INTERSPEECH 2009 BRIGHTON

Speech enhancement in a 2-dimensional area based on power spectrum estimation of multiple areas with investigation of existence of active sources

[1]*Yusuke Hioka,* [1]*Ken'ichi Furuya,* [1]*Yoichi Haneda,* and [2]*Akitoshi Kataoka*

[1]NTT Cyber Space Laboratories, NTT Corporation, Tokyo, Japan
[2]Faculty of Science and Technology, Ryukoku University, Shiga, Japan

Abstract

A microphone array that emphasizes sound sources located in a particular 2-dimensional area is described. We previously developed a method that estimates the power spectra of target and noise sounds using multiple fixed beamformings. However, that method requires the areas where the noise sources are located to be restricted. We describe the principle of this limitation then propose a procedure that investigates the possibility of the existence of a sound source in a target area and other areas beforehand to reduce the number of unknown power spectra to be estimated.

Index Terms: microphone array, power spectrum estimation, speech enhancement, 2-dimensional area, estimating source position

10.21437/Interspeech.2009-412

1. Introduction

With the popularisation of teleconference and voice recognition systems, hands-free microphone systems have become indispensable devices due to their convenience. However, they often suffer from SNR deterioration due to the speaker's mouth being positioned a few metres away from the microphone, which is much farther than the mouth's position during use of a conventional handset or close-talking microphone. In particular, an intercom with a hands-free microphone may be used in a public space where various nonstationary noise sources are located around the speaker. In such an environment, the microphone might pick up confidential conversations that should not be sent to the other speaker. Thus, a hands-free microphone that picks up only the speech located in a particular 2-dimensional area, i.e., the area where the speakers are located, is required.

For that purpose, using a microphone array[1] is a well-known strategy, and various conventional methods have been reported. However, few works[2, 3, 4, 5] have dealt with the problem of discriminating sounds whose sources are located in the same direction but at different distances. One major approach reported is beamforming. Fixed beamforming[2] has the advantage of robustness against the change in noise source positions because its design is independent of the input signal; however, its noise suppression performance is limited. Adaptive beamforming is a different beamforming strategy that achieves better noise-suppression performance after sufficient adaptation calculation[3, 4]. However, its performance is sometimes severely degraded by noise source movement because the convergence speed of the adaptation lags behind the changes in noise position. There is also a method based on independent component analysis (ICA)[5], but it also lacks the robustness against rapid change in noise positions because the ICA also includes an iterative calculation. As the beamformer is recognized as a linear spatial filter, a restricted design may cause instability in its response.

Another framework known as "fixed beamforming with nonlinear postfilter"[6] achieves sufficient noise suppression through the nonlinearity of the postfilter. Saruwatari *et al.* proposed a method within this framework that estimates the noise power spectrum for postfiltering using complementary beamforming[7]. However, even this method also encounters instability of the linear system due to its estimation of the noise power spectrum in the linear processing framework. In addition, this method deals with sound emphasis in a particular direction but only in a 2-dimensional area.

The authors previously proposed an algorithm[8] that estimates sound spectra whose sources are located in the target sound areas using a pair of small microphone arrays. As both the spectrum estimation and postfiltering are performed in the power spectrum domain, which belongs to nonlinear processing, the method sufficiently emphasizes the speech without suffering from the instability problem. One major drawback of the method is that the area where the noise source is located must be restricted. Motivated by this situation, in this paper, we determine the maximum number of areas for which the method is able to simultaneously estimate sound power spectra. Then we propose a method to investigate areas that include active sound sources in order to avoid restriction of the noise source positions.

2. Power spectra estimation in 2-dimensional area

Many of the conventional speech enhancement procedures based on nonlinear postfilter designs necessitate the power spectra of target speech and noise to be estimated *a priori*[6]. For example, a well-known Wiener filter[9] is defined by $G(\omega) = \lambda_{\mathrm{S}}/(\lambda_{\mathrm{S}} + \sum \lambda_{\mathrm{N}})$, where λ_{S} and $\sum \lambda_{\mathrm{N}}$ are the power spectra of the speech and noise, respectively. In this section, we propose a method for estimating λ_{S} and λ_{N} when the sound sources are located in a particular angular sector. We then extend the method to cover 2-dimensional cases.

