4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), April 2003, Nara, Japan

BLIND SOURCE SEPARATION USING ICA AND BEAMFORMING

Xuebin Hu and Hidefumi Kobatake
Tokyo University of Agriculture & Technology.
2-24-16 Naka-cho, Koganei-shi, Tokyo 184-8588, Japan

ABSTRACT

Time-frequency domain blind source separation (BSS) leads to an important problem that generally the independence assumption between source signals collapses in frequency domain due to inadequate samples. It consequently degrades the performance of all the ICA-based BSS methods. To remedy the defect, we propose introducing the beamforming into the conventional BSS system taking the advantage that the null beamforming does not depend upon the assumption of independence but only upon the estimation of the directions of arrival (DOA). We set up a criterion on the performance of separation. It is used to compare the separation results by ICA and beamforming, and select the result that is thought better. The separations at certain bins are greatly improved, which results in a better separation.

1. INTRODUCTION

Blind source separation has received extensive attention in signal and speech processing, machine intelligence, and neuroscience communities. The goal of BSS is to recover the unobserved source signals without any prior information given only the sensor observations that are unknown linear mixtures of the independent sources. A variety of successful ICA methods have been developed for this purpose [1-5].

Due to the multi-path effect and reverberation in real environment, computationally blind speech separation is often implemented in time-frequency domain. A number of approaches for the delayed and convoluted sources separation have been reported [6-9]. The commonly mentioned disadvantage in frequency domain implementation is that the standard ICA indeterminacy of scaling and permutation appears at each output frequency bin. Various methods using different continuity criteria as listed in [10] have been reported.

There is another disadvantage, which is not addressed yet but is thought to be important. It is the assumption of independence between source signals collapsed in frequency domain [11]. This is because the BSS is normally implemented on a short time period of observations due to the dynamic mixing process in real environment. Frequency domain implementation leads to much less samples than in time domain. As a result, the estimation of statistics often includes large error. For example, the correlation function between source signals can no longer be expected to be zeros. We say that the frequency representations of the source signals corresponding to the observations of limited samples are *correlated*. The existing correlation obviously disobeys the commonly adopted basic assumption of ICA that the sources should be independent with each other. It decreases the performance of ICA in different degree, sometimes very seriously, at various frequency bins whatever the separation method is used. Furthermore, it degrades the performance of the solution to the ambiguity of scaling and permutation problem.

For remedying the unfavorable effect caused by the low-independence problem, we propose to incorporate the beamforming into conventional ICA-based BSS method. It is expected that beamforming could produce a better separation to replace ICA when ICA can not do its job properly due to low-independence problem. The null beamforming is established on the estimated directions of arrival (DOA) of sources and has the advantage that it does not depend on the assumption of independence between source signals. It could give a better separation if a good estimation of DOA is available. We first set up a criterion on the performance of separation and use it to compare and select the separation that is thought better between those achieved by ICA and beamforming. Because the DOA at each frequency is generally different in acoustic environment and also in consideration of the estimation error of it, we adopt a search-and-error scheme in using the beamforming.

The rest of this paper is organized as follows. Section 2 summarizes the conventional ICA-based method and the directivity pattern. Section 3 details the low-independence problem in frequency domain BSS. The section 4 describes the proposed criterion on the performance of separation and section 5 gives the proposed method. In section 6, more clarification of DOA and beamforming are made. Section 7 gives the simulation test results, and following the discussion on the results of the experiments, the paper is concluded.

2. PREPARATION

2.1. Conventional ICA-based BSS

The formation of conventional ICA-based BSS could be summarized as follows. Source signals are assumed to be independent with each other, with zero mean, and are denoted by a vector $\boldsymbol{s}(t) = (s_1(t), \cdots, s_L(t))^T$. When the signals are recorded in a real environment, the observations can be approximated with convolutive mixtures of source signals,

$$\boldsymbol{x}(t) = \boldsymbol{A} * \boldsymbol{s}(t) = \left(\sum_l a_{lk} * s_l(t) \right), \tag{1}$$

where $\boldsymbol{A}$ is an unknown polynomial matrix, a_{lk} is the impulse response from source l to microphone k, and the symbol * refers to convolution. In frequency domain, the convolutive mixing problem is decomposed into multiple instantaneous mixing problems.

