International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:7, 2007

On-line Speech Enhancement by Time-Frequency Masking under Prior Knowledge of Source Location

Min Ah Kang, Sangbae Jeong, and Minsoo Hahn

***Abstract*—**This paper presents the source extraction system which can extract only target signals with constraints on source localization in on-line systems. The proposed system is a kind of methods for enhancing a target signal and suppressing other interference signals. But, the performance of proposed system is superior to any other methods and the extraction of target source is comparatively complete. The method has a beamforming concept and uses an improved time-frequency (TF) mask-based BSS algorithm to separate a target signal from multiple noise sources. The target sources are assumed to be in front and test data was recorded in a reverberant room. The experimental results of the proposed method was evaluated by the PESQ score of real-recording sentences and showed a noticeable speech enhancement.

***Keywords*—**Beamforming, Non-stationary noise reduction, Source separation, TF mask.

I. Introduction

SPEECH enhancement algorithms are used to increase the recognition rates and improve the quality of services. They are concerned with the processing of noisy speech to improve the quality of the signal.

Stationary noise reduction has been achieved through various techniques including Wiener and Kalman filtering. But non-stationary noises including speech or music signals cannot be effectively reduced. Microphone array-based speech enhancement has been recently spotlighted because of its wide usability such as high-quality hands-free telecommunication systems and its powerfulness in reducing non-stationary noises. In a sense, we can think of the microphone array-based algorithms as a kind of target source extraction systems.

The conventional source extraction system is achieved by the generalized sidelobe canceller (GSC). It is considered the most practical algorithm because of its simplicity and effectiveness [1]-[3]. But it has some insufficiency in noise reduction. Also, the blind source separation (BSS) can be used for the target source extraction. In the BSS, the separated signal is one of the sources. The TF mask-based BSS has better performances for the under-determined condition than other methods including independent component analysis (ICA)-based ones. Also, TF mask-based techniques can run on-line easily in contrast to ICA-based methods which have long time delay when they are operated on-line. In conventional TF masking methods, the mask is made by clustering the inter-channel gain ratio (IGR) and the inter-channel time delay (ITD) [4]. But this approach is impractical because the clustering cannot be successful in real environments and has the problem in determining which of the separated sources the target source is. Also, the conventional techniques using TF masks have been applied to extract the target source from mixed speech signals because speech signals are almost W-DO (Windowed-Disjoint Orthogonal) in the TF domain [4]. They have poor performances in music noise environments. But, the proposed method makes up for these defects and is efficient to extract the frontal target source completely in various noisy environments.

In this thesis, we propose the TF mask extracting completely a target source by applying prior knowledge of its location to the 2-dimensional distribution defined by the IGR and the ITD. Because we assume the look direction, that is, the angle between the source and the front of the microphone array is zero, our goal is to extract only the frontal signal in multiple noise environments. The main idea, that is, focusing on the location of the target source is similar to beamforming. So, the proposed system has a beamforming concept but uses the TF mask rather than adaptive filters to separate the target signal in multiple noise environments.

The organization is as follows. Section 2 explains the conventional source extraction algorithms. Section 3 presents the proposed algorithm. Experimental results and conclusions are given in Section 4 and Section 5, respectively.

II. Conventional Algorithms

The conventional source extraction algorithms have the two approaches. Firstly, the beamforming algorithm has been used in non-stationary noise reduction. The purpose of the beamformer is to minimize the effects of noise at the array output using the prescribed frequency response in the direction of the target signal. The GSC is considered the effective beamforming algorithm because of its simplicity and capacity of noise reduction. Secondly, the BSS algorithm can be used to separate the target signal from non-stationary noise sources.

Manuscript received August 30, 2007.

The authors are with the School of Engineering, Information and Communications University, Daejeon, Republic of Korea (e-mails: dsiqueen@icu.ac.kr, sangbae@icu.ac.kr, mshahn@icu.ac.kr).

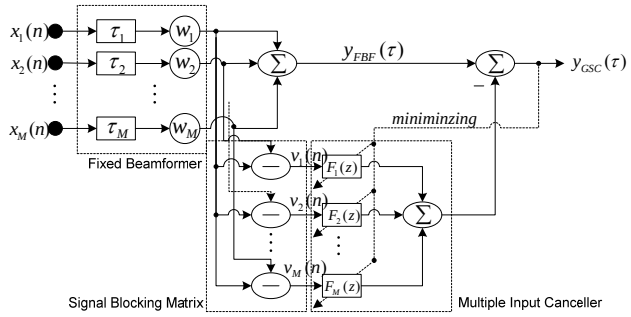

Fig. 1 GSC structure

A. Generalized Sidelobe Canceller

The GSC is constrained to extract the target signal with a filter having a predetermined phase response and gain. The target signal is identified by time delay and weight steering the microphone outputs. Its direction is pointed by the time delay elements $\tau_1, \tau_2, \cdots, \tau_M$ and weights $w_1, w_2, \cdots, w_M$.

