THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA VOL.23, NO.2E 2004. 6 pp. 56~67

Multi-channel Speech Enhancement Using Blind Source Separation and Cross-channel Wiener Filtering

Gil-Jin Jang*, Changkyu Choi*, Yongbeom Lee*, Jeongsu Kim*, Sangryong Kim*
*Human Computer Interaction Laboratory, Samsung Advanced Institute of Technology
(Received June 2 2004; accepted June 30 2004)

Abstract

Despite abundant research outcomes of blind source separation (BSS) in many types of simulated environments, their performances are still not satisfactory to be applied to the real environments. The major obstacle may seem the finite filter length of the assumed mixing model and the nonlinear sensor noises. This paper presents a two-step speech enhancement method with multiple microphone inputs. The first step performs a frequency-domain BSS algorithm to produce multiple outputs without any prior knowledge of the mixed source signals. The second step further removes the remaining cross-channel interference by a spectral cancellation approach using a probabilistic source absence/presence detection technique. The desired primary source is detected every frame of the signal, and the secondary source is estimated in the power spectral domain using the other BSS output as a reference interfering source. Then the estimated secondary source is subtracted to reduce the cross-channel interference. Our experimental results show good separation enhancement performances on the real recordings of speech and music signals compared to the conventional BSS methods.

Keywords: *Blind source separation (BSS), Spectral subtraction, Wiener filtering, Adaptive noise cancellation (ANC).*

I. Introduction

Separation of multiple signals from their superposition recorded at several sensors is an important problem that shows up in a variety of applications such as communications, biomedical and speech processing. The class of separation methods that require no source signal information except the number of mixed sources is often referred to blind source separation (BSS)[1]. In real recording situations with multiple microphones, each source signal spreads in all directions and reaches each microphone through "direct paths" and "reverberant paths." The observed signal by the j th microphone input is expressed as

$$x_j(t)=\sum_{i=1}^{N}\sum_{\tau=0}^{\infty}h_{ji}(\tau)s_i(t-\tau)+n_j(t)=\sum_{i=1}^{N}h_{ji}(t)*s_i(t)+n_j(t) \quad (1)$$

where $s_i(t)$ is the i th source signal, N is the number of sources, $x_j(t)$ is the observed signal, and $h_{ji}(t)$ is the transfer function from source i to sensor j. The noise term $n_j(t)$ refers to the nonlinear distortions due to the characteristics of the recording devices. The assumption that the sources never move often fails due to the dynamic nature of the acoustic objects [1]. Moreover the practical systems should set a limit on the length of an impulse response, and the limited length is often a major performance bottleneck in realistic situations[2].

This paper proposes a post-processing technique for eliminating the remaining cross-channel interference at the BSS output. Our method is motivated by adaptive noise cancellation (ANC)[3]. The proposed method considers one BSS output as noisy signal and the other as reference noise source, and performs cancellation in the power spectral domain as the conventional spectral subtraction methods do[4]. The advantage of the power spectral subtraction is the

Corresponding author: Gil-Jin Jang, (giljin.jang@samsung.com)
Human Computer Interaction Laboratory, Samsung Advanced Institute of Technology Mt. 14-1, Nongseo-Ri, Giheung-Eup, Yongin-Si, Gyeonggi-Do, 449-712, South Korea

effective absorption of small amount of mismatch between the actual filter and the estimated one, and the generation of cleanly denoised signals. The disadvantage is the introduction of the musical noises due to the below-zero spectral components as a result of the subtraction. With the help of source absence/ presence detection prior to the subtraction, we reduce the error of the cancellation factor estimation and hence minimize the musical noises. Experimental results show that our proposed method has a superior performance to the conventional spectral subtraction on the output of the frequency-domain BSS method in realistic conditions.

II. Frequency-domain Blind Source Separation

The frequency-domain blind source separation algorithm for convolutive mixtures is to transform the original time-domain filtering architecture into an instantaneous BSS problem in the frequency domain [5]. Using short time Fourier transform, (1) is rewritten as

$$\mathbf{X}(\omega,n)=\mathbf{H}(\omega)\mathbf{S}(\omega,n)+\mathbf{N}(\omega,n)\ , \tag{2}$$

where ω is a frequency index, $\mathbf{H}(\omega)$ is the $N\times N$ square mixing matrix, $\mathbf{x}(\omega,n)=\left[X_1(\omega,n)X_2(\omega,n)\mathrm{K}\ X_N(\omega,n)\right]^T$ and $X_j(\omega,n)=\sum_{\tau=0}^{T-1}e^{-12\pi\omega\tau/T}x_j(t_n+\tau)$, representing the DFT of the frame of size T with shift length $\lfloor T/2\rfloor$ starting at $t_n=\lfloor T/2\rfloor(n-1)+1$ where $\lfloor\cdot\rfloor$ is the flooring operator, and corresponding expressions apply for $\mathbf{S}(\omega,n)$ and $\mathbf{N}(\omega,n)$ --- in our paper, we denote lowercase letters with argument t for the time-series, and capital letters with argument ω and n for the Fourier transform at frequency ω for the n th frame. When the letters are boldfaced, they are column vectors whose components are accompanying the same arguments. The unmixing process can be formulated in a frequency bin ω:

$$\mathbf{Y}(\omega,n)=\mathbf{W}(\omega)\mathbf{X}(\omega,n)\ , \tag{3}$$

where $N\times 1$ vector $\mathbf{Y}(\omega,n)$ is an estimate of the original source $\mathbf{S}(\omega,n)$ disregarding the effect of the noise $\mathbf{N}(\omega,n)$. The convolution operation in the time domain corresponds to the element-wise complex multiplication in the frequency domain. The instantaneous ICA algorithm we use is the non-holonomic information maximization [6]:

$$\Delta\mathbf{W}\propto\left[\varphi(\mathbf{Y})\mathbf{Y}^H-\mathrm{diag}\left(\varphi(\mathbf{Y})\mathbf{Y}^H\right)\right], \tag{4}$$

where ^{H} is the Hermitian transpose, and the polar nonlinear function $\varphi(\mathbf{Y})$ is component-wisely defined as $\varphi(Y_i)=Y_i/|Y_i|$ [7]. A disadvantage of this decomposition is that there arises the permutation problem in each independent frequency bin. The problem is solved by the time-domain spectral smoothing[5].

