A TALKER TRACKING METHOD USING TWO MICROPHONES BASED ON THE SOUND SOURCE LOCALIZATION

†*Kenji Suyama, and* ††*Kota Takahashi*

†Faculty of Engineering, Tokyo DENKI University,
2-2 Kandanishikicho, Chiyoda-ku, Tokyo JAPAN,
E-mail: *suyama@cck.dendai.ac.jp*,

††The University of Electro-Communications,
1-5-1 Chofugaoka, Chofu-shi, Tokyo JAPAN
E-mail: *kota@ice.uec.ac.jp*

ABSTRACT

In this paper, we propose a novel method for a target talker tracking using two microphones, which is often referred as the two channel microphone array. The target sound is adaptively extracted by the frequency domain generalized sidelobe canceler(FDGSC). Then, the results of sound source localization method, which is one of methods based on the time difference of arrival between two microphones and enables us to localize multiple sources in a realtime processing, utilized for the adjustment of the direction of target sound in the beamformer of the FDGSC. A superior performance of the proposed method are shown by several numerical experiments.

1. INTRODUCTION

Recently, a two–channel microphone array technique[1]-[3], which is a sound capture system using only two microphones, is often used as a front-end of the speech recognition and speech communication. Although the two–channel microphone array is attractive since it can be easily implemented using a stereo input of PC's sound card and so on, a restriction of few microphones prevents a realization of high spatial resolution of the array. Thus, it is difficult to acquire a high S/N improvement in the noisy environment.

The generalized sidelobe canceler(GSC)[4] was often utilized to improve the S/N of received sound. The GSC is constructed from two parts. One of them is a fixed beamformer, and forms a peak of the directional pattern of array toward the target sound direction. The other is the sidelobe canceler, and forms a null toward the noisy sound direction. In general, since the target source is considered to be fixed to a specific direction, which is often referred as the look–direction, the problem of target sound extraction falls into forming the null by adjusting the sidelobe canceler. However, in a case that the target source is moving, we have to adjust not only the sidelobe canceler but also the beamformer, that is the look–direction.

In reference [5], the frequency domain generalized sidelobe canceler(FDGSC) with a target source tracking mechanism was used for such the situation. The FDGSC is realized on the frequency domain, and the beamformer and the sidelobe canceler are adaptively applied to the received sounds in each the frequency divided using the discrete Fourier transform(DFT). In the method, the direction of target sound source is estimated adaptively in a criterion of the minimization of power of the target sound at the input of sidelobe canceler under several conditions. In those conditions, it is included that the S/N of the received sound is always higher than 0 at every sections of the sound signal. In a case that both the target and the noisy sound are the speech signals, it is very tight condition for us due to existence of the silent section of speech signals.

On the other hand, we apply the results of the sound source localization method for adjusting the direction of target sound. In the method, the data selection(DS) method[6] is used as the sound source localization technique. Although the DS method is based on the time-difference-of-arrival(TDOA) between two microphones, which is often considered as the method for the single source localization, the method enables us to localize multiple sound sources in a realtime signal processing. Moreover, the method is executed on the frequency domain and easily incorporated into the FDGSC. In this paper, several superior performance of the method are shown through the numerical experiments.

2. A TARGET TRACKING PROBLEM

As shown in the figure 1, the target sound signal $s(t)$ and noisy sound signal $n(t)$, which are both speech signals, are collected by the two microphones separated by d. The direction of $s(t)$ is $\theta_s(t)$, and moves with times. Then, a moving range of $\theta_s(t)$ is known as Θ_s and an initial direction of $\theta_s(t)$ is known as $\theta_s(0)$. The direction of $n(t)$ is fixed to θ_n.

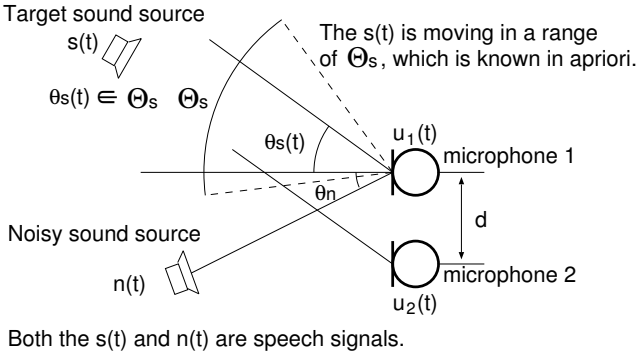

Figure 1: Problem description of the target sound tracking using two microphones.

