REAL-TIME BINAURAL BLIND SOURCE SEPARATION

Changkyu Choi

Human & Computer Interaction Laboratory
Samsung Advanced Institute of Technology
San 14-1, Nongseo-ri, Kiheung-eup, Yongin-city, Kyonggi-do 449-712, Korea
Email: flyers@sait.samsung.co.kr
http://www.hcilab.com/lab_publication.html, http://myhome.naver.com/flyers

ABSTRACT

Binaural blind source separation algorithm for noisy mixtures is proposed. We consider ambient background noise signals and nonstationary target source signals. The proposed blind source separation combines signal estimation from noisy observations with source identification through mixing parameter estimation. The sparseness property of target source signals enables the proposed noisy, underdetermined, binaural blind source separation. A minimum mean square error estimator in frequency domain is implemented to estimate the signal spectra from noisy observations. The K-means clustering algorithm is utilized to identify the sources. With the help of calculating the signal absence probability for each frequency bin, noises are effectively eliminated from the target source signals and the mixing parameter estimation becomes more accurate in noisy environments.

1. INTRODUCTION

Source signals we are targeting, for example, speech, music, and a mixed sound composed of both, are naturally nonstationary. Noise is considered as an unwanted signal to be eliminated in order to get a high-quality target source signals. It can be a stationary background noise, an interfering speaker, a bursting sound when you close the door, or an intermittent sound like telephone rings. Previous works to separate high-quality sound source signals are grouped into several categories in accordance with the number of microphones used, the presence or absence of the noise sources, and the number of target sound sources to be separated. The work we are aiming here is to separate multiple sound sources from multiple noisy observations. Pioneering works like [1], [2], [3], and [4] dealt with the underdetermined blind source separation (BSS) problem, which means that the number of observations are less than the number of target sources. Most of the works assumed instantaneous mixing and were closer to the off-line algorithms except [4].

The sparseness of the short-time Fourier transform (STFT) coefficients of source signals is the main key to solve the underdetermined BSS problem in an echoic environment. This property is also known as the w-disjoint orthogonality [2]. It is reported that several sound sources satisfying this property can be separated using two microphones. However, they didn't deal with the problem of extracting target signals from noisy observations. In this paper, the problem of 'noisy' underdetermined BSS is investigated. The proposed algorithm consists of source identification and signal estimation blocks as seen in Fig 1. In the source identification block, we use the soft K-means clustering algorithm for estimating the mixing parameters, in which signal presence probability of each frequency bin is incorporated to accurately compute the cluster centers. With the help of this clustering algorithm, the proposed algorithm works online. In the signal estimation block, we utilize the minimum mean square error (MMSE) short-time spectral amplitude estimator for estimating the signal source components. To cope with the fluctuating and even abruptly changing nature of target source signals, *a priori* SNR is re-estimated by the decision-directed method.

The rest of the paper is organized as follows. Section 2 mentions about the proposed algorithm to separate N sources from M microphone observations when $M \leq N$. Source identification and source signal estimation are the main issues of this section. Experimental results of the proposed method with two microphones are reported in section 3. Conclusions will be drawn in section 4.

2. BINAURAL BLIND SOURCE SEPARATION

We consider the problem that given M microphone observations find N independent sources where $M \leq N$. To simplify the problem let $M = 2$. In an environment of small

This work was partly supported by the Critical Technology-21 Program of Korean Ministry of Science and Technology.

