respectively.

To estimate performances of the proposed method, we use the Source-to-Interferences Ratio (SIR) for time-invariant filters allowed distortions [13,14]. In [13], SIR is computed for each estimated source by comparing it to the given true source. In this approach, therefore, the mixing system and demixing technique do not need to be known and the only assumption we make is that the true source signals are known. That is the reason this performance criterion can be applied on all usual Blind Audio Source Separation problems. Table 1 shows the average SIRs of estimated signals in the above simulation experiments.

Table 1. Experimental results of extracting main content with various evaluations

Experiment	HA	BA	JV
Exp. 1	21.20-dB	21.02-dB	21.23-dB
Exp. 2	20.00-dB	21.69-dB	22.63-dB
Exp. 3	23.06-dB	22.86-dB	24.03-dB

From Table.1, we can see that the results of separation in the first experiment are good. All three algorithms obtain very similar results. In the following experiment, we note that the BA and JV algorithms permit to obtain better results than HA algorithm. And

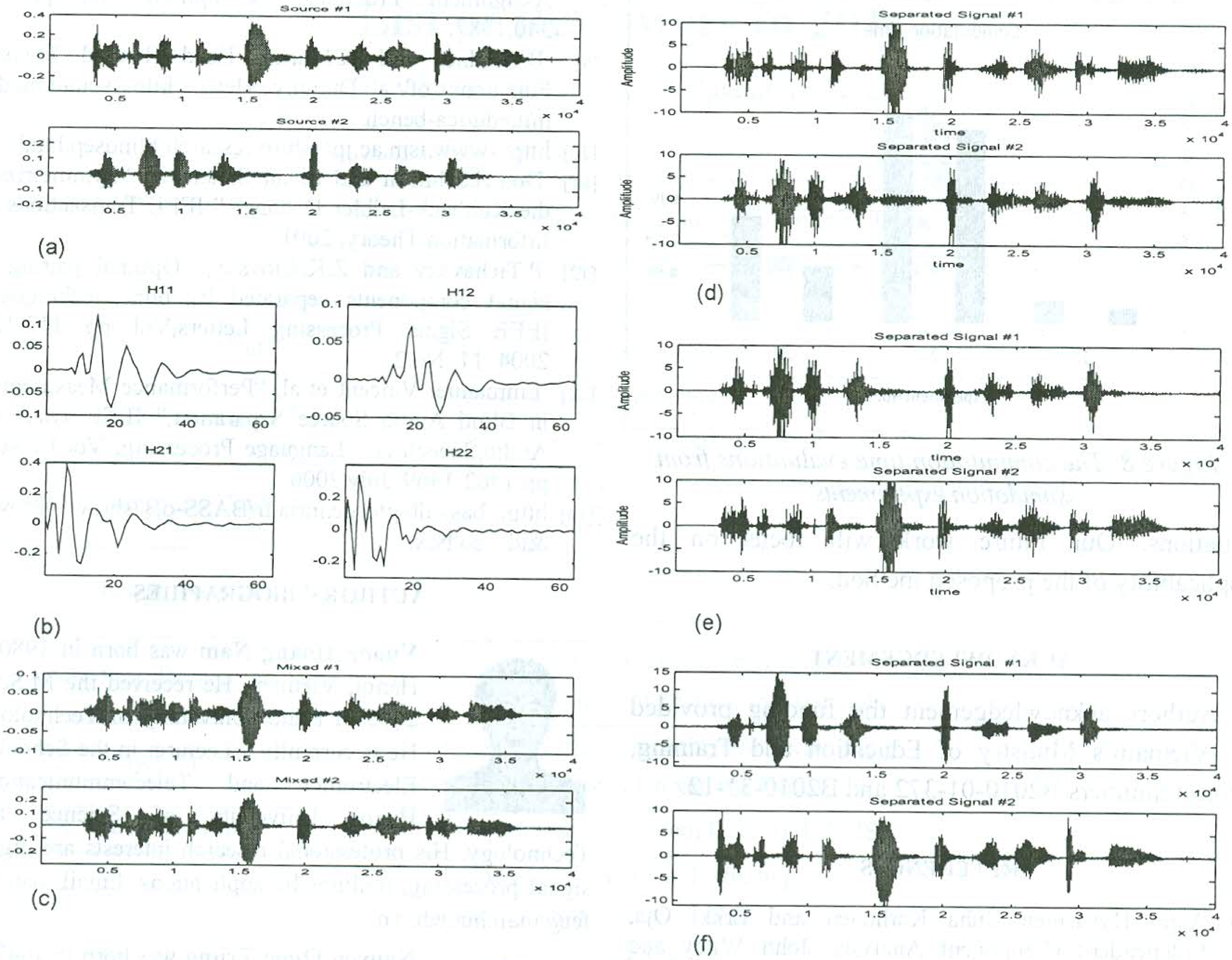

Figure 7. The results obtained from the last experiment

in the last experiment, the result obtained from JV is the best while the remains are similar.

Fig.8 shows the computation time for solving permutation problem in the above simulation experiments.

From the results obtained with the AP approach on artificial and benmark data we can conclude it reach good performance to all the applied AP algorithms. We also see that the JV algorithm presents better performance than the rest. Moreover, this algorithm have been proven to be uniformly faster than the other algorithms (HA,BA). These are the reasons the AP using Jonker-Volgenant algorithm turned out to be the best choice in our experiments.

V. CONCLUSIONS

In this paper, we have discussed several aspects of the general audio source separation problem and proposed some solutions on the permutation problem. We have presented a FD-ICA, which is the most successful approach up to now, using the Assigment Problem approach. In our work, by using Jonker-Volgenant algorithm, we can reduce the computation time compare to previous results [5]. From the different experimental results we can conclude that the proposed method presents good separation performance as well as indicates a good potential to be implemented into BSS application in real world