The GSC structure is shown in Fig. 1. This processor consists of three distinct substructures which are depicted as the dotted blocks. Using a set of fixed weights $w_1, w_2, \cdots, w_M$, the fixed beamformer produces non adaptive beamformed signal $y_{FBF}(n)$,

$$y_{FBF}(n) = \mathbf{w}^T x(n) \tag{1}$$

where $\mathbf{w}^T = [w_1, w_2, \cdots, w_M]$.

The adaptation process consists of the two processing blocks, the signal blocking matrix and the multiple input canceller. They adjust the beamforming coefficients to new values as each array sensors received new samples. In the signal blocking matrix, the each fixed beamformer output is subtracted by other channel outputs. Then, the in-phased target signals in the fixed beamformer outputs are cancelled and leave only interference signals. The multiple input canceller reduces the interference signals in the fixed beamformer outputs through the adaptation process with the signal blocking matrix outputs. The least mean square (LMS) algorithm is used to adapt the filter coefficients. The LMS adaptation is processed by (3).

$$\begin{aligned} J(n) &= E\left[y^2_{GSC}(n) \right] \\ &= E\left[\left(y_{FBF}(n) - \sum_{m=0}^{L-1} f^{(n)}(m) v'(n-m) \right)^2 \right] \end{aligned} \tag{2}$$

$$\begin{aligned} f^{(n+1)}(m) &= f^{(n)}(m) - \mu \cdot \frac{\partial J(n)}{\partial f^{(n)}(m)} \\ &= f^{(n)}(m) - \mu \cdot y_{GSC}(n) v'(n-m) \end{aligned} \tag{3}$$

B. Blind Source Separation

The BSS is applied to estimate original source signals using only the information of the mixed signals observed in each input channel. It has broadly two approaches: the ICA-based BSS and the TF mask-based BSS.

1) ICA-based BSS

Assuming that the original signals are independent we can apply an ICA algorithm to reconstruct the unknown sources. Bell and Sejnowski proposed the information maximization approach (infomax) as the ICA algorithm [9].

The infomax principle separates many independent sources using the separating matrix W. The separating coefficients are learned by minimizing the mutual information between components of $y(t) = g(u(t))$, where g is a nonlinear function approximating the cumulative density function of the sources. Minimizing the mutual information between components of y is equal to maximizing the entropy of y. The stochastic gradient ascent algorithm [9] adapts the separating coefficients to maximize the entropy of the output.

But, to operate the ICA-based BSS on-line, usually long buffering of signals are required. This causes inevitable long time delay.

2) TF mask-based BSS

The TF mask-based BSS is another method used to separate sources from a mixture. It can run on-line easily and has the better performance for under-determined condition than the ICA-based BSS. When the source signals do not overlap in the TF domain, high-quality reconstruction can be obtained. So, the conventional TF mask is constructed under the assumption that the speech sources are W-DO [4]. However, when there are overlaps between the sources like music sources, it has poor performance.

When the sources are W-DO, at most one source will be active at any TF point (τ, ω). Equation (4) shows that we can demix an arbitrary number of sources from only one of the mixtures if we can construct the corresponding mask, M_j, for each source.

$$s_j = M_j x_1 \tag{4}$$

The masks, M_j, are constructed from the IGR and ITD of the mixtures, $x_1(t)$ and $x_2(t)$. Then, its IGR and ITD are (5) and (6).

$$g(\tau, k) = \left| \frac{X_2(\tau, k)}{X_1(\tau, k)} \right| \tag{5}$$

$$\delta(\tau, k) = -\frac{N}{2\pi k} \angle \frac{X_2(\tau, k)}{X_1(\tau, k)} \tag{6}$$

Where $X_1(\tau, k) = FFT\left[x_1(\tau)\right]$, $X_2(\tau, k) = FFT\left[x_2(\tau)\right]$, τ is the time and N is the FFT size.