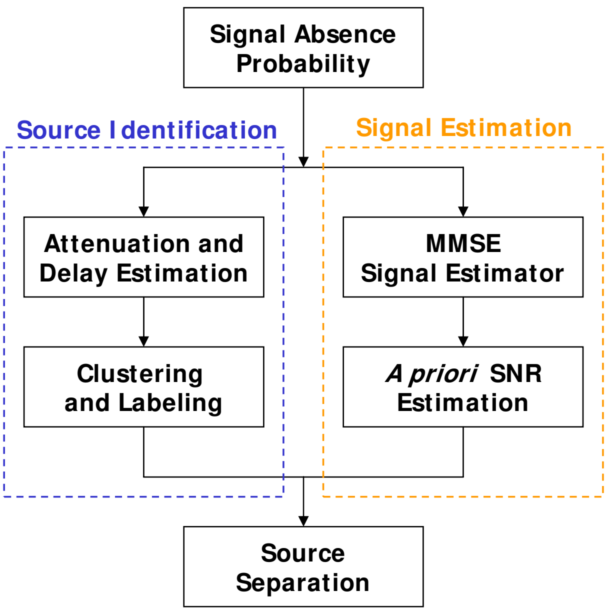


Fig. 1. Flowchart of the proposed noisy, underdetermined, binaural blind source separation algorithm.

reverberations, without loss of generality, we can model the relationship between the source signals, $s_j(t)$ and microphone observations, $x_i(t)$ by

$$x_1(t) = \sum_{j=1}^{N} s_j(t) + n_1(t), \quad (1)$$

$$x_2(t) = \sum_{j=1}^{N} a_j s_j(t - d_j) + n_2(t), \quad (2)$$

where a noise $n_1(t)$ is added in the mixture $x_1(t)$ from the background noise source, a noise $n_2(t)$ through another propagating channel is added in the mixture $x_2(t)$, a_j is a relative attenuation parameter, and d_j is the relative propagation delay between microphones due to the direction of arrival. Because we assumed the small reverberation environment, attenuation and delay parameters merge into the definition of the sources in the first mixture.

In frequency domain, assuming that the noisy signal observation $\mathbf{X}(m)$ of the first mixture is a sum of source signal $\mathbf{S}(m)$ and additive noise $\mathbf{N}(m)$, we consider the statistical model employing global hypotheses, H_0 and H_1, which indicate source signal absence and presence at m-th frame, respectively.

$$\begin{array}{ll} H_0: & \mathbf{X}(m) = \mathbf{N}(m) \\ H_1: & \mathbf{X}(m) = \mathbf{S}(m) + \mathbf{N}(m) \end{array} \quad (3)$$

The source signal in H_1 is a sum of all the individual source signals, i.e. $\mathbf{S}(m) = \sum_{j=1}^{N} \mathbf{S}_j(m)$. Moreover, since source signal absence and presence arise in each frequency bin, we further consider the statistical model employing local hypotheses, $H_{0,k}$ and $H_{1,k}$ for each frequency bin, which indicate source signal absence and presence at k-th frequency bin of the m-th frame, respectively.

$$\begin{array}{ll} H_{0,k}: & X_k(m) = N_k(m) \\ H_{1,k}: & X_k(m) = S_k(m) + N_k(m). \end{array} \quad (4)$$

Again, the source signal in $H_{1,k}$ is a sum of all the individual source signal components, i.e. $S_k(m) = \sum_{j=1}^{N} S_k^{(j)}(m)$, where $S_k^{(j)}(m)$ represents the j-th source signal component at k-th frequency bin of the m-th frame. We assumed here that the source signals have the property of 'sparsity' of the short-time Fourier transform (STFT) [1]. This is also known as the w-disjoint orthogonality as in [5]. Therefore, we consider $S_k(m) = S_k^{(j)}(m)$ for some j. It is also assumed that $X_k(m)$ and $N_k(m)$ have zero-mean complex Gaussian densities [6]. The Gaussian density model is motivated by the central limit theorem. For strong mixture, its distribution gets closer to the Gaussian density regardless of the shape of the individual source density.

2.1. Signal Absence Probability (SAP)

For the m-th frame the global SAP, $p(H_0|\mathbf{X}(m))$ and the local SAP, $p(H_{0,k}|X_k(m))$ are computed by

$$p(H_0|\mathbf{X}(m)) = \frac{p(H_0, \mathbf{X}(m))}{p(\mathbf{X}(m))} = \frac{1}{\prod_{k=1}^{M} [1 + q_k \Lambda_k(m)]}, \quad (5)$$

and

$$p(H_{0,k}|X_k(m)) = \frac{1}{1 + q_k \Lambda_k(m)}, \quad (6)$$

respectively, in which q_k is the ratio defined by

$$q_k = \frac{p(H_{1,k})}{p(H_{0,k})}, \quad (7)$$

and $\Lambda_k(m)$ is the likelihood ratio computed for the k-th frequency bin of the m-th frame as