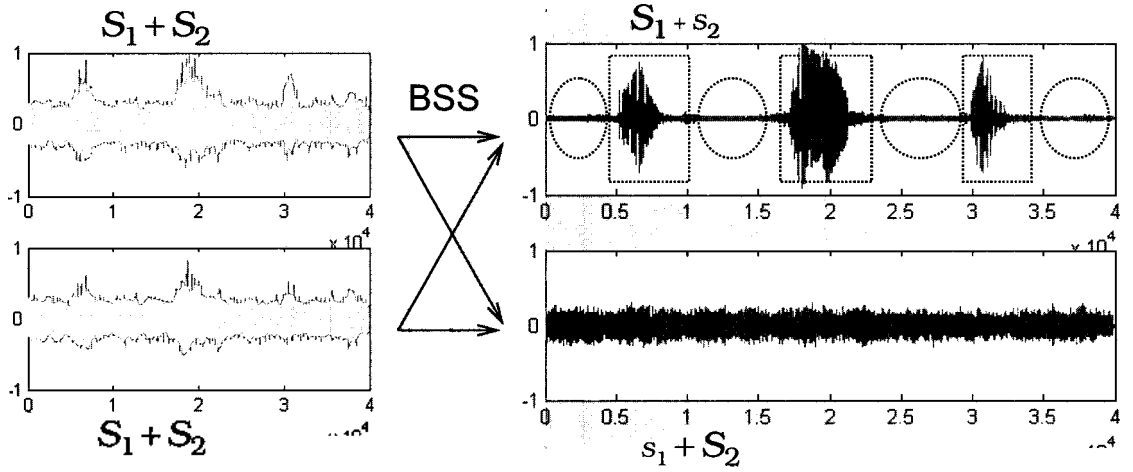


Fig. 0. The separability of the ordinary BSS algorithm. Left two signals are sensor inputs, and right two signals are BSS outputs. The original sources are rock music and speech signals[8]. There exists no speech signal in the ellipse-marked parts but still remains a small amount of rock music signal.

III. Adaptive Cross-Channel Interference Cancellation

3.1. Cross-Channel Interference Detection

Figure 1 illustrates inputs and outputs of the ordinary BSS. The output signals still contain cross-channel interference that is audible and identifiable by human listeners. However, in the first output, if we assume that the speech signal is present only in the region enclosed by rectangles (call them active blocks), apparently the region enclosed by ellipses (inactive blocks) contains the music signal only. The existence of the cross-channel interference can be described by the presence of the primary source. When the primary source is present, it often coexists with the secondary source, and the interference occurs. The presence probability of the primary source is modeled by complex Gaussian distributions[9], and used in properly estimating the interference cancellation factors regarding the cross-channel output as a reference noise source.

For each frame of the i th BSS output, two hypotheses $H_{i,0}$ and $H_{i,1}$ are given which respectively indicate the absence and presence of the primary source:

$$\begin{aligned} H_{i,0} &: \forall\omega, Y_i(\omega,n) = \textstyle\sum_{i\neq j} G_{ij}(\omega)S_j(\omega,n), \\ H_{i,1} &: \forall\omega, Y_i(\omega,n) = G_{ii}(\omega)S_i(\omega,n) + \textstyle\sum_{i\neq j} G_{ij}(\omega)S_j(\omega,n), \\ & \quad where \quad G_{ij}(\omega) = \textstyle\sum_{k=1}^{N} W_{ik}(\omega)H_{kj}(\omega) \end{aligned} \tag{5}$$

where $W_{ik}(\omega)$ is (i,k) -component of the unmixing matrix $\mathbf{W}(\omega)$, and $H_{kj}(\omega)$ is (k,j) -component of the mixing matrix $\mathbf{H}(\omega)$. Although $G_{ii}(\omega)=1, G_{ij}(\omega)=0$ for $i\neq j$ when ideal separation is obtained, it is not always possible due to the nature of the instantaneous ICA algorithm and the additive noises in the real recordings. The hypothesis $H_{i,0}$ means that the primary source is absent in the i th BSS output at frame n, and $H_{i,1}$ means that the primary source as well as the secondary sources coexist, that is, interference occurs. Conditioned on a set of all the frequency components for frame n, $Y_i(n)=\{Y_i(\omega,n)|\omega=1,\mathrm{K},T\}$, the source absence and presence probabilities are given by

$$p(H_{i,m}|Y_i(n)) = \frac{p(Y_i(n)|H_{i,m})p(H_{i,m})}{p(Y_i(n)|H_{i,0})p(H_{i,0}) + p(Y_i(n)|H_{i,1})p(H_{i,1})}, \tag{6}$$

where $p(H_{i,0})$ is a priori probability for source i absence (inactive frames), and $p(H_{i,1})=1-p(H_{i,0})$ is that of source i presence (active frames). Assuming the probabilistic independence among the frequency components,

$$p(Y_i(n)|H_{i,m}) = \prod_{\omega} p(Y_i(\omega,n)|H_{i,m}) \ . \tag{7}$$

Then source absence probability becomes

$$p(H_{i,0}|Y_i(n)) = \left[1 + \frac{1-P(H_{i,0})}{P(H_{i,0})}\prod_{\omega}^{T}\frac{p(Y_i(\omega,n)|H_{i,1})}{p(Y_i(\omega,n)|H_{i,0})}\right]^{-1} . \tag{8}$$

The posterior probability of $H_{i,1}$ is obviously source presence probability indicating the amount of cross-channel interference at the i th BSS output, which is easily computed by $p(H_{i,1}|Y_i(n))=1-p(H_{i,0}|Y_i(n))$. In the following sections, we explain the cancellation of the cross-channel interference and the statistical models for the component densities $p(Y_i(\omega,n)|H_{i,m})$.

3.2. Cross-Channel Interference Cancellation

Adaptive noise cancellation (ANC) is one of the powerful techniques when a reference noise source is given. Because the assumed mixing model of ANC is a linear FIR filter architecture, direct application of ANC may not model the mismatch of the linear filter to the realistic conditions --- nonlinearities due to the sensor noise and the infinite filter length. Therefore we add a nonlinear feature adopted in conventional spectral subtraction [4]:

$$\begin{aligned} |U_i(\omega,n)|^a &= f\left(|Y_i(\omega,n)|^a - \alpha_i \textstyle\sum_{j\neq i} b_{ij}(\omega)|Y_j(\omega,n)|^a\right), \\ \angle U_i(\omega,n) &= \angle Y_i(\omega,n) \end{aligned} \tag{9}$$

where $Y_i(\omega,n)$ is the i th component of the BSS output $\mathbf{Y}(\omega,n)$, $b_{ij}(\omega)$ is the cross-channel interference cancellation factor for frequency ω from channel j to i. A positive constant a is first introduced by Weiss [10] to incorporate the statistical property of the power spectrum to enhance the spectral subtraction performance. The value of a is typically between 1 and 2. α_i is also a positive constant and usually called over-subtraction factor [11], and used to suppress residual musical noise at the subtracted spectrum. In our work, α_i is assigned a value according to the source absence probability. The spectral flooring function $f(\cdot)$ is defined by [11]

$$f(x)=\begin{cases} x & \text{if} \quad x \geq \varepsilon \\ \varepsilon & \text{if} \quad x < \varepsilon \end{cases}, \tag{10}$$

where the positive constant ε sets a lowerbound on the spectrum value and prevents below-zero power spectrum. The nonlinear operator $f(\cdot)$ suppresses the remaining errors of the BSS, but may introduce musical noises as most of spectral subtraction techniques suffer. The subtraction in (9) can be expressed in another form, such that

$$|U_i(\omega,n)|^a = H_i(\omega)|Y_i(\omega,n)|^a \ , \tag{11}$$

where $H_i(\omega)$ is called Wiener filter and approximated by (9) as

$$H_i(\omega) \approx f\left(\frac{|Y_i(\omega,n)|^a - \alpha_i \sum_{j \neq i} b_{ij}(\omega)|Y_j(\omega,n)|^a}{|Y_i(\omega,n)|^a} \right), \tag{12}$$