When $s(t)$ and $n(t)$ are enough far from the array, the received sound of microphone 1 and 2, $u_1(t)$ and $u_2(t)$ can be described as

$$u_1(t) = s(t) + n(t) \quad (1)$$
$$u_2(t) = s(t - \tau_s(t)) + A_n n(t - \tau_n) \quad (2)$$

where $\tau_s(t) = d \sin\theta_s(t)/c$, $\tau_n = d \sin\theta_n/c$, and c is a sound velocity. The purpose of the target tracking is to extract $s(t)$ from $\mathbf{u}(t) = (u_1(t), u_2(t))^T$.

3. TARGET SOUND EXTRACTION USING THE FDGSC

In this section, a method of the target sound extraction using the frequency domain generalized sidelobe canceler(FDGSC) shown in figure 2 is described.

The received sound, $\mathbf{u}(t)$, are transposed to $U_m(k,n)$ by N-point DFT, where $U_m(k,n)$ presents the frequency component at the frame n and the frequency k of $u_m(t)$.

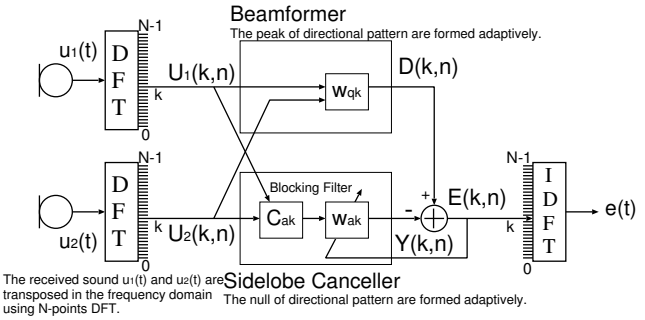

Figure 2: Target sound extraction using the FDGSC.

The FDGSC are constructed from following two parts. One of them is the beamformer part, in which, $\mathbf{U}(k,n) = (U_1(k,n), U_2(k,n))^T$ is transposed to be an equi-phase by following procedure,

$$D(k,n) = \frac{1}{2}\mathbf{w}_{qk}^H \mathbf{U}(k,n) \tag{3}$$

where $\mathbf{w}_{qk} = 0.5(1, e^{-j\omega_k \tau_s(n)})^T$, $D(k,n)$ is the output of beamformer, and H means the Hermite transpose. If $\tau_s(n)$ is set appropriately, a peak of the array pattern is directed toward the target sound.

The another part of FDGSC is the sidelobe canceler part. At there, $\mathbf{U}(k,n) = (U_1(k,n), U_2(k,n))^T$ is transposed to $X(k,n) = \mathbf{C}_{ak}^H \mathbf{U}(k,n)$, where $\mathbf{C}_{ak} = 0.5(1, -e^{-j\omega_k \tau_k(n)})^T$. According to this operation, the target sound component included in $\mathbf{U}(k,n)$ are eliminated. $X(k,n)$ is used as the input signal of the adaptive filter with a weight w_{ak}, the output signal is calculated as $w_{ak}^* X(k,n)$, where $*$ presents the complex conjugate operation. Finally, we can get $e(t)$ as the extracted sound by the inverse DFT(IDFT) of $E(k,n)$ described as,

$$E(k,n) = D(k,n) - w_{ak}^*(n) X(k,n). \tag{4}$$

Then, w_{ak} is updated with following the Normalized Least Mean Square(NLMS) algorithm,

$$w_{ak}(n+1) = w_{ak}(n) + \mu \frac{X^*(k,n)E(k,n)}{\sum_k |X(k,n)|^2} \tag{5}$$

where μ presents the step-size parameter.

If $\tau_s(n)$ can be estimated appropriately at the frame n, we can extract $s(t)$ by the above-mentioned procedure. Therefore, it is important to develop the estimation method of $\tau_s(n)$. In next section, we will describe a new estimation method for $\tau_s(n)$ based on the data selection procedure[6].

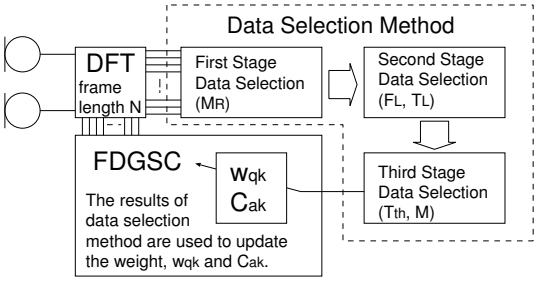

Figure 3: The procedure of the three phase data selection method.

4. SOUND SOURCE LOCALIZATION BASED ON DATA SELECTION

In this section, a new method for moving sound source localization using data selection procedure[6] is described. Fig.3 shows a block diagram of the method. This method is based on the time-difference-of-arrival(TDOA) between two microphones, and contains the following two features: (1) no *a priori* information about the number of sound sources is required, that is, if multiple sources exist simultaneously, we can estimate each the direction-of-arrival(DOA) separately using this method, (2) the real time processing is possible.