$$\Lambda_k(m) = \frac{p(X_k(m)|H_{1,k})}{p(X_k(m)|H_{0,k})}. \quad (8)$$

In deriving (5), the following equations were utilized

$$p(H_0, \mathbf{X}(m)) = \prod_{k=1}^{M} [p(X_k(m)|H_{0,k}) p(H_{0,k})], \quad (9)$$

and

$$
\begin{aligned}
p(\mathbf{X}(m)) &= \prod_{k=1}^{M} p(X_k(m)) \\
&= \prod_{k=1}^{M} [p(X_k(m)|H_{0,k})p(H_{0,k}) \\
&\quad + p(X_k(m)|H_{1,k})p(H_{1,k})] .
\end{aligned} \tag{10}
$$

Because the STFT coefficients are uncorrelated, it is assumed that the spectral components, $X_k(m)$ are statistically independent.

We compare the global SAP with a threshold that can be set by the user. If the global SAP exceeds the threshold, the noise power estimate is updated. If the global SAP does not exceed the threshold, the noise power estimate remains the same. The local SAP is used for the frequency bin selection. Using those selected frequency bin components we can reliably estimate the attenuation and delay parameters for source identification.

2.2. Source Identification

In microphone array signal processing the time delay between adjacent microphones can be computed by searching the main peak of their cross correlation. To identify the source correctly we should acquire the information about the distance between the source and the sensor. Subspace methods like MUSIC can find the direction of arrivals and the attenuations, or equivalently, the delays and the distances. However, the MUSIC algorithm is not adequate for underdetermined mixing problems. DUET [2] algorithm can find the attenuation and the delay factors from mixture observations for underdetermined mixing problems. Therefore, we adopt the DUET as a base framework for estimating the attenuation and delay parameters.

2.2.1. Attenuation and Delay Estimation

Because the sources are assumed to be sparse in STFT coefficients, at most one source of N sources will dominantly occupy the k-th frequency bin. Based on the local SAP (6), we can determine which frequency bins are occupied by the source signals. At those frequency bins the following relationship holds for some j,

$$
\begin{bmatrix} X_{1,k}(m) \\ X_{2,k}(m) \end{bmatrix} \approx \begin{bmatrix} 1 \\ a_j e^{-i\omega d_j} \end{bmatrix} S_k^{(j)}(m), \tag{11}
$$

where $\omega = \frac{2\pi k}{L}$, L is the number of FFT points. Thus, the attenuation and delay parameters can be computed by analyzing the ratio of $X_{1,k}(m)$ and $X_{2,k}(m)$. The mixing parameters are estimated by

$$
\begin{aligned}
\mu^{(k)} &\equiv (a(k,m), d(k,m)) \\
&= \left(\left| \frac{X_{2,k}(m)}{X_{1,k}(m)} \right|, -\frac{1}{\omega} \angle \frac{X_{2,k}(m)}{X_{1,k}(m)} \right) .
\end{aligned} \tag{12}
$$

The original DUET algorithm constructed a 2-D histogram of amplitude-delay estimates and looked at the number and location of the peaks in the histogram to determine the number of sources and their mixing parameters. However, constructing the 2-D histogram requires a lot of samples and is adequate mainly for off-line computation. Rather than constructing the histogram we use the K-means clustering algorithm. The mixing parameter estimates are obtained by the cluster centers.

2.2.2. Soft K-means Clustering

At the m-th time frame, we have $L/2$ pairs of mixing parameter estimates. In the K-means algorithm, K parameter vectors called the means μ_j, which correspond to the estimated mixing parameters, are initialized to random values. Then, the algorithm is an iterative two-step algorithm [7].