3.3. Probability Model and Cancellation Factor Update

If the subtraction in (9) successfully removes the cross-channel interference, the spectral magnitude $U_i(\omega,n)$ would be zero in inactive frames. We evaluate the posterior probability of $Y_i(\omega,n)$ given each hypothesis by the complex Gaussian distributions of $U_i(\omega,n)$:

$$p(Y_i(\omega,n)|H_{i,m}) \cong p(U_i(\omega,n)|H_{i,m}) \propto \exp\left[-\frac{|U_i(\omega,n)|^2}{\lambda_{i,m}(\omega)} \right], \tag{13}$$

where $\lambda_{i,m}(\omega)$ is the variance of the subtracted frames. When $m=1$, it is the variance of the primary source, and when $m=0$ it is of the secondary source. The variance $\lambda_{i,m}(\omega)$ is updated at every frame by the following probabilistic averaging formula:

$$\lambda_{i,m} \Leftarrow \{1-\eta_\lambda p(H_{i,m}|Y_i(n))\}\lambda_{i,m} + \eta_\lambda p(H_{i,m}|Y_i(n))|U_i(\omega,n)|^2 , \tag{14}$$

where the positive constant η_λ defines the adaptation frame rate. The value of η_λ is empirically determined. The primary source signal is expected to be at least "emphasized" by BSS. Hence we assume that the amplitude of the primary source should be greater than that of the interfering one, which is primary in the other BSS output channel. While updating the model parameters, it might happen that the variance of the enhanced source, $\lambda_{i,1}(\omega)$, becomes smaller than $\lambda_{i,0}(\omega)$. Such cases are undesirable, so we explicitly change two models when

$$\Sigma_\omega \lambda_{i,0}(\omega) > \Sigma_\omega \lambda_{i,1}(\omega) \ . \tag{15}$$

The next step is updating the interference cancellation factors. First we compute the difference between the spectral magnitude of Y_i and Y_j at frequency ω and frame n:

$$\delta_i(\omega,n) = |Y_i(\omega,n)| - \Sigma_{j \neq i} b_{ij}(\omega)|Y_j(\omega,n)| \ . \tag{16}$$

We define a cost function J by ν-norm of the difference multiplied by the frame probability:

$$J(\omega,n) = p(H_{i,0}|Y_i(n)) \cdot |\delta_i(\omega,n)|^\nu \ . \tag{17}$$

The gradient-descent learning rules for b_{ij} at frame n is

$$\Delta b_{ij}(\omega) \propto -\frac{\partial J(\omega,n)}{\partial b_{ij}(\omega)} = p(H_{i,0}|Y_i(n)) \cdot |\delta_i(\omega,n)|^{\nu-1} \cdot Y_j(\omega,n) \ . \tag{18}$$

According to the earlier findings about the statistical property of natural sounds, ν is set to be less than 1 for highly kurtotic speech signals [12], greater than 1 for music signals [13], and 2 for pure Gaussian random noises. In the case of the speech signal mixtures, we assign $\nu = 0.8$ for $p\left(H_{i,1}\middle|Y_i(n)\right)$, and $\nu = 1.5$ for $p\left(H_{i,0}\middle|Y_i(n)\right)$ to fit to the distribution of the musical noises that are frequently observed in the inactive frames as a result of spectral subtraction.

3.4. Stepwise Description of the Proposed Algorithm

We design an online, multi-channel source separation algorithm by combining the frequency-domain BSS algorithm and the adaptation formulas presented in the preceding sections. The whole procedure is summarized as the following steps:

Inputs: $\mathbf{X}(\omega,n)$ - N mixture signals at frequency $\omega = 1,\ldots,T$ and frame $n = 1,\ldots,\#\ frames$

$p\left(H_{i,0}\right)$ - prior probability of the null hypothesis

Intermediate outputs (to be adapted online):

At each frequency ω, for source (or channel) i and j

$\mathbf{W}(\omega)$ - demixing matrix

$b_{ij}(\omega)$ - interference cancellation factor

$\lambda_{i,0}(\omega), \lambda_{i,1}(\omega)$ - variances of the sources for Hypotheses 0 and 1

Final outputs: interference-cancelled results $U_i(\omega,n)$ at each frequency ω and frame n

Procedures:

1. Take some initial values for $\mathbf{W}(\omega)$, $b_{ij}(\omega)$, and $\lambda_{i,0}(\omega), \lambda_{i,1}(\omega)$
2. For each frame n,
   2.1 For each frequency ω
      2.1.1 Compute BSS outputs
      $$\mathbf{Y}(\omega,n) = \mathbf{W}(\omega)\mathbf{X}(\omega,n)$$
      2.1.2 Update BSS separation matrix
      $$\Delta\mathbf{W} \propto \left[\varphi(\mathbf{Y})\mathbf{Y}^H - \mathrm{diag}\left(\varphi(\mathbf{Y})\mathbf{Y}^H\right)\right]$$
      2.1.3 Eliminate cross-channel interference and