2.1. Power spectrum estimation using fixed beamformings

We briefly explain here the basic structure of power spectrum estimation. When we have N fixed beamformings whose complex directivity gains are assumed to be constant within each M angular sector, depicted by dashed-line in Fig. 1, the output of the n-th beamforming is given by $Y_n(\omega) = \sum_m a_{nm}(\omega) S_m(\omega)$. Here, $a_{nm}(\omega)$ is the constant complex directivity gain of n-th beamforming in the m-th angular sector, and $S_m(\omega)$ is the spectrum of the signal located in the m-th angular sector. ω denotes the frequency bin. If only the directivity gain to the m-th angular sector is not zero while the other gains are all zero, the beamforming is able to estimate the spec-

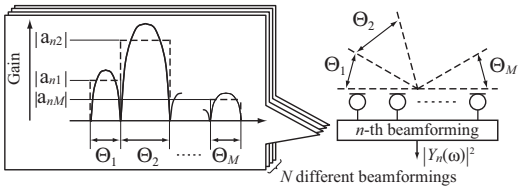

Figure 1: *Angular sectors and directivity of n-th beamforming*

trum of each sound source separately. However, this is often impossible because the beamforming that satisfies the $a_{nm}(\omega)$ condition cannot be designed in practice. This is because the beamforming works to retrieve the phase as well as the power spectrum of the signal.

Our method estimates only the power spectrum using multiple beamforming outputs, given by

$$\begin{bmatrix} |Y_1(\omega)|^2 \\ \vdots \\ |Y_N(\omega)|^2 \end{bmatrix} \simeq \begin{bmatrix} |a_{11}(\omega)|^2 \cdots |a_{1M}(\omega)|^2 \\ \vdots \quad \ddots \quad \vdots \\ |a_{N1}(\omega)|^2 \cdots |a_{NM}(\omega)|^2 \end{bmatrix} \begin{bmatrix} |S_1(\omega)|^2 \\ \vdots \\ |S_M(\omega)|^2 \end{bmatrix}. \quad (1)$$

The power spectrum of signals in each angular sector is given by solving this simultaneous equation as long as the columns of the matrix are linearly independent. Because the experimental directivity shape is not like the ideal design assumed above, the average gains of actual directivity (depicted by solid-line in Fig. 1) for each angular sector are used for $a_{nm}(\omega)$. Therefore, the left side of Eq. (1) is an approximation of the right side. In the following, we explain the expansion of this power spectrum estimation for sound sources located in a 2-dimensional area.

2.2. Definition of *Area*

In general, a microphone array requires a sufficiently large aperture size to discriminate the distances of the sources. To satisfy this requirement, we introduce a pair of separated microphone arrays, "array-L" and "array-R". When M_L and M_R angular sectors are defined for respective arrays, in this placement, a 2-dimensional *Area* is defined by the combination of angular sector $\Theta_{\mathtt{a}m}$ for both array-L and array-R, where $\mathtt{a}$ and m indicate the array that either takes the value "L" or "R", and the number of angular sectors, respectively. For example, when three angular sectors ($M_L = M_R = 3$) are assumed, as shown in Fig. 2, the *target Area* S and eight *noise Area*s RR, R, C, L, LL, NC, NR, and NL are defined. Note that in the following we denote the Areas in an *italic* font, e.g., "*S*", while the arrays are denoted in a Roman font, e.g., "L". In addition, we define the sum of the power spectra of signals whose sources are located in the Area $\mathtt{b}$ as $P_{\mathtt{b}}(\omega, l)$ ($\mathtt{b} \in \{LL, L, C, R, RR, NR, NC, NL\}$). Regarding the Area, the following discussion uses the definition in Fig. 2 unless otherwise stated.