In assignment step, each data point $\mu^{(k)}$, $k = 1, \cdots L/2$ is given a soft 'degree of assignment' to each of the means. We call the degree to which point $\mu^{(k)}$ is assigned to cluster j the responsibility $r_j^{(k)}$, which is the responsibility of cluster j for point $\mu^{(k)}$. The rule for assigning the responsibilities can be viewed as a 'soft-min' by

$$
r_j^{(k)} = \frac{exp\left(-\beta d(\mu_j, \mu^{(k)})\right)}{\sum exp\left(-\beta d(\mu_j, \mu^{(k)})\right)} \tag{13}
$$

where $d(\cdot)$ is a distance operator, and β is a decaying parameter and is inversely proportional to the noise variance. In update step, the means are adjusted to match the sample means of the data points that they are responsible for.

$$
\mu_j = \frac{\sum_k r_j^{(k)} \mu^{(k)}}{\sum_k r_j^{(k)}} \equiv (a_j(m), d_j(m)) , \tag{14}
$$

where $a_j(m)$ and $d_j(m)$ are the attenuation and delay parameters of the j-th source at the m-th frame, respectively. Moving averages of these parameters are taken for the estimates for the next frame. Therefore, tracking the target sources is possible when the target sources are slowly moving. Initial guess of the cluster centers for the next frame is set to the ones obtained at the current frame. It speeds up the convergence of the clustering algorithm.

Alternatively, we can use the 'hard' K-means clustering algorithm to lessen the computational burden to compute the exponential function for all the data points. While the assignment $r_j^{(k)}$ in the soft K-means algorithms involved a 'soft-min' over the distance, the rule for assigning the responsibility in the hard K-means algorithm can take on values between 0 and 1. However, even in this case, only when calculating the cluster mean after the convergence, the responsibility $r_j^{(k)}$ is assigned to a value proportional to the magnitude of the spectral component that is involved to compute the data point $\mu^{(k)}$. This is because spectral components of large magnitude are less vulnerable to corrupting noise.

2.3. Signal Estimation

2.3.1. MMSE Short-Time Spectral Amplitude Estimator

Let $X_k(m)$ denote the k-th spectral component of the first mixture $x_1(t)$, and $S_k(m)$ and $N_k(m)$ denote the source and noise component therein.

$$X_k(m) = S_k(m) + N_k(m), \tag{15}$$

where $X_k(m) = R_k e^{i\theta_k}$ and $S_k(m) = A_k e^{i\alpha_k}$. The MMSE estimation problem can be reduced to be that of estimating A_k from X_k, because the short-time spectral amplitude rather than its waveform is of major importance in human perception. The constrained complex exponential estimator is found to be the complex exponential of the noisy phase, i.e. $\alpha_k = \theta_k$ [6]. Since the spectral components are assumed to be statistically independent, the MMSE estimator $\hat{A}_k$ of A_k is obtained as follows:

$$\begin{aligned} \hat{A}_k &= E\{A_k|\mathbf{X}\} \\ &= E\{A_k|X_k\} \\ &= p(H_{1,k}|X_k)E\{A_k|X_k, H_{1,k}\} \\ &= (1 - p(H_{0,k}|X_k))\, E\{A_k|X_k, H_{1,k}\} \\ &= \frac{q_k\Lambda_k}{1 + q_k\Lambda_k} E\{A_k|X_k, H_{1,k}\} \end{aligned} \tag{16}$$

and

$$E\{A_k|X_k, H_{1,k}\} = \frac{\int_0^\infty \int_0^{2\pi} A_k p(X_k|A_k, \alpha_k) d\alpha_k dA_k}{\int_0^\infty \int_0^{2\pi} p(X_k|A_k, \alpha_k) d\alpha_k dA_k}, \tag{17}$$

where the probability density function (pdf), $p(X_k|A_k, \alpha_k)$ is given by

$$p(X_k|A_k, \alpha_k) = \frac{1}{\pi\sigma_{N_k}^2} exp\left(-\frac{1}{\sigma_{N_k}^2}|X_k - A_k e^{i\alpha_k}|^2\right). \tag{18}$$

Since complex Gaussian densities are assumed for the pdfs of the mixture and the noise, $p(A_k)$ becomes a Rayleigh distribution. $p(\alpha_k)$ is independent of $p(A_k)$ and is uniform over $[-\pi, \pi)$

$$p(A_k, \alpha_k) = p(A_k)p(\alpha_k) = \frac{A_k}{\pi\sigma_{S_k}^2} exp\left(-\frac{A_k^2}{\sigma_{S_k}^2}\right). \tag{19}$$