In this method, two kinds of sound source signals as the target signals are assumed. These signals have special characters on a time-frequency(TF) plane. One of them is a signal whose energy localizes in a specific frequency on the TF plane. We refer this kind of signal as a f-localized signal in this method. A vowel of the speech signals is one of the f-localized signals. The other is a signal whose energy localizes in a specific time on the TF plane. We refer this kind of signal as a t-localized signal in this method. A consonant of the speech signals is one of the t-localized signal.

In the method, $\mathbf{U}(k,n)$ are available for the estimation of DOA based on the TDOA. At k and n, we can estimate the TDOA $\tau_k(n)$ by $\tau_k(n) = (\angle U_2(k,n) - \angle U_1(k,n))N/2\pi k$, and estimate the DOA by $\theta_k(n) = \sin^{-1}(c\tau_k(n)/d)$. Then, the energy distribution of $\mathbf{U}(k,n)$ can be classified on the TF plane into two regions; (I) the region where only a single source energy exists, and (II) the region where multiple source energy exists. If only the results estimated in region (I) can be selected, we can obtain results to be considered as correct one since the method based on the TDOA is originally suitable for the single source localization. The problem is how to select those results. In this method, those results can be selected using the three phase data selection.

4.1 First stage data selection

In the first stage data selection, we select only $\tau_k(n)$ estimated from $\mathbf{U}(k,n)$ to be considered that only the single source energy contain on the TF plane.

As the index for selection, we use the min/max ratio, M_R, of the amplitude of $U_1(k,n)$ and $U_2(k,n)$, and it is defined as following,

$$M_R = \frac{\min\{|U_1(k,n)|, |U_2(k,n)|\}}{\max\{|U_1(k,n)|, |U_2(k,n)|\}} \in [0,1]. \tag{6}$$

As depicted in figure 4, when only the single source energy is included in $\mathbf{U}(k,n)$, it can be considered that $|U_1(k,n)|$ and $|U_2(k,n)|$ have nearly similar amplitude. It can be expected that the correct results are obtained by using such the

$\mathbf{U}(k,n)$. Then, M_R approaches to 1. Therefore, we provide the threshold M_0 for the selection, only $\tau_k(n)$ estimated from $\mathbf{U}(k,n)$ satisfying $M_R \geq M_0$ are selected, and the others are rejected.

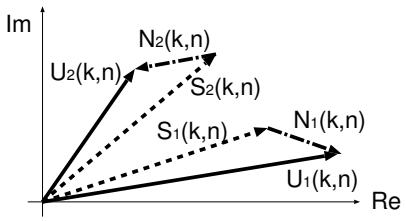

Figure 4: The amplitude of $\mathbf{U}_k(n)$.

4.2 Second stage data selection

In the second stage data selection, multiple estimation results on the TF plane are evaluated based on the $f-$localized signal or the $t-$localized signal simultaneously.

In the case of the f-localized signal, it can be considered that several results along the time(frame) axis in one frequency band k locate a same direction. Therefore, the threshold value F_L is provided to detect those results. If F_L results locate a same direction at frequency band k continuously, we infer that those results have been estimated from the f-localized signal. Then, those results are selected as shown in figure 5.

On the other hand, in the case of the t-localized signal, it can be considered that several results along the frequency axis at one time frame n estimate a same direction. Then, the threshold value T_L is provided to detect those results. If T_L results locate a same direction simultaneously, it is inferred that those results have been estimated from the t-localized signal. Then, those results are selected.

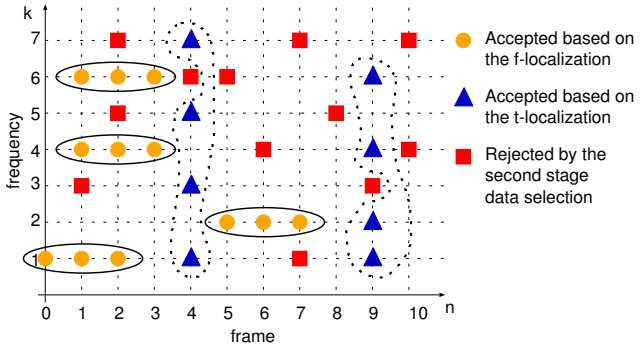

Figure 5: The second stage data selection.

4.3 Third stage data selection

When the second stage data selection have been applied, we can obtain the results as shown in figure 6. In the third data selection, we select only the results of the target sound signal.