The positions of the arrays and Areas are arbitrarily determined under the following restrictions: (a) the distance between the target Area and each microphone array is nearly the same ($d_L \simeq d_R$), (b) the distance between two microphone arrays is unknown but longer than that between the target Area and each microphone array ($d_{inter} \geq d_L$ or d_R), and (c) the target Area is located in front of each microphone array. Furthermore, the received signal can be recognized as a plane wave because the aperture length of each microphone array is sufficiently short. In addition, we recognize that each sound signal is uncorrelated.

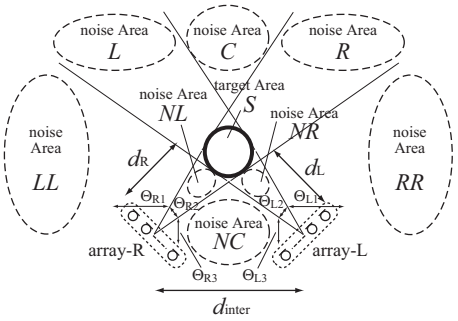

Figure 2: *Definition of Areas*

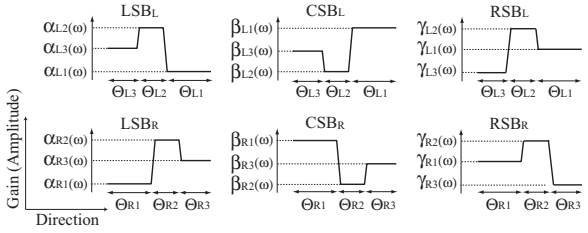

Figure 3: *Beampatterns of fixed beamformers: left suppression beamformer (LSB), centre suppression beamformer (CSB), and right suppression beamformer (RSB).*

2.3. Power spectrum estimation of Areas

With regard to the single array case mentioned in Sec. 2.1, we introduce fixed beamformings applied to both array-L and array-R. When each beamforming is designed to have a directivity shape with a directivity null pointing towards one of the angular sectors $\Theta_{\mathtt{a}m}$ as depicted in Fig. 3, the relationship between the input and output power spectra of the beamformings is described by Eq. (2). Here, $\mathbf{Y}(\omega, l)$ and $\mathbf{Z}(\omega, l)$ are column vectors that consist of the output power spectra of fixed beamformings and the unknown power spectra of each Area, respectively. Due to limited space, the notations ω and l, which denote the frequency bin and frame index, are omitted in Eq. (2).

$$\underbrace{\begin{bmatrix} |Y_{\mathrm{LL}}|^2 \\ |Y_{\mathrm{CL}}|^2 \\ |Y_{\mathrm{RL}}|^2 \\ |Y_{\mathrm{LR}}|^2 \\ |Y_{\mathrm{CR}}|^2 \\ |Y_{\mathrm{RR}}|^2 \end{bmatrix}}_{\mathbf{Y}(\omega,l)} \simeq \underbrace{\begin{bmatrix} \alpha_{\mathrm{L2}}^2 & \alpha_{\mathrm{L3}}^2 & \alpha_{\mathrm{L2}}^2 & \alpha_{\mathrm{L3}}^2 & \alpha_{\mathrm{L1}}^2 & \alpha_{\mathrm{L3}}^2 & \alpha_{\mathrm{L1}}^2 & \alpha_{\mathrm{L1}}^2 & \alpha_{\mathrm{L2}}^2 \\ \beta_{\mathrm{L2}}^2 & \beta_{\mathrm{L3}}^2 & \beta_{\mathrm{L2}}^2 & \beta_{\mathrm{L3}}^2 & \beta_{\mathrm{L1}}^2 & \beta_{\mathrm{L3}}^2 & \beta_{\mathrm{L1}}^2 & \beta_{\mathrm{L1}}^2 & \beta_{\mathrm{L2}}^2 \\ \gamma_{\mathrm{L2}}^2 & \gamma_{\mathrm{L3}}^2 & \gamma_{\mathrm{L2}}^2 & \gamma_{\mathrm{L3}}^2 & \gamma_{\mathrm{L1}}^2 & \gamma_{\mathrm{L3}}^2 & \gamma_{\mathrm{L1}}^2 & \gamma_{\mathrm{L1}}^2 & \gamma_{\mathrm{L2}}^2 \\ \alpha_{\mathrm{R2}}^2 & \alpha_{\mathrm{R1}}^2 & \alpha_{\mathrm{R1}}^2 & \alpha_{\mathrm{R2}}^2 & \alpha_{\mathrm{R1}}^2 & \alpha_{\mathrm{R3}}^2 & \alpha_{\mathrm{R2}}^2 & \alpha_{\mathrm{R3}}^2 & \alpha_{\mathrm{R2}}^2 \\ \beta_{\mathrm{R2}}^2 & \beta_{\mathrm{R1}}^2 & \beta_{\mathrm{R1}}^2 & \beta_{\mathrm{R2}}^2 & \beta_{\mathrm{R1}}^2 & \beta_{\mathrm{R3}}^2 & \beta_{\mathrm{R2}}^2 & \beta_{\mathrm{R3}}^2 & \beta_{\mathrm{R2}}^2 \\ \gamma_{\mathrm{R2}}^2 & \gamma_{\mathrm{R1}}^2 & \gamma_{\mathrm{R1}}^2 & \gamma_{\mathrm{R2}}^2 & \gamma_{\mathrm{R1}}^2 & \gamma_{\mathrm{R3}}^2 & \gamma_{\mathrm{R3}}^2 & \gamma_{\mathrm{R2}}^2 & \gamma_{\mathrm{R3}}^2 \end{bmatrix}}_{\mathbf{T}(\omega)} \underbrace{\begin{bmatrix} P_S \\ P_{LL} \\ P_L \\ P_{NL} \\ P_C \\ P_{NC} \\ P_{RR} \\ P_R \\ P_{NR} \end{bmatrix}}_{\mathbf{Z}(\omega,l)} \quad (2)$$