After some calculations (see Appendix A of [6]) we get

$$\begin{aligned} \hat{A}_k = & \frac{q_k\Lambda_k}{1 + q_k\Lambda_k} \cdot \Gamma(1.5)\frac{\sqrt{v_k}}{\gamma_k} exp\left(-\frac{v_k}{2}\right) \\ & \cdot \left[(1 + v_k)I_0\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right] R_k. \end{aligned} \tag{20}$$

$\Gamma(\cdot)$ denotes the gamma function, with $\Gamma(1.5) = \sqrt{\pi}/2$; $I_0(\cdot)$ and $I_1(\cdot)$ denote the modified Bessel functions of zero and first order, respectively. v_k is defined by

$$v_k = \frac{\xi_k}{1 + \xi_k}(\gamma_k + 1), \tag{21}$$

where ξ_k and γ_k are defined by

$$\xi_k = \frac{\sigma_{S_k}^2}{\sigma_{N_k}^2} \tag{22}$$

$$\gamma_k = \frac{R_k^2}{\sigma_{N_k}^2} - 1. \tag{23}$$

ξ_k and γ_k are known as the *a priori* and *a posteriori* SNR, respectively. Due to the target source nonstationarity, the variance of each spectral component of the target source is time-varying. Thus, *a priori* SNR should be re-estimated in each analysis frame. In general, *a posteriori* SNR should be also re-estimated because in real environment there is no noise that is definitely stationary.

It is interesting to examine the asymptotic behavior of $\hat{A}_k$ at high SNR ($\xi >> 1$). In this case, the MMSE short-time spectral amplitude estimate $\hat{A}_k$ converges to the Wiener amplitude estimator.

$$\hat{A}_k \approx \frac{\xi_k}{1 + \xi_k} R_k \tag{24}$$

2.3.2. A Priori and A Posteriori SNR Update

The real acoustic environment is continuously changing. To adapt the time-varying nature of source and noise signals, the signal powers, i.e. the signal variances, are updated at every time frame. Based on the global SAP, the noise variance, $\sigma_{N_k}^2$ is updated only when the target signals are absent. Source signal variance $\sigma_{S_k}^2$ is always updated. Under the general stationarity assumption of background noise, we use the long-term power spectra of the background noise as the estimates for $\sigma_{N_k}^2$.

$$\sigma_{N_k}^2(m) = \zeta_{N_k}\sigma_{N_k}^2(m-1) + (1 - \zeta_{N_k})|N_k(m)|^2, \tag{25}$$

where $0 < \zeta_{N_k} < 1$ is a smoothing parameter. Because the MMSE estimator is a function of *a priori* and *a posteriori* SNRs, they should be re-estimated [8]. *A posteriori* SNR, γ_k is updated using the moving-averaged noise variance estimate (25). However, *a priori* SNR is not a simple ratio of the source and the noise variances because of the nonstationarity of the source signals. For abruptly changing source signals, we use the 'decision-directed' method by

$$\xi_k(m) = \zeta_{SNR} \frac{\hat{A}_k(m\text{–}1)}{\sigma^2_{N_k}(m\text{–}1)} + (1-\zeta_{SNR}) P\left(\gamma_k(m)\right), \quad (26)$$

where $\hat{A}_k(m\text{–}1)$ is the amplitude estimate of the kth spectral component at the $(m-1)$ analysis frame, and $P(\cdot)$ is an operator which is defined by

$$P(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \quad (27)$$

$P(\cdot)$ is used to ensure the positiveness of the proposed estimator in case $\gamma_k(m)$ is negative. As γ_k defined in (23) is not necessarily positive, the operator P guarantees that ξ_k is always nonnegative or, equivalently, that the expression of the gain given by (20) is valid. By doing so, the smoothing effect of the *a priori* SNR is achieved when the *a posteriori* SNR is fluctuating. If there is an abrupt change in the *a posteriori* SNR, the *a priori* SNR can follow it with a one frame delay [9]. This property of the *a priori* SNR helps it to reduce the musical noise effect.