When $\left(g(\tau, k), \delta(\tau, k)\right)$ is obtained at each non-zero TF point, the TF points, (τ, ω), with the same label are grouped by the clustering. But, the clustering cannot be successful in real situation. So, the conventional TF-masked BSS is impractical so that it should be compensated.

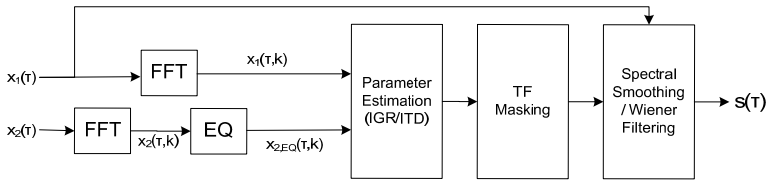

Fig. 2 Block diagram of proposed method.
($x_1(\tau)$: mixed signal from channel 1, $x_2(\tau)$: mixed signal from channel 2, EQ: Equalizer, $s(\tau)$: extracted target source)

III. Proposed Source Extraction System

Entire organization of our target source extraction system is shown in Fig. 2. The proposed method is performed in the frequency domain because it is easy to analyze the location information for each frequency component. Therefore, for the stereo signals, $x_1(\tau)$ and $x_2(\tau)$, which are noisy, they are converted into the frequency domain by the Fast Fourier Transform (FFT). Then, the mask for the target source extraction has to be constructed. In order to make the mask, the ITD and the IGR are required. The equalizer helps the parameters to be estimated exactly independent of external factors. After the construction of the mask, it is applied to the spectrum of one noisy signal, $x_1(\tau,k)$ or $x_2(\tau,k)$. Because the TF masking technique can separate the spectrum of noisy signal into that of noise signal and that of the target speech signal, noises can be reduced by the Wiener filter. Therefore, finally, the proposed system extracts only target speech source using Wiener filter. The following subsections describe each step of the proposed method in detail.

A. Microphone Equalizer Design

The gain coefficients of each microphone can be different and the frequency characteristics of each input channel can also be different. If so, the key parameters for mask construction, the IGR and the ITD, won't be estimated exactly. Because the frontal signal is inputted with same gain at the same time through two microphones, the IGR and the ITD should be one and zero, respectively. But the different gain coefficients of microphones and frequency characteristics make the IGR and the ITD of the frontal signal measured wrongly. The equalizer is required to be unaffected by these external factors. The equalizer makes two input channels have same gain and no time delay. It is designed by minimizing the cost function, J_k, given in (7). a_k is the coefficient which compensates magnitude and phase differences between input signals for k^{th} spectral bin.

$$J_k = \sum_{\tau=0}^{T} \left| X_1(\tau,k) - \alpha_k X_2(\tau,k) \right|^2 \tag{7}$$

where τ is the time , T is the number of total speech frames, k is the index of spectral bin and $X_1(\tau,k) = FFT\left[x_1(\tau)\right]$, $X_2(\tau,k) = FFT\left[x_2(\tau)\right]$.

Since a_k is the complex number, we obtain a_k satisfying (8).

$$\frac{\partial J_k}{\partial \alpha_k^*} = 0 \tag{8}$$

$$\alpha_k = \frac{\sum_{\tau=0}^{T} X_1(\tau,k) X_2^*(\tau,k)}{\sum_{\tau=0}^{T} \left| X_2(\tau,k) \right|^2} \tag{9}$$

The coefficient a_k obtained in (9) is estimated in the training mode, that is, noiseless condition and is applied to one input channel in the test mode.

$$X_{2,EQ}(\tau,k) = \alpha_k X_2(\tau,k) \tag{10}$$

In (10), the second input channel, $X_{2,EQ}(\tau,k)$, compensated with the coefficient a_k have the same magnitude and no time delay with $X_1(\tau,\omega)$. Strictly speaking, $X_1(\tau,\omega)$ and $X_{2,EQ}(\tau,k)$ become inputs to estimate parameters in following section.

B. Parameter Estimation

The proposed algorithm uses the IGR and the ITD to construct the spectral mask for the extraction of the target signal. Then the IGR and the ITD are calculated by (5) and (6), respectively.

C. TF Mask Construction

Using the IGR and the ITD, we show how to construct the mask corresponding to the target source in this sub-section. The IGR and the ITD, $\left(g(\tau,k),\delta(\tau,k)\right)$ are distributed pair-wisely in the 2-dimensional coordinate. In case a frontal signal, inter-channel gain ratio is one and inter-channel time delay is zero. Therefore, under the assumption that the target source is located in front, the pair $\left(g(\tau,k),\delta(\tau,k)\right)$ of the target signal is distributed in the near (1,0) of the 2-dimensional coordinate. Then, the mask which selects only components within the circle is constructed with two approaches. One is to make the hard-decision mask and another is to make the soft-decision mask for better performance.