The matrix $\mathbf{T}(\omega)$ consists of column vectors given by Eq. (3), where each vector is composed of a set of squared directivity gains for each Area.

$$\mathbf{T}(\omega) := \begin{bmatrix} \mathbf{t}_S & \mathbf{t}_{LL} & \mathbf{t}_L & \mathbf{t}_{NL} & \mathbf{t}_C & \mathbf{t}_{NC} & \mathbf{t}_{RR} & \mathbf{t}_R & \mathbf{t}_{NR} \end{bmatrix} \quad (3)$$

As $\mathbf{Y}(\omega, l)$ and $\mathbf{T}(\omega)$ consist of known parameters, we have power spectra estimated using $\mathbf{Z}(\omega, l)$, which is given by solving the simultaneous equation.

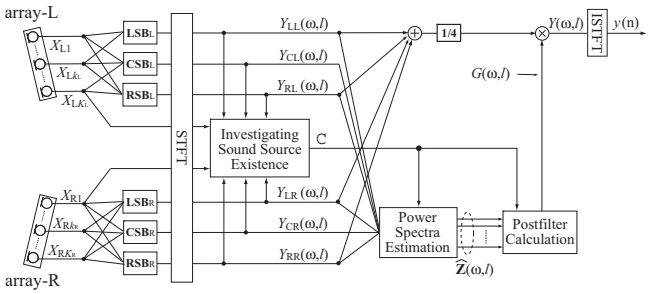

Figure 4: *Process flow of proposed method.*

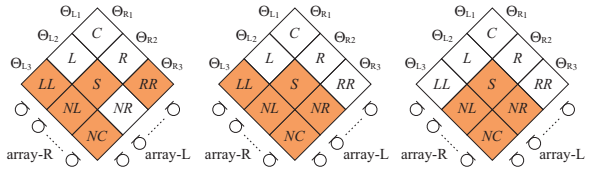

Figure 5: *Example of Area restriction for estimation.*

2.4. Discussion on maximum number of power spectra to be estimated

The problem in solving Eq. (2) is that the simultaneous equation is under-determined because the rank of $\mathbf{T}(\omega)$ is smaller than the number of unknown variables in $\mathbf{Z}(\omega, l)$. This is caused by the lack of degrees of freedom in specifying a particular Area, which is determined by the number of angular sectors defined at each array.