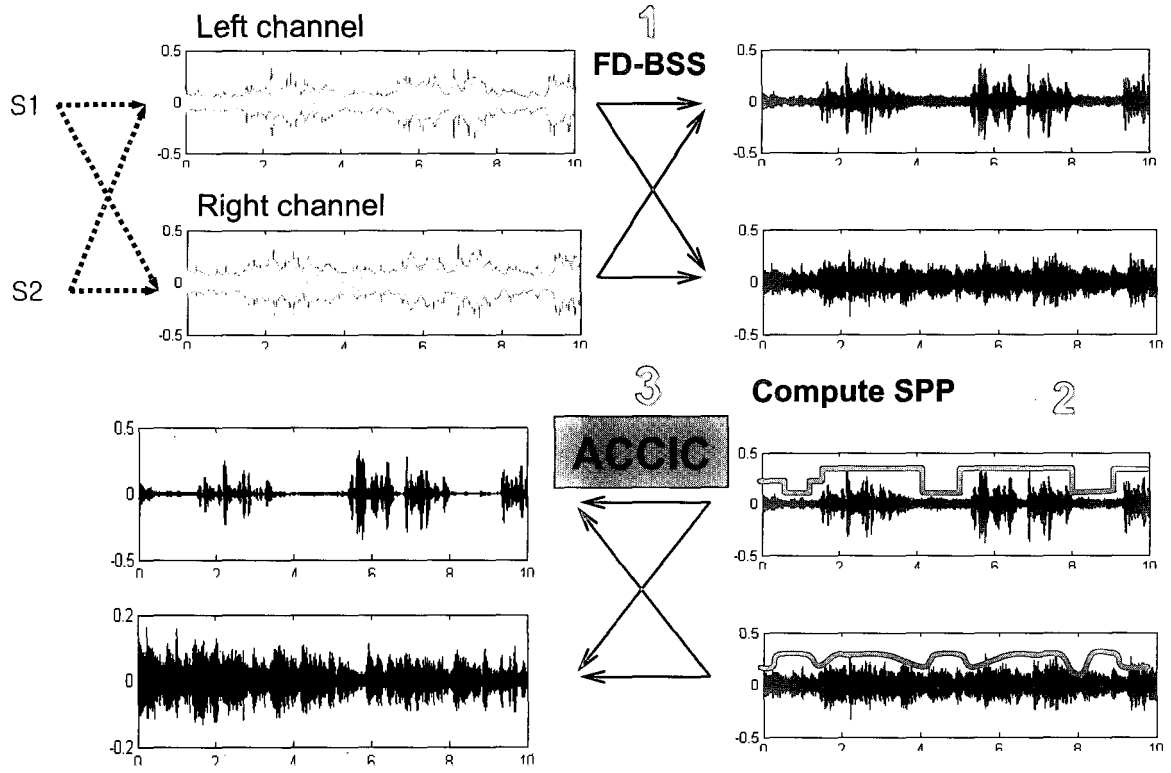


Fig. 0. Block diagram of the whole separation algorithm. (1) The input microphone recordings are initially separated by FD-BSS algorithm. In the resultant signals, one primary source and the other secondary source signals are identified at each output. (2) Source presence probabilities (SPPs) are computed for each frame of the FD-BSS outputs. (3) The proposed method (ACCIC; adaptive cross-channel interference cancellation) is performed on a single frame basis, that is, either buffering or batch processing is required.

generate output by current estimates

$$|U_i(\omega,n)|^a = f\left(|Y_i(\omega,n)|^a - \alpha_i \sum_{j\neq i} b_{ij}(\omega)|Y_j(\omega,n)|^a\right)$$

$$\angle U_i(\omega,n) = \angle Y_i(\omega,n)$$

2.1.4 Compute posterior probability of the BSS outputs

$$p(Y_i(\omega,n)|H_{i,m}) \propto \exp\left[-\frac{|U_i(\omega,n)|^2}{\lambda_{i,m}(\omega)}\right]$$

2.2 Compute source absence probability

$$p(H_{i,0}|Y_i(n)) = \left[1+\frac{1-P(H_{i,0})}{P(H_{i,0})}\prod_{\omega}^{T}\frac{p(Y_i(\omega,n)|H_{i,1})}{p(Y_i(\omega,n)|H_{i,0})}\right]^{-1}$$

2.3 For each of frame n, compute

$$\delta_i(\omega,n) = |Y_i(\omega,n)| - \sum_{j\neq i} b_{ij}(\omega)|Y_j(\omega,n)|$$

2.4 Update interference-canceling factor

$$\Delta b_{ij}(\omega) \propto p(H_{i,0}|Y_i(n)) \cdot |\delta_i(\omega,n)|^{\nu-1} \cdot Y_j(\omega,n)$$

2.5 Update variances

$$\lambda_{i,m} \leftarrow (1-\eta_\lambda p(H_{i,m}|Y_i(n)))\lambda_{i,m} + \eta_\lambda p(H_{i,m}|Y_i(n))|U_i(\omega,n)|^2$$

2.6 Change the values of the variances when

$$\sum_{\omega}\lambda_{i,0}(\omega) > \sum_{\omega}\lambda_{i,1}(\omega)$$

3. Repeat all steps in 2 until the end of input.

Whenever a frame is entered, steps 2.1 through 2.6 are performed and their outputs $U_i(\omega,n)$ are generated with no delay. The separated outputs are transformed to the time domain by performing inverse Fourier transform and overlap-addition on $U_i(\omega,n)$. Since the algorithm is completely online and adaptive, applications requiring real-time responses can adopt the proposed method. The whole procedure is illustrated in Figure 2. For compactness purpose, we refer to the frequency-domain blind source separation algorithm as FD-BSS, and the proposed algorithm as ACCIC (adaptive cross-channel interference cancellation) in the rest of the paper.

IV. Evaluation

We conducted experiments designed to demonstrate the performance of the proposed method. To show the validity of the proposed method, we measured the separation quality improvement of ACCIC for the FD-BSS outputs. We compared the results with those of the ordinary spectral subtraction method and showed the superior improvements.

4.1. Data

The data are recorded in a normal office room.

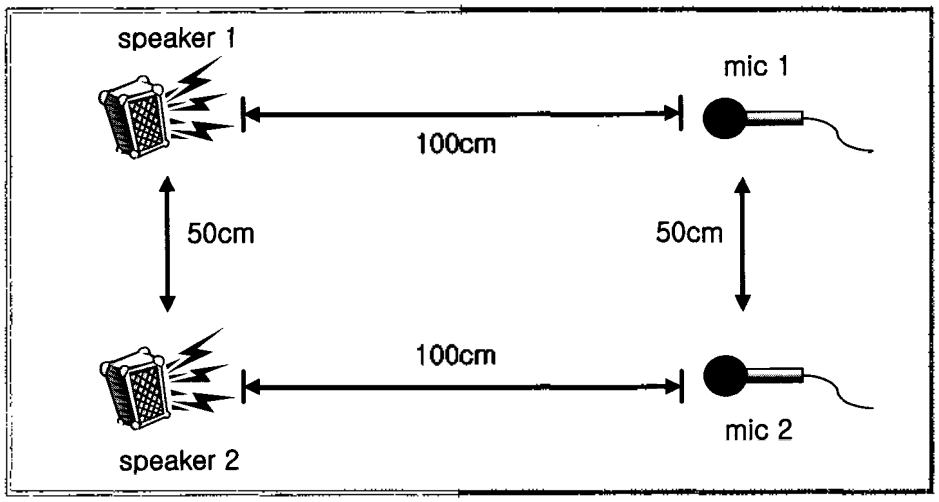


Fig. 0. Recording setup. Two loudspeakers are placed 50 cm apart from each other. One speaker plays primary source signal and the other plays secondary source simultaneously. Two omnidirectional microphones are placed 50 cm apart from each other and 100 cm from the speakers. They collect the convolutively mixed audio sounds. We assume that the source signal whose recorded amplitude is larger in microphone 1than in microphone 2 is the primary source, and the other source is secondary source. The recording was done in a normal office room.

Table 1. Source signals that are used in the experiment. A female speech signal 'f1' and a male speech 'm1' is played at 'speaker 1' in our recording configuration shown in Figure 3. At 'speaker 2', 5 different sounds of male and female speech, and 3 musics with different characteristics, are played.

	Symbol	Type	Description
Played at 'speaker 1'	'f1'	Female speech	Reading a given text in a laboratory
	'm1'	Male speech	Reading a given text in a laboratory
Played at 'speaker 2'	'g1'	Pop song	Smooth and calm; a background singer, and two people interchangeably crooning
	'g2'	Rock music	Loud; a male vocal vigorously singing
	'g3'	Instrumental music	Very soft: composed of slow piano playing only
	'f2'	Female speech	Reading a given text in a laboratory
	'm2'	Male speech	Reading a given text in a laboratory

Microphones and speakers are placed in a rectangular arrangement as shown in Figure 3. Two loudspeakers that are placed 50 cm apart from each other play the different sources and two omnidirectional microphones placed 100 cm apart simultaneously record the mixed source signals at a sampling rate of 16 kHz. The distance between microphones is 50 cm, and between the speakers is 50 cm. The left speaker plays one of male and female speech signals, and the right speaker plays one of 5 different sounds at a time. The speech signals are a series of full sentence utterances, and the music signals are a pop song, a rock with vocal sounds, and a soft instrumental music. See Table 1 for details. The length of the frame for FD-BSS was 512 samples, and the same length is used for the cross-channel interference cancellation.