2.4. Source Separation

From the source identification and signal estimation methods described above we have indices k which correspond to each j-th source and the estimated spectra of the sources

$$\hat{S}_k(m) = \hat{A}_k e^{\imath \hat{\alpha}_k}, \quad (28)$$

where $\hat{\alpha}_k = \theta_k$. Source separation is done with a proper pairing of k and $\hat{S}_k(m)$. Based on (4), we can rewrite the equation by

$$\hat{S}_k(m) = \hat{S}_k^{(1)}(m) + \cdots + \hat{S}_k^{(N)}(m). \quad (29)$$

Based on the assumption of the sparseness of source signals, it further reduces to

$$\hat{S}_k(m) = \hat{S}_k^{(j)}(m). \quad (30)$$

To separate the sources, one creates the label vector corresponding to each source and applies the each label vector to the estimated spectrum vector $\hat{\mathbf{S}}(m)$ to produce the separated source spectrum. Define the label vector $\mathcal{L}_j(m)$, which maps the relationship between the support of source j and $\hat{S}_k^{(j)}(m)$. Each component of the label vector has the value 0 or 1. Then, one obtains spectrum vector of the j-th source at the m-th frame by

$$\hat{\mathbf{S}}_j(m) = \mathcal{L}_j(m) \odot \hat{\mathbf{S}}(m), \quad (31)$$

where $\odot$ represents the component-wise multiplication of vectors.

3. EXPERIMENTS

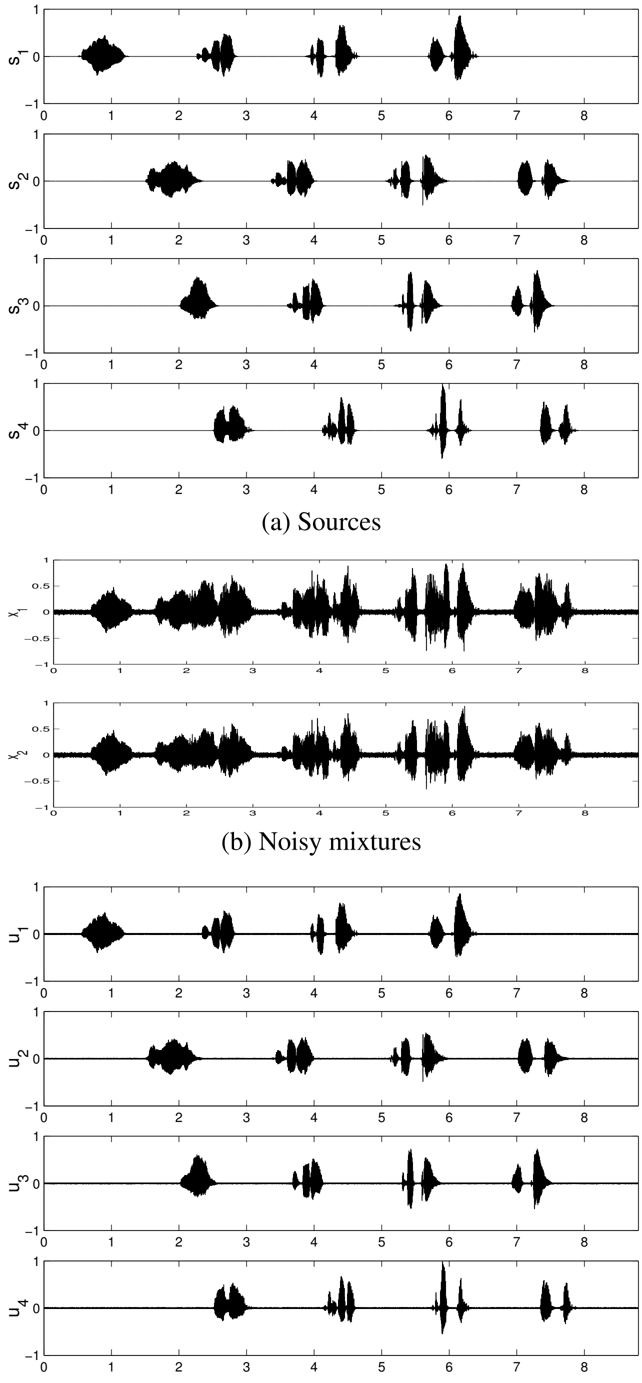
To verify the performance of the proposed noisy, underdetermined blind source separation algorithm, we performed experiments on speech signals. Clean speech data were previously recorded at 16kHz in an anechoic room. 40 words for computer commands and Korean names uttered by four male speakers were used for making the target signals. 40 words pronounced by each speaker split into 10 files. Each file consists of four words. Thus, 10 sets of target sources were prepared. White Gaussian noise (WGN) was used as a broad-band noise. According to the SNR, WGN was simply added sample by sample after adjusting the signal levels by the method described in the ITU-T recommendation P.830. Actually, the mixing was done on a laptop computer. Source identification and signal estimation were done frame by frame. Figure 2 shows an experimental result of the proposed algorithm. As one can see, the proposed method could separate individual sources well while reducing the noises significantly and effectively in real-time. The quality of the recovered speech was quite good to hear.