In the design of beamforming, in principle, at least K microphones are required for directivity control in K directions. Therefore, the minimum number of microphones required for each array is decided by the number of angular sectors. In the case of Fig. 2, we need at least $M_L + M_R (= 6)$ microphones. Normally, we can estimate the power spectra of 6 sound sources using this array, but we lose one degree of freedom in the directivity design because the distance between the array-L and array-R is unknown. Thus, the rank of $\mathbf{T}(\omega)$ is reduced to $M_L + M_R - 1 (= 5)$. Consequently, to modify Eq. (2) so that it is determined, we need to reduce the number of unknown variables in $\mathbf{Z}(\omega, l)$ down to 5, which is achieved by restricting the number of Areas where sound sources exist. Figure 5 shows examples of this Area restriction where only the power spectrum of sound sources in the shaded Area is estimated. The Areas are located in all three angular sectors for both arrays in Fig. 5(a). In this case, the rank of $\mathbf{T}(\omega)$ is 5, which means Eq. (2) is determined. However, in some cases, the rank of $\mathbf{T}(\omega)$ becomes smaller than 5, although the number of variables in $\mathbf{Z}(\omega, l)$ is reduced. We found that this occurs depending on the combination of Areas selected, as that includes sound sources. For example, there is no Area including a sound source in the angular sector Θ_{L1} in Fig. 5(b). Because only two degrees of freedom of array-L are used, the rank of $\mathbf{T}(\omega)$ is reduced to 4. The same problem occurs if the number of Areas is restricted to 4, as in Fig. 5(c). Eventually, such a problem will not occur if the number of Areas is reduced to the smaller of M_L or M_R.

3. Proposed method

The process flow of the proposed method is shown in Fig. 4. The method mainly consists of the following steps: (a) investigating the possibility of an active sound source existing, (b) estimating the power spectrum, and (c) postfiltering. These steps are explained in the following.

3.1. Investigating possibility of existence of active sound source

Let us remember that a nonstationary signal such as speech satisfies the sparseness assumption on the time-frequency plane[10]. This means the spectrum of input signals at an arbitrary frequency bin and frame is composed of a few sound sources. With this assumption, the method preliminarily estimates the noise positions by investigating the possibility of each Area including active sound sources or not. Using this information, we omit the Area that does not include active sound sources from the power spectra estimation. This enables reformulation of the simultaneous equation to be over-determined, where the equation can be solved by the least squares method, by reducing the number of unknown variables.

The possibility is investigated by using the amount of input-output (I/O) level differences of the beamformings. As each beamforming in Fig. 3 is designed to point its directivity null towards a particular angular sector, the I/O level difference decreases if an active sound source is located in the corresponding angular sector. Because each Area is defined by the combination of the angular sectors of array-L and array-R, we can specify the Area where an active sound source exists by examining the amplitude of $IO_{\mathtt{b}}(\omega, l) = \frac{|Y_{\mathtt{X}L}(\omega,l)||Y_{\mathtt{X}R}(\omega,l)|}{|X_{Lk}(\omega,l)||X_{Rk}(\omega,l)|}$ for Area $\mathtt{b}$, where $\mathtt{X} = \{C, L, R\}$, and k is an arbitrary microphone of each array. The smaller this value is, the higher the possibility that a sound source exists.

3.2. Power spectrum estimation under Area restriction

The number of unknown variables in $\mathbf{Z}(\omega, l)$ is reduced by using the investigated possibilities of active sound source existence. The rank of $\mathbf{T}(\omega)$ is not more than 5, so we first discard 4 out of 9 Areas whose product of I/O level differences is large. Thus, we newly introduce a vector $\mathbf{T}_5(\omega) := [\, \mathbf{t}_{\mathtt{c}1}\ \mathbf{t}_{\mathtt{c}2}\ \mathbf{t}_{\mathtt{c}3}\ \mathbf{t}_{\mathtt{c}4}\ \mathbf{t}_{\mathtt{c}5}\,]$, where the set $\mathtt{C} := \{\mathtt{c}1, \cdots, \mathtt{c}5\}$ consists of the names of remaining (not discarded) Areas listed in ascending order of the product of I/O level differences, i.e., $IO_{\mathtt{c}1} < IO_{\mathtt{c}2} < \cdots$. Then, the power spectra of the remaining Areas are estimated by solving the modified simultaneous equation