$$\boldsymbol{X}(f,t) = \boldsymbol{A}(f)\boldsymbol{S}(f,t). \tag{2}$$

The instantaneous mixing problem then can be solved using any desired ICA method. With the derived unmixing filter $\boldsymbol{W}(f)$, we recover the source signals by

$$\hat{\boldsymbol{S}}(f,t) = \boldsymbol{PDW}(f)\boldsymbol{X}(f,t). \tag{3}$$

where, $\boldsymbol{P}$ and $\boldsymbol{D}$ are the solution to the ambiguity of permutation and scaling. Then we can either transfer the bin unmixing filters into time domain or directly transfer the separated frequency components into time domain to recover the source signals.

In the conventional BSS method, the unmixing matrix $W(f)$ is derived by the following learning rule [3]:

$$\begin{aligned} W_{i+1}(f) = W_i(f) + \eta[& diag(\langle \Phi(Y(f,t))Y^H(f,t) \rangle_t) \\ & - \langle \Phi(Y(f,t))Y^H(f,t) \rangle_t] W_i(f) \end{aligned} \tag{4}$$

where $\langle \cdot \rangle_t$ denotes the time averaged operator, η denotes the step-size learning factor, and the nonlinear vector function $\Phi(\cdot)$ is defined as [6]:

$$\begin{aligned} \Phi(Y(f,t)) &= [\Phi(Y_1(f,t)), \cdots, \Phi(Y_L(f,t))]^T \\ \Phi(Y_l(f,t)) &= [1 + \exp(-Y_l^{(R)}(f,t))]^{-1} \\ &\quad + j \cdot [1 + \exp(-Y_l^{(I)}(f,t))]^{-1} \end{aligned} \tag{5}$$

where $Y_l^{(R)}(f,t)$ and $Y_l^{(I)}(f,t)$ are the real and imaginary parts of $Y_l(f,t)$, respectively.

2.2. Directivity pattern [6]

For utilizing the directivity pattern to estimate the DOA, a linear array is assumed. The coordinates of the microphones are designated as $d_k (k = 1, \cdots, K)$, and the directions of arrival of multiple sound sources are designated as $\theta_l (l = 1, \cdots, L)$. Hereinafter, we assume a two channel system, i.e., K=L=2. Fig.1 illustrates the configuration of the linear spaced array and sound sources.

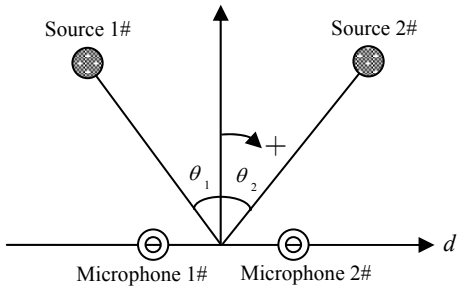

Fig.1 Illustration of the configuration

In the standpoint of the array signal processing, and omitting the permutation and scaling operators, $\boldsymbol{P}$ and $\boldsymbol{D}$, the resultant output signals of Eq.(3) could be seen as obtained by multiplying array signals X_1(t) and X_2(t) by the weight W_{kl} and adding them. It implies that directivity patterns are produced in the array system. The unmixing matrix W_f that is achieved by ICA could be used to estimate the DOAs of sound sources with respect to each frequency bin. The directivity of the array system, $F_l(f,\theta)$, is given by [Johnson and Dudgeon, 1993]

$$F_l(f,\theta) = \sum_{k=1}^{K} W_{lk}^{(ICA)}(f) \exp[j2\pi f d_k \sin\theta / c], \tag{6}$$

where $W_{lk}^{(ICA)}(f)$ is the element of $W_{i+1}(f)$. In the directivity patterns, directional nulls exist in only two particular directions. The DOA of the l-th source, $\hat{\theta}_l$, is estimated as

$$\hat{\theta}_l = \frac{2}{N} \sum_{m=1}^{N/2} \theta_l(f_m), \tag{7}$$

where N is the length of DFT, $\theta_l(f_m)$ denotes the DOA of the l-th source at the m-th frequency bin, which are given by

$$\begin{aligned} \theta_1(f_m) &= \min[\arg\min_\theta |F_1(f_m,\theta)|, \arg\min_\theta |F_2(f_m,\theta)|] \\ \theta_2(f_m) &= \max[\arg\min_\theta |F_1(f_m,\theta)|, \arg\min_\theta |F_2(f_m,\theta)|] \end{aligned} \tag{8}$$