1) Hard-decision Mask

The hard-decision approach for mask construction is to separate each spectral bin into noise bin or source bin using the binary mask. The target signal components are selected within a circle centered at (1,0). $M_H(\tau,k)$ is the binary mask to extract the target signal at time τ for the k^{th} spectral bin. It is

found by

$$M_H(\tau,k)=\begin{cases}1 & if\ (g(\tau,k)-1)^2+\delta(\tau,k)^2<r^2\\ 0 & otherwise\end{cases} \quad (17)$$

where r is the pre-determined threshold that specifies if a spectral bin is from the target signal or not.

2) Soft-decision Mask

In the soft-decision approach, we assume that the frequency bins which have both components of noise and target speech exist. So, the soft-decision mask considering these bins is constructed instead of the binary mask. From the distribution of the IGR and ITD, we can know that the compartment of noise spectral bins and source spectral bins are more dependent on ITD than IGR. Then, the bins corresponding to the ITD which are less than the critical value are classified as source bins and the rest bins are classified as noise bins. But we assume that the bins which have both components of noise and source are located within an oval centered at (1,0). So, the target signal components located within the oval are selected with a proper proportion, η, between zero and one. The proportion and the critical value of ITD are determined by experiments in training mode. $M_S(\tau,k)$ is the soft-decision mask to extract the target signal at time τ for the k^{th} spectral bin. It is found by (18).

$$M_S(\tau,k)=\begin{cases}\eta & if\ \dfrac{(g(\tau,k)-1)^2}{a^2}+\dfrac{\delta(\tau,k)^2}{b^2}<1\\ 1 & if\ \dfrac{(g(\tau,k)-1)^2}{a^2}+\dfrac{\delta(\tau,k)^2}{b^2}>1,\ |\delta(\tau,k)|<\phi\\ 0 & otherwise\end{cases} \quad (18)$$

where a and b are the pre-determined thresholds that classifies the bins which have both components of noise and source . ϕ is the critical value of ITD that specifies if a spectral bin is from the target signal or not.

Before constructing the soft-decision mask, firstly, the standard deviation of ITD is calculated each bin from the training mode. If it is above pre-determined threshold, β, source signals are negligible in the bins corresponding on it. Therefore, we apply the zero mask, $M_S(\tau,k)=0$, to these bins. But if the standard deviation of ITD is below the threshold, the soft-decision mask as (18) is applied.

D. Wiener Filter based Noise Reduction

After the mask for the extraction of the target source is constructed, it is applied to a noisy signal of the two input channels. Then, the TF masking procedure is performed by

$$S(\tau,k)=M(\tau,k)X_1(\tau,k) \quad (19)$$

where $M(\tau,k)$ is the mask value at time τ for the k^{th} spectral bin.

The only frequency bins of target source are extracted by the spectral masking and we obtain the power spectrum of target source through the procedure. Also, we can obtain the power spectrum of noise source from unextracted bins.

If the spectrum of noise and the spectrum of target source are separated from ones of noisy signal, noises can be reduced by Wiener filtering. Therefore we obtain the impulse response for the Wiener filter using the two spectrum separated by TF masking procedure. The impulse response is found by (20).

$$h(\tau)=IDFT\left[\frac{SNR(k)}{1+SNR(k)}\right] \quad (20)$$

Then, the filter response is convoluted to the noisy signals as (21) and we obtain the extracted target speech signal.

$$s(\tau)=x_1(\tau)*h(\tau) \quad (21)$$

But the separated spectrum has the undefined frequency bins. The undefined frequency bins cause the problem in calculating SNR. So we solve it by the spectral smoothing technique which convolutes the frequency response of hamming window.