$$\mathbf{T}_5(\omega)\mathbf{Z}_5(\omega, l) \simeq \mathbf{Y}(\omega, l), \tag{4}$$

where $\mathbf{Z}_5(\omega, l) := \left[\, P_{\mathtt{c}1}\ P_{\mathtt{c}2}\ P_{\mathtt{c}3}\ P_{\mathtt{c}4}\ P_{\mathtt{c}5}\,\right]^T$. The equation is now over-determined, so it can be solved by the least squares method as $\widehat{\mathbf{Z}_5}(\omega, l) = \mathbf{T}_5^+(\omega)\mathbf{Y}(\omega, l)$, where $^+$ and $\widehat{\cdot}$ denote the Moore-Penrose pseudo inverse and estimated value, respectively. If $\mathbf{T}_5(\omega)$ is still rank deficient, we further discard the Area with the least possibility of containing a sound source from the estimation until the rank of the matrix corresponds to the number of Areas to be estimated.

3.3. Modified postfilter based on selected Areas

For emphasizing the sound sources in the target Area, we multiply the sum of fixed beamforming outputs whose directivity nulls are not pointing towards the target Area, i.e., $\mathrm{LSB}_{\mathtt{a}}$ and $\mathrm{RSB}_{\mathtt{a}}$. The postfilter $G(\omega, l)$ is calculated using the power spectra of the target and noise Area estimated by well-known conventional methods, such as Wiener filtering[9].

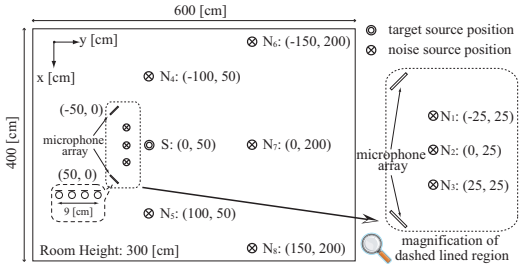

Figure 6: *Positions of microphone array and sound sources.*

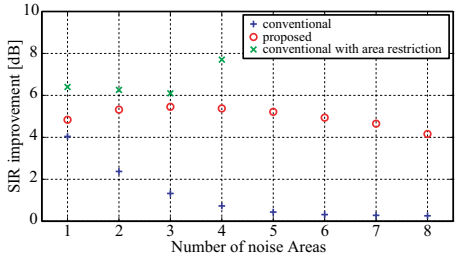

Figure 7: *Change of average SIR improvement depending on number of noise sources.*

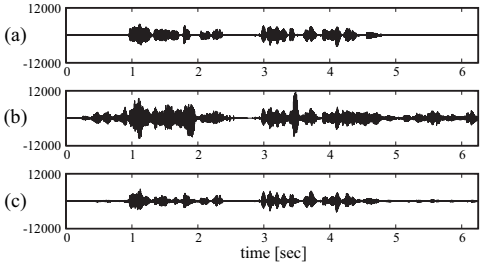

Figure 8: *Observed waveforms in experiment: (a) target speech, (b) input signal, (c) output of proposed method.*

4. Experimental evaluation

Experiments were performed in a room whose reverberation time was approximately 250 ms. The frame was shifted every 16 ms, and the length of each frame was 32 ms at 16 kHz. For the angular sectors, we set $\Theta_{L1} = [10^\circ, 70^\circ]$, $\Theta_{L2} = [-10^\circ, 10^\circ]$, and $\Theta_{L3} = [-50^\circ, -10^\circ]$ for array-L, and their symmetric sets were adopted for array-R. The positions of microphone arrays and sound sources are shown in Fig. 6. The target speaker is positioned at S, which is located inside the target Area, and the noise sources are positioned at N_1 to N_8. Each microphone array was composed of four microphones in a linear and equispaced configuration with an intermicrophone distance of 3 cm. For the postfilter design, we used Wiener filtering, i.e., $G(\omega, l) := \widehat{P_S}(\omega, l) / \sum_{i \in \mathtt{C}} \widehat{P}_i(\omega, l)$, except for the case where the target Area was not included in the set $\mathtt{C}$. In that case, we recognized that the target sound source was not active; thus, the $G(\omega, l)$ was set to a small constant. For comparison, we also evaluated the conventional method[8], which assumes that no noise exists in the Areas NL, NR, and NC.