2.3 Beamforming

We use the null beamforming [6] in our method. In the case that the look direction is $\hat{\theta}_1$ and the directional null is steered to $\hat{\theta}_2$, the elements of the matrix for signal separation are given as

$$\begin{aligned} W_{11}^{(BF)} = & \exp[-j2\pi f_m d_1 \sin\hat{\theta}_1 / c] \\ & \times \{\exp[j2\pi f_m d_1 (\sin\hat{\theta}_2 - \sin\hat{\theta}_1)/c] \\ & - \exp[j2\pi f_m d_2 (\sin\hat{\theta}_2 - \sin\hat{\theta}_1)/c]\}^{-1} \end{aligned} \tag{9}$$

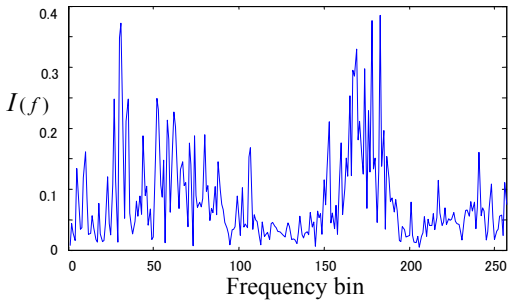

Fig.2 Correlation distribution between two speech samples, "Good morning" and "Konbanwa".

$$W_{12}^{(BF)} = -\exp[-j2\pi f_m d_2 \sin\hat{\theta}_1 / c] \times\{\exp[j2\pi f_m d_1(\sin\hat{\theta}_2 - \sin\hat{\theta}_1)/c] - \exp[j2\pi f_m d_2(\sin\hat{\theta}_2 - \sin\hat{\theta}_1)/c]\}^{-1} \quad (10)$$

And in the case that the look direction is $\hat{\theta}_2$ and the directional null is steered to $\hat{\theta}_1$, the elements of the matrix for signal separation are given as

$$W_{21}^{(BF)} = -\exp[-j2\pi f_m d_1 \sin\hat{\theta}_2 / c] \times\{-\exp[j2\pi f_m d_1(\sin\hat{\theta}_1 - \sin\hat{\theta}_2)/c] + \exp[j2\pi f_m d_2(\sin\hat{\theta}_1 - \sin\hat{\theta}_2)/c]\}^{-1} \quad (11)$$

$$W_{22}^{(BF)} = \exp[-j2\pi f_m d_2 \sin\hat{\theta}_2 / c] \times\{-\exp[j2\pi f_m d_1(\sin\hat{\theta}_1 - \sin\hat{\theta}_2)/c] + \exp[j2\pi f_m d_2(\sin\hat{\theta}_1 - \sin\hat{\theta}_2)/c]\}^{-1} \quad (12)$$

3. THE PROBLEM

Due to the dynamic mixing process in real environment, BSS is normally implemented on a short time period of observations. Frequency domain implementation leads to much less samples than that in time domain. As a result, there often exists large estimation error in the second and higher order statistics. For example, the correlation function between the source signals can no longer be expected to be zeros. The frequency components of source signals, corresponding to the observations of the limited samples, are *correlated*.

For evaluation of the correlation between $s_1(f,t),\cdots,s_L(f,t)$, we define that,

$$\boldsymbol{V}(f) = diag\left(\left\langle S(f,t)S^{H}(f,t)\right\rangle\right) - \left\langle S(f,t)S^{H}(f,t)\right\rangle, \quad (13)$$

where $\langle\cdot\rangle$ denotes the expectation operator, and $\left\langle S(f,t)S^{H}(f,t)\right\rangle$ is the normalized covariance matrix. The correlation between the frequency components of source signals, denoted as $I(f)$, is quantified by the Frobenius norm of $\boldsymbol{V}(f)$. It is further normalized and defined as,

$$I(f) = \|\boldsymbol{V}(f)\| = \sqrt{\sum_{l\neq k}\left|\boldsymbol{V}_{lk}(f)\right|^2 \Big/ (N^2 - N)} \quad (14)$$