4.2. Quality Measures and Alternative Method Description

The separation results are measured by signal to interference ratio (SIR), which we define the logarithm of the ratio of the primary source power to the secondary source power in a channel:

$$\mathrm{SIR}(u_i)[\mathrm{dB}] = 10\log_{10}\left[\frac{E_1(u_i)}{E_2(u_i)}\right] \cong 10\log_{10}\left[\frac{E_{1+2}(u_i)-E_2(u_i)}{E_2(u_i)}\right],$$

where $E_1(u_i)$ and $E_2(u_i)$ are the average power of primary and secondary source in signal u_i, and $E_{1+2}(u_i)$ is the average power when cross-interference occurs. When the two sources are uncorrelated, we can approximate $E_1 \approx E_{1+2} - E_2$. Because the exact signal is unable to obtain, we use the source absence and presence probabilities to evaluate the source powers:

$$E_2(u_i) = \frac{\sum_n p(H_{i,0}|Y_i(n))\langle u_i(t)^2\rangle_n}{\sum_n p(H_{i,0}|Y_i(n))}$$

$$E_{1+2}(u_i) = \frac{\sum_n p(H_{i,1}|Y_i(n))\langle u_i(t)^2\rangle_n}{\sum_n p(H_{i,1}|Y_i(n))},$$

where $\langle u_i(t)^2\rangle_n$ is the average sample power of frame n.

To show the benefit of the proposed ACCIC, we adopt conventional spectral subtraction method [4] as an alternative. This is simply performing spectral subtraction after FD-BSS, in Step 3 of Figure 2, instead of ACCIC. We use the source presence probabilities computed during the execution of ACCIC, represented by solid lines in the lower-right part of Figure 2, in estimating the average noise spectra. The implementation details are as follows:

Inputs: $Y(\omega,n)$ - N BSS output signals at frequency $\omega = 1,...,T$ and frame $n = 1,...,\#\ frames$

$p(H_{i,0}|Y_i(n))$ - posterior probability of the null hypothesis, at frame n, computed by ACCIC

Intermediate outputs (to be adapted online):

$\bar{N}_i$ - average noise spectra estimates at frame n and source i

Final outputs: $\hat{U}_i(\omega,n)$ - noise-reduced results at each frequency ω and frame n

Procedures:

Take initial value for $\bar{N}_i$, by an average of the first K frames' spectra

From frame K+1 to end of the input,

Update average noise spectra by weighted averaging

$$\bar{N}_i \Leftarrow \left(1 - c_N p\left(H_{i,0} \middle| Y_i(n)\right)\right)\bar{N}_i + c_N p\left(H_{i,0} \middle| Y_i(n)\right) Y_i(n)$$

where c_N is the adaptation frame rate for noise spectra

Subtract noise spectra and obtain current source estimates

$$\left|\hat{U}_i(\omega,n)\right|^a = f\left(\left|Y_i(\omega,n)\right|^a - \alpha_i \left|N_i(\omega,n)\right|^a\right)$$

$$\angle \hat{U}_i(\omega,n) = \angle Y_i(\omega,n)$$

The constant c_N is set to be 0.02, meaning an average over 50 = 1/0.02 frames. The conditions ($a, \alpha_i, f(\cdot)$) are same as ACCIC. Since spectral subtraction is a popular speech enhancement technique, we consider it as a conventional background denoising approach for BSS. Similar effort has been tried in [14], using wavelet filterbanks instead of spectral subtraction. The performances over the real recordings are reported in the following sections.

4.3. Experimental Results

Table 2 reports the SIRs of the input signals, outputs of FD-BSS, noise-reduced results of spectral subtraction on the FD-BSS outputs, and the interference-canceled results with the proposed ACCIC. The music signals are active in the whole time courses, so the estimated interference energy $E_2(u_i)$ is not reliable. Therefore we calculate the SIRs only for the first channels where the speech signals f1 and m1 are primary sources. To show the performance of the proposed method on the different sets of mixtures, we split the SIR results of *speech + music* mixtures and *speech + speech* mixtures and summarized separately.

By applying the FD-BSS, there were about 3.6 dB average SIR improvements for *speech + music* mixtures and 4.9 dB for *speech + speech* mixtures, respectively. With the help of spectral subtraction, there were observed 2.3 dB and 2.1 dB more improvements. However, when the proposed ACCIC is performed, there were additional 3.9 dB and 4.9 dB average SIR improvements, which are 6.2 dB and 7.0 dB improvements from FD-BSS results. Our results show that ACCIC improves the performance of *speech + speech* separation more than that of *speech + music* separation. This is because speech signals are sparser than music signals, and therefore interferences occur less frequently in speech mixtures, which matches well to our hypothesis model in (5).

Figure 4 plots the results of the proposed method for f1-g2 mixture recordings (*speech + music*), and Figure 5 plots the results for f1-m2 mixture (*speech + speech*). The waveform in the left of the first row is observed by microphone 1 in Figure 3, and the waveform in the right of the first row is by microphone 2. The microphone observations are inputted to FD-BSS and the second row waveforms are produced. The major source of the left of the second row is speech, since the left microphone input is already speech-major. Similarly the major source of the right of the second row is music. The second row signals are inputted to both of the spectral subtraction and the proposed method. The source presence probabilities are computed and represented in the third row. Based on the computed probabilities, the post-processing methods remove the leftover cross-channel interference signals. The two waveforms in the fourth row are the noise-reduced results of spectral subtraction, and the waveforms in the fifth row are the final denoised results of the proposed method. When the primary source is not detected (source presence probability is close to 0), both successfully cancel out the background source component. However when the primary source is detected, ACCIC shows better separability than spectral subtraction. By listening to the left channel of Figure 4, from 10 sec to 15 sec, it is observed that ACCIC significantly removed the music sound from the speech signal that is left in the FD-BSS outputs, however spectral subtraction could rarely reduce the interfering music sound. Besides, although music is the primary source at the right channel of the FD-BSS outputs, music signal of small amplitude (from 4 sec to 8 sec) disappeared since that duration was identified as "background" and suppressed. Because spectral subtraction usually uses slowly-varying noise estimates, it is good for the reduction of stationary sources such vehicle noise, but non-stationary sources such as music or speech signals are difficult to handle.