In section 2.1, the SAP is calculated over the mixture signal. However, for the signals with a low SNR it is undesirable to use the mixture signal. Even for the two noisy observations, a beamforming technique, for example, a delay-and-sum beamformer can improve the SNR by simply computing the time delay between the sensor observations and adjusting them. Using the beamformer output to calculate the SAP is highly recommended in a low SNR environment.

4. CONCLUSIONS

A noisy, underdetermined blind source separation problem has been addressed in this paper. The proposed algorithm can separate N individual sources only from binaural noisy observations. This is possible for the source signals which has the sparseness property of short-time Fourier transform coefficients. This assumption states that at most one source occupies a given time-frequency point. However, when the broad-band noise signal corrupts the clean mixture of source signals, for example, the mixture consists of source signals and a white Gaussian noise, the assumption we made is easy to be broken. This shortcoming can be overcome by incorporating the signal absence probability for each frequency bin with the estimation of the attenuation and the delay parameters. The signal absence probability is also involved in estimating the target source signals buried in an ambient background noise. The minimum mean square error short-time spectral amplitude estimator is utilized as the source signal estimator. Because the short-time spectral amplitude rather than its waveform is of major importance in human perception, the phase estimator is not used, i.e. the phase remains intact.

There are many situations that the sparseness property is not valid. At a given time-frequency point, more than one sources could co-exist. We are now investigating this problem as well as the problem of coherent source signals separation in a highly reverberant environment.



(a) Sources

(b) Noisy mixtures

(c) Separated sources

Fig. 2. Four original sources, two noisy mixtures, and four denoised and separated sources. A white Gaussian noise of 10dB was used in this experiment.

5. REFERENCES

[1] P. Bofill and M. Zibulevsky, "Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation*, 2000, pp. 87–92.

[2] A. Jourjine, S. Rickard, and Ö. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, 2000, vol. 5, pp. 2985–2988.

[3] T-W. Lee, M. Lewicki, M. Girolami, and T. Sejnowski, "Blind source separation of more sources than mixtures using overcomplete representations," *IEEE Signal Processing Letters*, vol. 6, no. 4, pp. 87–90, 1999.

[4] S. Rickard, R. Balan, and J. Rosca, "Real-time time-frequency based blind source separation," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation*, 2001, pp. 651–656.

[5] S. Rickard and Ö. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *Proc. Int. Conf. on Acoust., Speech, Signal Processing*, 2002, vol. 1, pp. 529–532.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[7] D. Mackay, *Information Theory, Inference and Learning Algorithms*, electronic version is available at http://www.inference.phy.cam.ac.uk/mackay/MyBooks.html.

[8] C. Choi, S. Choi, and S-R. Kim, "Speech enhancement using sparse code shrinkage and global soft decision," in *Proc. Int. Conf. on Independent Component Analysis and Blind Signal Separation*, 2001.

[9] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 345–349, 1994.