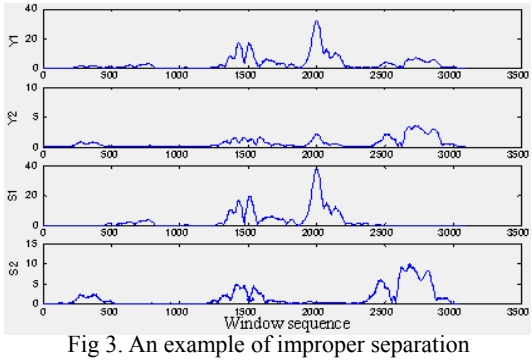

Fig 3. An example of improper separation

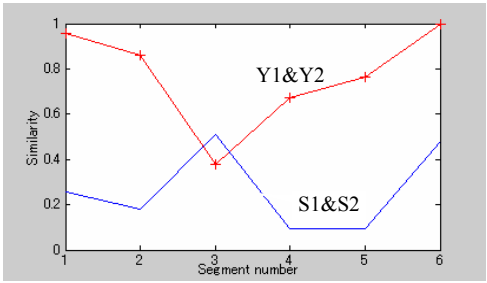

Fig 4. Comparison of segmental similarities between improperly separated signals and original signals.

The higher the value $I(f)$ is, the lower the independence will be. Fig. 2 shows the correlation distribution against frequency between the two speech samples, "Good morning" and "Konbanwa". The signals are about one second in length, and the sampling rate is 16kHz. The DFT length is 512 samples (32msec) and the window shift is 5 samples. Hamming window is used. It is noted that the correlation may vary in some degree when using different frame length and window shift. But it will not cause significant change to it.

Regardless of the ICA method, the off-diagonal elements of the covariance matrix of the separated signals are to be minimized as practical, ideally minimized to zeros. In other words, ICA will make $I(f)$ close to or equal to zero. The existing correlation between the sources as shown in Fig.2 apparently shows that ICA will not work perfectly in such case. It degrades the performance of ICA, sometimes makes the separation completely failed.

4. CRITERION

ICA assumes the limited data of $\boldsymbol{S}(f,t)$ is independent and separates the frequency mixtures into mutually independent signals. However $\boldsymbol{S}(f,t)$ is often not independent and sometimes the separation completely failed. In such case, the separated signals are still mixed with each other. Taking an improper separation (Fig.3) as an example, Y1 and Y2 are the separated results achieved

by ICA, and S1, S2 are the components of original sources. The waveforms are the norms of the complex-valued signals respectively. Although Y1 and Y2 are uncorrelated with each other, comparing with S1 and S2, it is obvious that Y1 and Y2 are quite similar at certain segment pairs, for example at [1900, 2200] and [2400, 3000].

The criterion is set up on the above observation. We divide the separated signals $\hat{\boldsymbol{S}}(f,t)$ into a number of continuous segments $\hat{\boldsymbol{S}}(f,l)$, and use the averaged segmental similarity (ASS) as the criterion to make the decision. The ASS is defined as,

$$ASS_f = \frac{1}{L}\sum_{l=1}^{L}\frac{\left\|\hat{S}_1(f,l)\cdot\hat{S}_2(f,l)^{\mathrm{H}}\right\|}{\left\|\hat{S}_1(f,l)\right\|\left\|\hat{S}_2(f,l)\right\|}. \qquad (15)$$

$$\hat{S}(f,l) = \hat{S}(f,(l-1)\times m+1:l\times m),$$

where, l is the segment sequence number and m is the segmental length.

Figure 4 shows the segmental similarities between Y1, Y2 and S1, S2, respectively. They are divided into six segment pairs with m equals to 500, respectively. We can see that the average of the segmental similarities of improper separation (Y1, Y2) is higher than that of complete separation (S1, S2). In other words, those with higher averaged segmental similarity tend to be improper separation.

In this paper, the proposed ASS is used as criterion to compare and select a better separation among the separations achieved by ICA and beamforming.

5. SEPARATION METHOD

We propose a separation method consisting of ICA and beamforming and expect beamforming produces better separation when the low-independence problem prevents ICA from a proper separation.