First, we evaluated the noise suppression performance with regard to the number of Areas including active sound sources. We measured the improvements in the signal-to-interference ratio (SIR)[11] when the number of Areas that included active noise sources changed. The average SIR improvement when the number of active noise Areas was changed is shown in Fig. 7. The results showed that the conventional method loses its ability to suppress noise as the number of Areas increases, while the proposed method maintains its effectiveness even if noise sources are located at every noise Area. We also showed the average SIR improvements of the conventional method when the number of Areas was restricted to 4. In comparison, the results of the proposed method were not much worse than those of the conventional method with restriction. Fig.8 shows the waveforms of the target speech, the input signal, and the output signals when noise source is located at the positions P_3, P_4, P_6, P_7, and P_8. As can be seen, the waveform of the output signal resembles that of the target speech.

In the above experiment in Fig.7 when the number of noise Area is 1, we also evaluated the amount of distortion in the extracted target sounds using the signal-to-distortion ratio (SDR)[11]. We found no large dispersion depending on the noise source position where the standard deviation of SDR was 0.55 dB. This means the quality of the target sounds extracted by the proposed method was independent of the noise source positions.

5. Concluding remarks

A method for estimating power spectra in a particular 2-dimensional area has been proposed. We first introduced our method of estimating power spectra using multiple fixed beamforming, and then extended the method to a 2-dimensional case and discussed the maximum number of Areas for which the method can simultaneously estimate power spectra. We have also proposed a way to overcome this limitation by investigating the source location and experimentally examined the effectiveness of this method.

6. References

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

[2] J. G. Ryan and R. A. Goubran, "Near-field Beamforming for Microphone Arrays," ICASSP 1997, pp. 363–366.

[3] Y. R. Zheng and R. A. Goubran, "Robust Near-field Adaptive Beamforming With Distance Discrimination," IEEE Trans. Speech and Audio Processing, Vol. 12, No. 5, pp. 478–488, 2004.

[4] Y. R. Zheng, P. Xie and S. Grant, "Robustness and Distance Discrimination of Adaptive Near Field Beamformings," ICASSP 2006, pp. 1041–1044.

[5] A. Ando, M. Iwaki, K. Ono and K. Kurozumi, "Separation of Sound Sources Propagated in the Same Direction," IEICE Trans. on Fundamentals, Vol. E88-A, No. 7, pp. 1665–1672", 2005.

[6] R. Zelinski, "A Microphone Array with Adaptive Post-filtering for Noise Reduction in Reverberant Rooms," ICASSP 1988, pp. 2578–2581.

[7] H. Saruwatari, S. Kajita, K. Takeda and F. Itakura, "Speech Enhancement Using Nonlinear Microphone Array Based on Complementary Beamforming," IEICE Trans. Fundamentals, Vol. E82-A, No. 8, pp. 1501–1510, 1999.

[8] Y. Hioka, K. Kobayashi, K. Furuya and A. Kataoka, "Enhancement of Sound Source Located within a Particular Area Using a Pair of Small Microphone Arrays," IEICE Trans. Fundamentals, Vol. E91-A, No. 2, pp. 561–574, 2008.

[9] J. Benesty, S. Makino and J. Chen, Speech Enhancement, Springer-Verlag, 2005.

[10] O. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-frequency Masking," IEEE Trans. Signal Processing, Vol. 52, No. 7, pp. 1830–1847, 2004.

[11] S. Araki, H. Sawada, R. Mukai and S. Makino, "Underdetermined Sparse Source Separation of Convolutive Mixtures with Observation Vector Clustering," ISCAS 2006, pp. 2594–2597.