IV. Experiments and Results

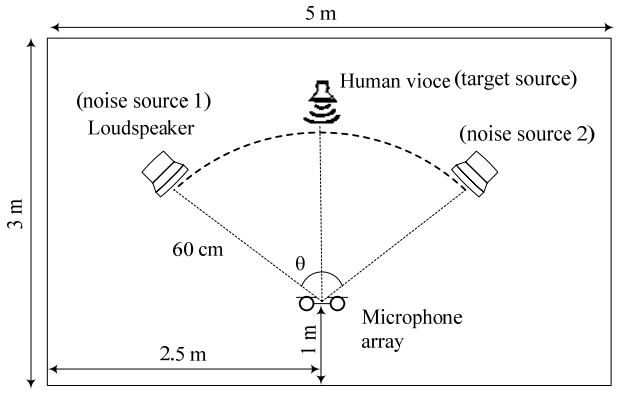

Fig. 3 Layout of experiments

A. Conditions

Experiments were conducted with 30 sentences. They were uttered by different speakers. Two microphones with the array aperture of 2 cm were applied. The target source was located in the distance of 60 cm in front of the microphone array. Two noise sources were located in both sides centering around the target source and speech or music signals were played with loudspeakers. The angles between two noise sources were 30°, 60°, 90°, 120°, 150°, 180°, respectively. The distance between the microphone array and the noises was also 60 cm. The 2-channel noisy signals were recorded in the $5m\times3m$ office environment as shown in Fig. 3. The pre-emphasis was applied to the input signals and the FFT was performed every 10 msec throughout the experiments. The sampling frequency and the FFT size were 16 kHz and 512. Also, the frame length was the same with the FFT size for no zero padding. The SNR is varied from 0 dB to 20 dB to evaluate the performance.

The experimental results was evaluated by the PESQ (Perceptual Evaluation of Speech Quality) score [5]. The higher PESQ score indicates the better quality of the extracted target speech. Also, the performance of proposed algorithm was compared with that of GSC algorithm in [10].

B. Results

1) Hard-decision results

In speech noise environments, compared with the PESQ score of original noisy signals, that of the hard-decision mask based system is increased by average 0.48 in SNR 0 dB, 0.4 in SNR 5 dB, 0.3 in SNR 10 dB, 0.25 in SNR 15 dB, 0.15 in SNR 20 dB, respectively. In music noise environments, it is increased by average 0.4 in SNR 0 dB, 0.38 in SNR 5 dB, 0.35 in SNR 10 dB, 0.3 in SNR 15 dB, 0.2 in SNR 20 dB.

2) Soft-decision results

The soft-decision mask based system is proposed for the better performance than the hard-decision mask based system.

In speech noise environments, the PESQ score is increased by average 0.53 in SNR 0 dB, 0.46 in SNR 5 dB, 0.37 in SNR 10 dB, 0.28 in SNR 15 dB, 0.17 in SNR 20 dB. In music noise environments, the result is increased by average 0.43 in SNR 0 dB, 0.4 in SNR 5 dB, 0.37 in SNR 10 dB, 0.3 in SNR 15 dB, 0.2 in SNR 20 dB.

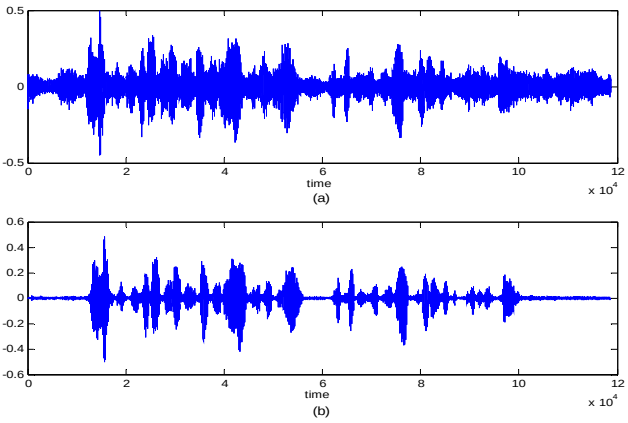

Fig. 4 (a) waveform of original noisy signal, (b) waveform of extracted target speech signal

3) Results Analysis

These increases in PESQ scores mean more reduced noise in human hearing. Fig.4 shows the waveforms of the original noisy signal and the target speech signal extracted by soft-decision mask. From the comparison of them, we can confirm the remarkable noise reduction.

C. Performance Comparison of Proposed Work vs. GSC

When extracting the target source from noisy signal, the GSC has been considered the most feasible algorithm. But it degrades the speech quality of target source in high SNR according to the increase of learning rate. We proposed the systems which have better performance than the GSC. The soft-decision based system with the equalizer is the best in the proposed systems and shows noticeable improvements. Fig. 5 shows the performance improvement comparison of the GSC and the proposed system in speech noise environments. The PESQ score increased as average 0.36. It is 6 times higher than that of GSC as 0.06. Fig. 6 shows the results in music noise environments. The PESQ score of the proposed system increased as average 0.4. It is 4 times higher than that of GSC as 0.1.