Fig. 5 depicts the processing flow of this method. The mixtures are firstly decomposed into frequency domain. In each frequency bin, the frequency mixtures are separated using ICA at first. The DOAs of sources are estimated using the directivity pattern. With the estimated DOAs, we construct the null beamforming.

Due to the reflections and reverberations in real environment, DOAs of sources actually are different at various frequency bins (see section 6.1). Furthermore, the derivation of DOAs of sources using Eq.(7) only provides an approximate estimation. As such, we use a search-and-error scheme to find the most adequate separation by beamforming at each frequency bin. The search range is set to $\hat{\theta}_k \pm \varphi (k = 1,2)$. The setting of the parameter φ depends on the acoustical environment. When the reverberation time is long, it might be better to

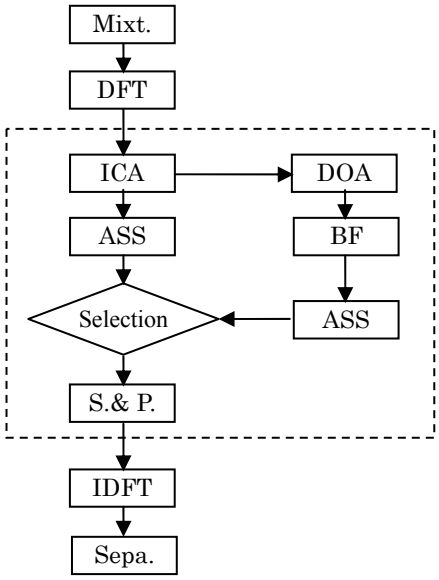

Fig.5 Process flow. BF: beamforming. S.&P.: scaling and permutation. Sepa.: separation. The operations within the broken line are implemented in each frequency bin.

set φ to a big value such as 30-50 degree. And when the reverberation time is short, it might be set to 10-20 degree. Experimentally, it is adequate to set the step size of search to 1-3 degree (see section 6.2). It is pointed out that the scheme does not cause heavy computation load and it may be implemented in parallel.

Among these searches, the separation that outputs the least ASS is selected. It is further compared with the result achieved by ICA as follows,

$$W_f = \begin{cases} W_f^{(ICA)} & ASS_f^{(ICA)} < ASS_f^{(BF)} \\ W_f^{(BF)} & ASS_f^{(ICA)} > ASS_f^{(BF)} \end{cases} \qquad (16)$$

The selected unmixing matrix W_f are transferred into time domain unmixing filter through inverse DFT after solving the indeterminacy of permutation and scaling. For the problem of irregular amplitude of the separated signals, the commonly adopted solution that is putting back the separated frequency components to the sensor with the inverse matrices is used. The indeterminacy of permutation is solved using the DOA information obtained in Eq.(8) [6].

6. DOA AND BEAMFORMING

6.1. About DOA

In real environment, the DOA of a sound source is generally varying at different frequency. The DOA at each frequency is determined by the impulse responses.

In a two microphone array system, the DOA of a source at a certain frequency (θ_f) is determined by the time delay between the two microphones.

$$H_2(\omega) = H_1(\omega)e^{-j\omega d \sin\theta_f / c} \quad (17)$$

where $H_1(\omega)$ and $H_2(\omega)$ denote the transfer functions from the source to the two microphones, respectively. The angle θ_f denotes the arriving angle of the source, d is space between the two microphones and c is the propagation speed of sound wave. From Eq.(17), we get

$$\theta_f = \sin^{-1}(phase(H_1(\omega)/H_2(\omega)) \times c/\omega d) \quad (18)$$

where *phase* denotes the operation that output the phase of a complex number.

Here, as an example, we verify the DOA of a source at each frequency using a pair of impulse responses from RWCP Sound Scene Database. The reverberation time is 300 msec. Source is in the angle of -20 degree. Fig.6 depicts the two impulse responses from the source to the two microphones. The space between the two microphones is 2.83 cm and sound speed c equals to 340 m/s. We use Eq.(18) to evaluate the DOA at each frequency. Fig.7 shows the result. Although DOA varies dramatically at some bins, most of them stay within a certain range with its center is close to the named angle. As such, for the impulse responses of 300 ms, φ equals to 15-25 degree could provide potential abilities to remedy the separations by ICA for most of the bins.