Table 1. Measured SIRs of the input signals, FD-BSS outputs, noise-reduced results by spectral subtraction, and the interference-canceled results by the proposed method. (a) 'speech + music' mixture separation results. (b) 'speech + speech' mixture separation results. 'mixture' columns indicate the type of sources mixed into the stereo input. 'f1' and 'f2' are female speakers, 'm1' and 'm2' are male speakers, and {'g1', 'g2', 'g3'} are three different music signals described in Table 1. 'Average' row lists the average SIRs of each method, and 'Increase' row lists the improvements from the preceding columns. The parenthesized values in the last column of the 'Increase' row are improvements of ACCIC over FD-BSS outputs. All the scalar values are in dB

(a) speech + music mixture separation

Mixture	Input	BSS only	BSS + Spectral Subtraction	BSS + proposed ACCIC
f1-g1	11.8	13.7	15.5	17.3
f1-g2	2.2	6.3	8.6	11.9
f1-g3	5.6	10.1	11.4	17.6
m1-g1	9.7	12.3	15.6	18.2
m1-g2	1.6	5.4	8.4	12.0
m1-g3	4.6	9.4	11.5	17.2
Average	5.9	9.5	11.8	15.7
Increase		+3.6	+2.3	+3.9 (+6.2)

(b) speech + speech mixture separation

Mixture	Input	BSS only	BSS + Spectral Subtraction	BSS + proposed ACCIC
f1-f2	4.7	8.4	9.6	15.4
f1-m2	7.6	11.3	13.8	18.0
m1-f2	2.9	10.1	11.9	17.5
m1-m2	7.5	12.5	15.2	19.5
Average	5.7	10.6	12.6	17.6
Increase		+4.9	+2.1	+4.9 (+7.0)

Our proposed method employs the information from the other channel and rapidly-varying cross-channel interfering sources are better suppressed with the help of the cross-channel cancellation factors (9).

The proposed ACCIC is most suitable to sparse sources such as speech signals. Since the probability density models described in (13) are based on the variance only, the density functions of sparse sources ---peaked at 0 and having rapidly changing amplitudes over time--- can be better modeled than the dense sources such as music signals. Since the source absence/presence probabilities computed by

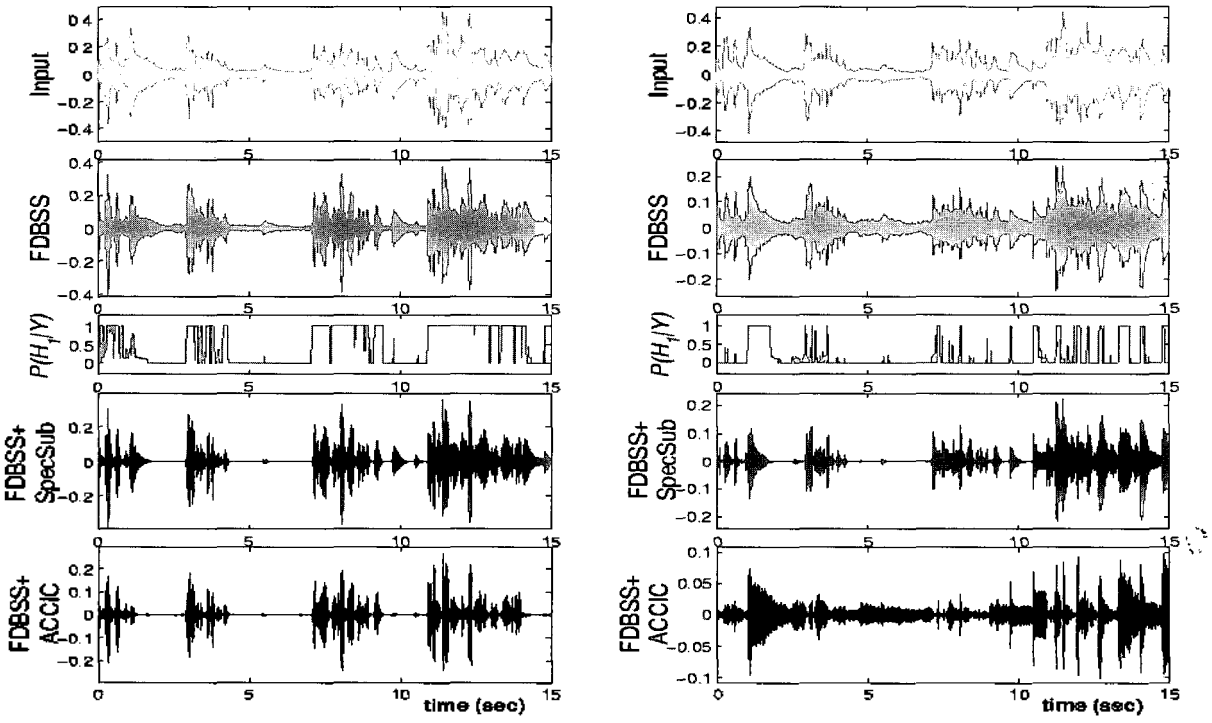


Fig. 0. Separation result of the developed method for female-music mixture (f1-g2). First row: waveform views of the microphone inputs (recorded signals). Second row: output of FD-BSS algorithm. Third row: source presence probabilities computed by (8). The probability values range from 0 to 1. Fourth row: output of spectral subtraction algorithm, on the FD-BSS outputs, using the computed source presence probabilities. Fifth row: output of the proposed adaptive cross-channel interference cancellation (ACCIC) algorithm. The measured SIRs are 5.6 dB (microphone recordings), 10.1 dB (FD-BSS), and 11.4 dB (FD-BSS + spectral subtraction), and 17.6 dB (FD-BSS + ACCIC). Wave files for all the data are available at http://myhome.naver.com/flyers

the model significantly affects the accuracy of estimating interference-canceling factors, the SIR improvements were greater in *speech + speech* mixtures than in *speech + music* mixtures.

4.4. Experiments with Beamformer Outputs

Our proposed method is able to be coupled with any kind of multi-channel source separation method, provided it produces at least 'emphasized' primary source and 'deemphasized' secondary sources. Figure 6 shows an example of applying ACCIC on the outputs of a linearly constrained minimum variance beamformer (LCMVBF) [15] instead of FD-BSS. The waveforms in the first row are beamformer outputs. Three speakers are talking with some overlaps and background music is played. This situation is that three source signals plus a noise are mixed and observed by three sensors, that is, in (1), $s_i(t)$'s are the speech sounds and $n_j(t)$ is the background music spread over all the sensors. The observed signals are then passed through a LCMVBF and outputs are produced. However, there still remain secondary speakers' speeches and the background music in an intelligible amount, although the amplitudes look insignificant compared to the target speaker's speech.

On the LCMVBF outputs, we perform spectral subtraction and ACCIC and displayed in the third and fourth rows of Figure 6. By listening to the ACCIC results, non-target speakers' speeches are almost removed and the background music is seldom audible. These results show that our method is able to cancel out common noise components as well, when the magnitude of the noise is a lot smaller than the primary source. This kind of noise frequently happens to be observed in the practical situations, and our method is more useful when the number of noise sources is not given. Although spectral subtraction reduces the uninterested speeches and background music to some extent, it produces a lot more musical noises and the overall quality is far poorer than ACCIC. When multiple observation channels are available, ACCIC is expected to show a lot of benefit such as the better noise reduction quality and less distortion in the target source.