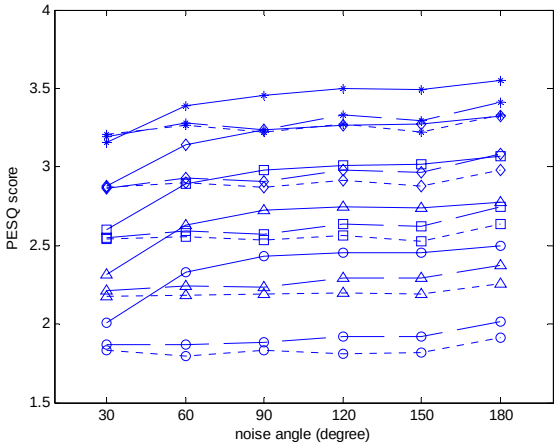

Fig. 5 Performance improvement of GSC and proposed system in speech noise environments (dotted line: original noisy, dashed line: GSC results, solid line: proposed system results, '○': SNR 0 dB, '△': SNR 5 dB, '□': SNR 10 dB, '◊': SNR 15 dB, '✳': SNR 20 dB)

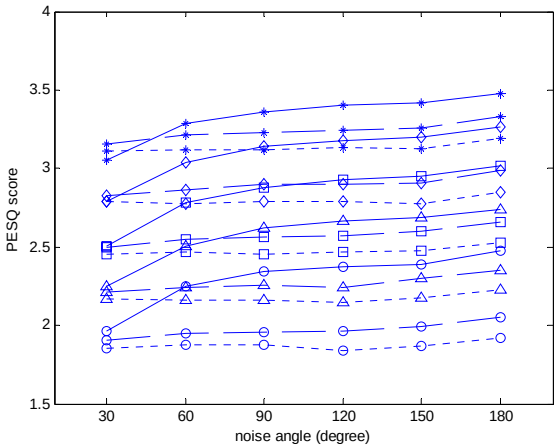

Fig. 6 Performance improvement of GSC and proposed system in music noise environments (dotted line: original noisy, dashed line: GSC results, solid line: proposed system results, '○': SNR 0 dB, '△': SNR 5 dB, '□': SNR 10 dB, '◊': SNR 15 dB, '✳': SNR 20 dB)

V. Conclusion

This paper proposed an improved TF masking technique to extract only the frontal target source corrupted by noise signals. The proposed system can run on-line and is robust in multiple noise environments although we have less microphones than sources. In order to evaluate its superiority, we have tested the system in a live situation where loudspeakers made speech sounds in a real reverberant room. From the results, it shows a remarkable rise of PESQ score.

Our system can be applicable to various real field applications as a preprocessor when non-stationary noise reduction is necessary. Moreover, it has an advantage in on-line implementation because of the complexity decrease. So, it can be implemented suitably in mobile cellular phones, navigation

systems, hands-frees, toys with speech recognition ability, etc. Also, it is useful to various speech coders.

As further works, we expect to apply the proposed algorithm to speech codecs and speech recognition systems. Also, we will add the adaptive process of equalizer for better performance.

REFERENCES

[1] M. Brandstein and D. Ward, *Microphone Arrays*, Springer, 2001.

[2] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 1991.

[3] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Process.,* vol.49, no.8, Aug. 2001, pp.1614-1626.

[4] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.,* vol. 52, no. 7, July 2004, pp.1830-1846.

[5] ITU-T, "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," ITU-T Recommendation P.862, February 2001.

[6] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of dominant target sources using ICA and time-frequency masking," *IEEE Trans. Signal Process. ,* vol. 14, no. 6, Nov. 2006, pp.2165-2173.

[7] H. Saruwatari, S. Kurita, and K. Takeda, "Blind source separation combining frequency-domain ICA and beamforming," *in Proc. ICASSP2001,* pp.2733-2736.

[8] G. Shi and P. Aarabi, "Robust digit recognition using phase-dependent time-frequency masking," *in Proceedings of ICASSP,* Hong Kong, Apr. 2003, pp.684-687.

[9] A. Bell and T. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol.7, Nov. 1995, pp.1129-1159.

[10] J. Yang-Won, K. Hong-Goo, L. Chungyong, Y. Dae-Hee, C. Changkyu, and K. Jaywoo, "Adaptive Microphone Array System with Two-Stage Adaptation Mode Controller," *in IEICE Trans. Fundamentals*, vol. E88-A, no. 4, Apr. 2005.