6.2. Step size of the search

For determining the step size of search approximately, a simple simulation using only the beamforming was conducted. A pair of source signals was mixed without reflection. The signal-to-noise ratio were calculated at different pairs of θ_1 and θ_2. Fig.8 shows the result. When both of θ_1 and θ_2 are a little apart away from their real angles in 1-3 degree, respectively, beamforming still provide good separation, the SNR is close to that of exact angles. This indicates that the step size in the search scheme might be set to about 1-3 degree.

7. SIMULATION TESTS

7.1. Simulation 1: Effect of the method

For evaluating the performance of the proposed method solely in remedying the defect from the independence problem in ICA, simulation test 1 was conducted in the simplest condition. The configuration was as follows (see Fig.1): The two sound sources were in the directions of -30 and +30 degrees, respectively. The distances from the sources to the center of array were 2 m. The space

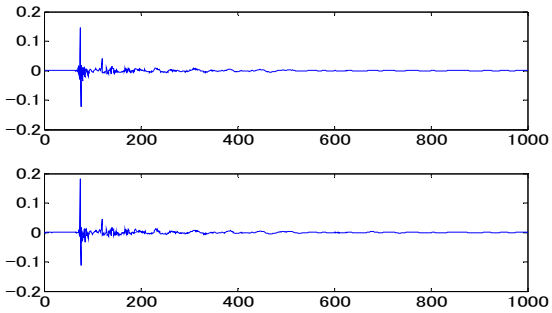

Fig.6 The two impulse responses from a source in the direction of -20 degree to the two microphones.

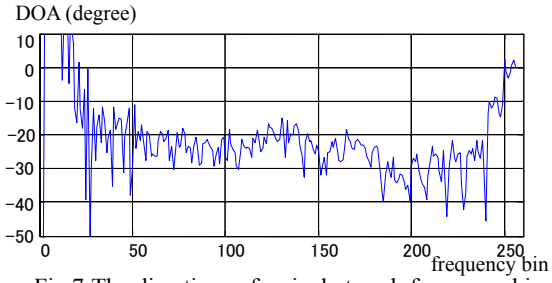

Fig.7 The directions of arrival at each frequency bin of a source in the named angle of -20 degree.

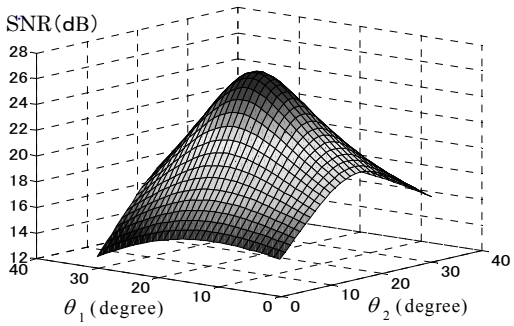

Fig.8 The distribution of SNR verse θ_1 and θ_2.

between the microphones was 8 cm. We used the sound samples from ASJ continuous speech corpus for research. The mixing filter was simulated according to the configuration without echoes.

The test parameters were as follows: sampling rate was 10 kHz, window length was 512 samples, window shift was 20 samples, and Hamming window was used. The signals used for learning the filter were 2.5 sec. The search range parameter φ was set to 2 degree.

Fig.9 shows an example of the improvement of SNR at each frequency bin accompanying by the decrease of ASS comparing to the ICA. At about half of the bins, beamforming provided lower ASS and replaced ICA. Although SNR unexpectedly decreased a little at a few bins, most of the replacement gave a better separation. This proves the effectiveness of the proposed criterion.

Twelve pair of different sound samples were used in

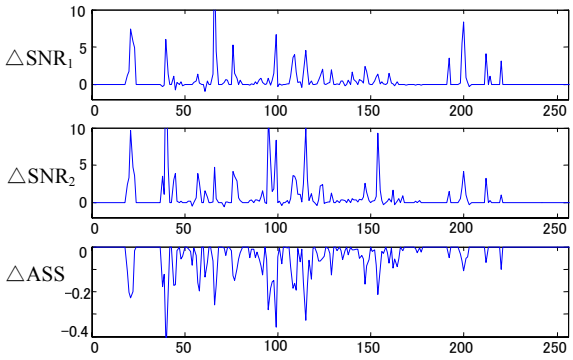

Fig.9 The upper two figures show the improvement of SNR in each frequency bin at source 1 and 2, respectively. The lower figure shows the decreasing of ASS. These are all compared with the conventional ICA. $\triangle SNR=SNR_{New}-SNR_{ICA}$. $\triangle ASS=ASS_{New}-ASS_{ICA}$;