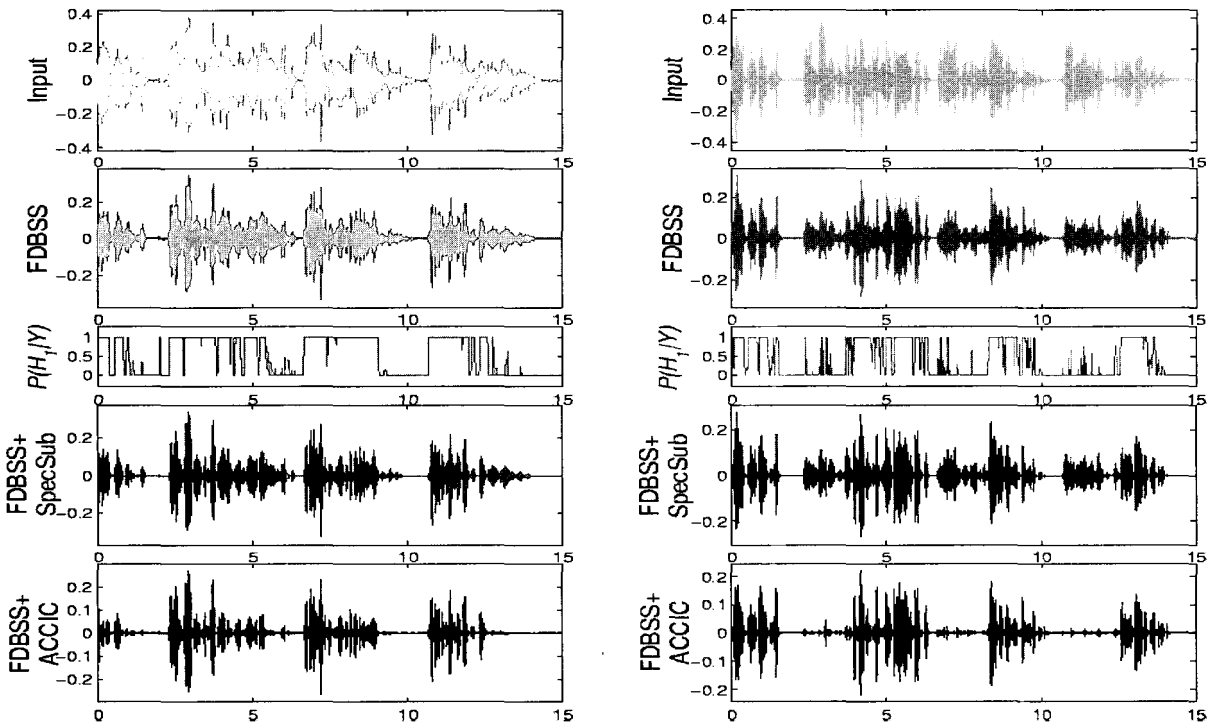


Fig. 0. Separation results of the proposed method for female-male mixture (f1-m2). The measured SIRs are 7.6 dB (microphone recordings), 11.3 dB (FD-BSS), 13.8 dB (FD-BSS + spectral subtraction), and 18.0 dB (FD-BSS ACCIC). Although more residual signals at the final results are observed than in the case of f1-g2, the listening quality and measured SIR are almost same as f1-g2 result. Wave files for all the data are available at http://myhome.naver.com/flyers.

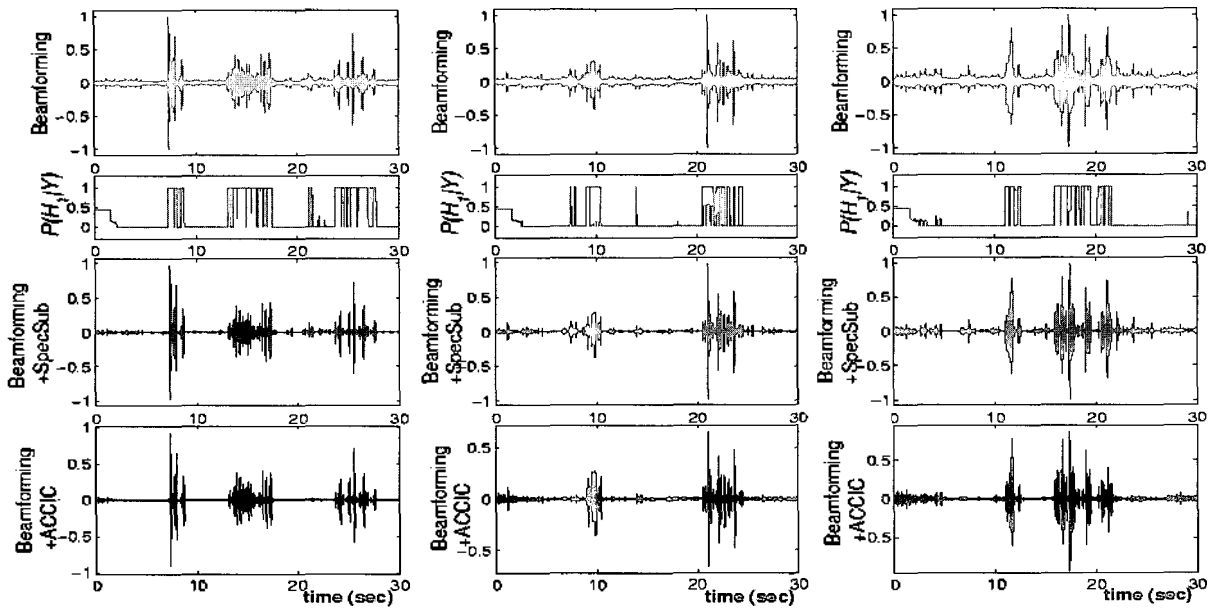


Fig. 0. Results of the proposed method for beamformer outputs. Three speakers are talking and a song is played in the background. First row: waveform views of beamformer outputs. Second row: source presence probabilities computed by (8). Third row: output of spectral subtraction algorithm, on the beamformer outputs, using the computed source presence probabilities. Fourth row: output of the proposed ACCIC algorithm. After running ACCIC, only a target speaker's speech is emphasized and the other 2 speakers' speeches and the background music are suppressed in all of the 3 channels. The measured SIRs are 13.8 dB (beamformer outputs), 20.1 dB (beamforming + spectral subtraction), and 22.8 dB (beamforming + ACCIC). Wave files for all the data are available at http://myhome.naver.com/flyers.