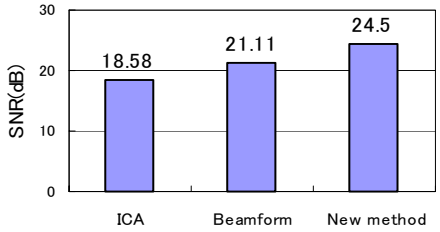

Fig.10 Separation results of simulation 1.

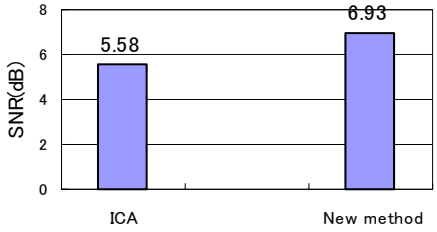

Fig.11 Separation results of simulation 2.

the test. Fig.10 shows the averaged results achieved by the proposed method and those of ICA and beamforming, respectively. The proposed method gave about 6.0 dB improvement than ICA and 3.4 dB higher than beamforming. In the simplest test condition, if bypassing the errors in estimation of DOA, beamforming could provide perfect result.

7.2. Simulation 2: Convolutive mixtures

In simulation 2, we used the same configuration but the impulse responses were used from RWCP Sound Scene Database in Real Acoustic Environment. The reverberation time is 300 msec. Test parameters were same with those used in simulation 1 with the exception that the window length was 1024, the search parameter φ and step were set to 15 and 2 degree, respectively.

Twelve trails on different sound samples were conducted. Fig.11 shows the averaged SNRs by ICA and the proposed method, respectively. Comparing with the conventional ICA-based BSS, about 1.5 dB improvement was achieved.

8. CONCLUSION

This paper addressed the low-independence problem in the frequency domain BSS and proposed a method that using beamforming to replace ICA when ICA cannot work well due to the problem. The null beamforming could work well in simplest condition, but it fails in complicated environment because of variation of DOA at frequency bins. However using it in a search and error scheme, it is possible in remedying the defect from the low-independence problem, which results in a better separation. Simulation tests proved the effectiveness.

REFERENCE

[1] A. Bell, and T. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation*, 7: 1129-1159, 1995.

[2] A. Hyvarinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation*, 9:1483-1492, 1997.

[3] S. Amari, A. Cichocki, and H. H. Yang, "A new learning algorithm for blind signal separation," In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, "*Advances in Neural Information Processing Systems*, vol. 8, pp. 757-763, MIT press, 1996.

[4] A. Belouchrani, K. Abed-Meraim, J. -F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. Signal Processing*, vol. 45(2), pp. 434-443, February 1997.

[5] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals", *IEE Proceedings F,* vol. 140, No. 6, pp. 362-370, 1993.

[6] H. Saruwatari, T. Kawamura, and K. Shikano, "Blind source separation based on fast-convergence algorithm using ICA and array signal processing," In *Proceeding of 3rd International Conference on ICA and BSS (San Diego),* pp. 412-417, Dec. 2001.

[7] S. Ikeda, N. Murata, "A method of blind separation based on temporal structure of signals," In *Proceedings of The Fifth International Conference on Neural Information Processing (ICONIP'98 Kitakyushu)*, pp. 737-742, 1998.

[8] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, 22:21-34, 1998.

[9] S. Kurita, H. Saruwatari, S. Kajita, K. Takeda, and F. itakura, "Evaluation of blind signal separation using directivity pattern under reverberant conditions," In *Proc. ICASSP 2000,* volume V, pp. 3140-3143, 2000.

[10] Kari Torkkola, "Blind separation for audio signals – are we there yet?" *Proc. Workshop on Independent Analysis and Blind Signal Separation,* Jan 11-16, 1999, France.

[11] S. Araki, S. Makino, R. Mukai, T. Nishikawa and H. Saruwatari, "Fundamental limitation of frequency domain blind source separation for convolved mixture of speech," Proc. of ICA2001, pp.132-137, Dec. 2001.