V. Conclusion

The ordinary BSS algorithms have inherent separation errors due to the mismatch between the assumed linear model and the real transfer functions. We proposed a post-processing technique that is applicable to such realistic environments. It has been a similar effort to compensate the separation errors for stationary noise sources [14]. In the proposed method, we deal with nonstationary natural noise sounds on the assumption that the number of sources is equal to the number of sensors, and each of the blind source separation system outputs has a primary source and a secondary source signal identified by their relative power. The proposed algorithm considers one BSS output as noisy signal and the other output as reference noise source, and the cancellation is done in the power spectral domain as the conventional spectral subtraction methods do. The advantage of the power spectral subtraction is that it effectively absorbs the small amount of mismatch between the actual filter and the estimated one, and generates cleanly denoised signals. The disadvantage is the introduction of the musical noises due to the half-wave rectification on the subtracted outputs. However, by comparing the separation results of the real recordings with the denoised results of conventional spectral subtraction, we showed that our method is more successfully removes non-stationary interfering sources as well as background common noise. The introduction of the interference canceling factors enables the extraction of the interfering source characteristics in the other channels, and helps accurate estimation of the interfering sources. The potential application areas would be noise reduction for automatic speech recognition especially in a vehicle, separation of the speakers in a mixed conversation, and reduction of background nonstationary noise in a distant voice communication.

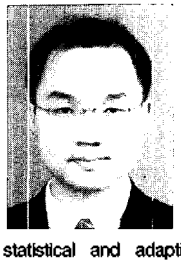
References

1. K. Torkkola, "Blind signal separation for audio signals - are we there yet?," in Proc. ICA99, (Aussois, France), pp.261-266, January 1999.
2. S. Araki, S. Makino, R. Aichner, T. Nishikawa, and H. Saruwatari, "Subband based blind source separation with appropriate processing

for each frequency band," in Proc. ICA2003, (Nara, Japan), pp.499-504, April 2003.
3. B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong, and R. C. Goodlin, "Adaptive noise cancelling: principles and applications," Proceedings of the IEEE, vol.63, pp. 1692-1716, December 1975.
4. S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acous., Speech and Signal Processing, ASSP, vol.27, no. 2, pp.113-120, 1979.
5. L. Parra and C. Spence, "Convolutive blind separation of non-stationary sources," IEEE Trans. Speech and Audio Processing, vol.8, pp.320-327, May 2000.
6. S. Choi, S. Amari, A. Cichocki, and R. wen LIU, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," in Proc. ICA99, (Aussois, France), pp.371-376, January 1999.
7. H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency-domain blind source separation," in Proc. ICASSP, (Orlando, Florida), May 2002.
8. T.-W. Lee, A. J. Bell, and R. Orglmeister, "Blind source separation of real world signals," in Proc. ICNN, (Houston, USA), pp.2129-2135, June 1997.
9. N. S. Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," IEEE Signal Processing Letters, vol.7, pp.108-110, May 2000.
10. M.R. Weiss and E. Aschkenasy, "Computerized audio processor," Final Report, Rome Air Development Center RADC-TR-83-109, May 1983.
11. M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by additive noise", In Proc. ICASSP'79, pp. 208-11, 1979.
12. G.-J. Jang, T.-W. Lee, and Y.-H. Oh, "Learning statistically efficient features for speaker recognition," in Proc. ICASSP, (Salt Lake City, Utah), May 2001.
13. A. J. Bell and T. J. Sejnowski, "Learning the higher-order structures of a natural sound," Network: Computation in Neural Systems, vol.7, pp.261-266, July 1996.
14. E. Visser, M. Otsuka, and T.-W. Lee, "A spatio-temporal speech enhancement scheme for robust speech recognition in noisy environments," Speech Communications, vol.41, pp.393-407, 2003.
15. C. Choi, D. Kong, S. M. Yoon, and H.-K. Lee, "Separation of multiple concurrent speeches using audio-visual speaker localization and minimum variance beamforming," in Proc. ICSLP, (Jeju, Korea), October 4-8, 2004.

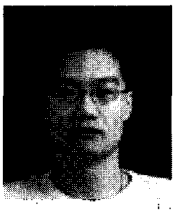
[Profile]

● Gil-Jin Jang



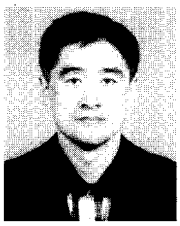
Gil-Jin Jang received his B.S., M.S., and Ph.D. degree in electrical engineering and computer science from the Korea Advanced Institute of Science and Technology (KAIST), in 1997, 1999, and 2004 respectively. He was a visiting research scholar at the University of California, San Diego, from September 2000 to February 2001. He is now a senior research engineer with the Computing Lab, Samsung Advanced Institute of Technology. His research interests include statistical and adaptive signal processing, speech enhancement, speech recognition, speech coding, and independent component analysis.

● Changkyu Choi



Changkyu Choi received his B.S., M.S., and Ph.D. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), in 1991, 1994, and 1999 respectively. He was a visiting researcher at the Mechanical Engineering Lab, Japan from January 1999 to February 1999, and a lecturer at Chungnam National University, Korea from March 1999 to August 1999. He was awarded best student paper from the Institute of Control, Automation, and Systems Engi-neers (ICASE), Korea in 1999. He is now a senior research engineer at the Interaction Lab, Samsung Advanced Institute of Technology (SAIT). His research areas are speaker localization and tracking using microphone array and active vision, sound source localization and beam-forming using microphone array signal, multi-channel speech enhancement and noise reduction, independent component analysis for face detection, and blind signal separation for speech and audio signals.

● Yongbeom Lee



Changkyu Choi received his B.S., M.S., and Ph.D. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), in 1991, 1994, and 1999 respectively. He was a visiting researcher at the Mechanical Engineering Lab, Japan from January 1999 to February 1999, and a lecturer at Chungnam National University, Korea from March 1999 to August 1999. He was awarded best student paper from the Institute of Control, Automation, and Systems Engineers (ICASE), Korea in 1999. He is now a senior research engineer at the Interaction Lab, Samsung Advanced Institute of Technology (SAIT). His research areas are speaker localization and tracking using microphone array and active vision, sound source localization and beam-forming using microphone array signal, multi-channel speech enhancement and noise reduction, independent component analysis for face detection, and blind signal separation for speech and audio signals.

● Jeongsu Kim

degree from the Korea Advanced Institute of Science and Technology (KAIST), in 1988, and 1990 respectively. From March 1990 to January 1993, he was affiliated with the Information and Communication Research Institute, Samsung Electronics Co., Ltd. He is now a senior research engineer with the Computing Lab, Samsung Advanced Institute of Technology. His research areas are speech recognition, speech synthesis, and dialog management. Jeongsu Kim received his B.S. from Yonsei University, Korea and his M.S.

● Sangryong Kim

Sangryong Kim recieved his B.S in electronic engineering from Hankuk Aviation University, Korea, M.S and Ph.D. degree in electronic engineering from the Korea Advanced Institute of Science and Technology (KAIST), in 1980, 1982, and 1989 respectively. He was affiliated with Samsung Electronics, Co., Ltd., as a technical research staff from 1989 to 1993. He was a director from 1993 to 1997 and Vice President from 1997 to 2004 with Human and Computer Interaction (HCI) Lab of Samsung Advanced Institute of Technology (SAIT). He was awarded Samsung Technical Excellence in 1994. He is now Vice President of the Interaction Lab at SAIT. His research interests are voice recognition for navigators, voice codec for ITU, automatic translation to Korean from English, voice recognition dialing systems, and computing system for blindly handicapped.