For the selection, we provide the threshold T_{th} to judge the fluctuation from the previous frame. If the fluctuation of results between nearest two frames are within T_{th}, we judge that such the results are estimated from same source

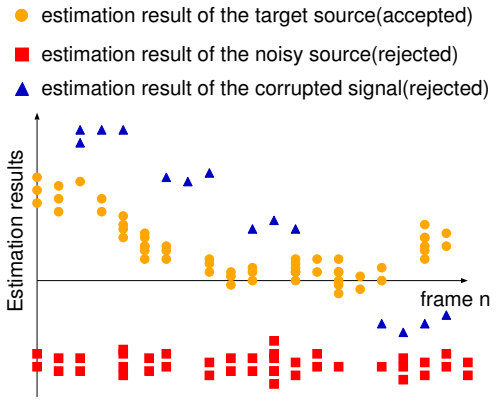

Figure 6: The estimation results after the second stage data selection.

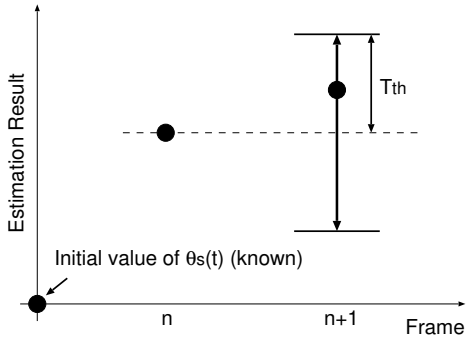

Figure 7: The third stage data selection.

as shown in figure 7. Then, the initial direction $\theta_s(0)$, which is known, is used for the results at the initial frame. We can obtain multiple results at one frame depicted in figure 6. Therefore, all results obtained at the same frame are averaged since only one result is required for the FDGSC.

By above-mentioned three phase data selection, we can the DOA of target sound source at each frame n, and those results are applied for the adjustment of $\mathbf{C}_{ak}$ in the FDGSC.

5. NUMERICAL EXPERIMENTS

The numerical experiments were performed to confirm the performance of the proposed method.

The speech signals were used as $s(t)$ and $n(t)$. The interval of two microphone d was set to 80[mm], and N was set to 128. The initial direction of $s(t)$ was set to $\theta_s(0) = 0°$, and the target sound source was moving within $\theta_s(t) : [-20°, 20°]$. The noisy sound source was set up at $\theta_n = 60°$.

When the S/N of received sound was set to 0[dB], the results of estimated DOA are shown in figure 8. Each threshold values for the three phase data selection were set to $M_0 = 0.77$, $T_L = 2$, $F_L = 4$, $T_{th} = 4.1$, which were decided at a preliminary experiment. As a comparison, the results by reference [5] were also included in the figure 8. Moreover, the waveform of the target sound $s(t)$, the noisy sound $n(t)$, and the extracted sound $e(t)$ were shown in figure 9, respectively. From these figures, even in the time section that the power of $n(t)$ was larger than one of $s(t)$, for example, $t : [0.65, 0.9]$, we have confirmed that the pro-

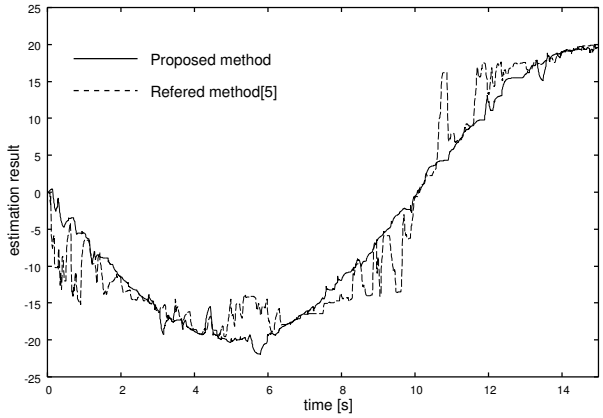

Figure 8: Estimation results by the three phase data selection when the S/N of received sound was 0[dB].

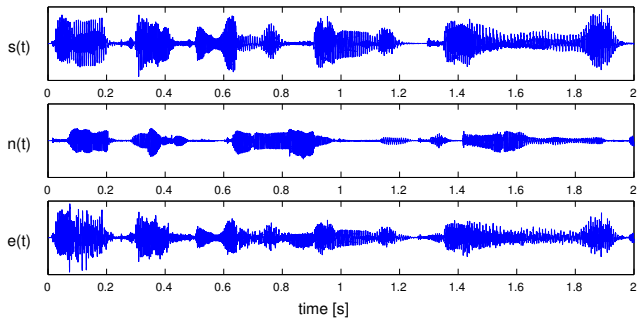

Figure 9: Waveform: $s(t)$ is the target sound, $n(t)$ is the noisy sound and $e(t)$ is the extracted sound.

posed method could estimate the direction of $s(t)$ precisely. Then, the S/N of $e(t)$ was 4.99[dB]. The figure 10 shows the estimation results when the S/N of received sound was set to -3[dB]. Then, $M_0 = 0.73$, $T_L = 2$, $F_L = 4$, $T_{th} = 7.9$ were used as the threshold values, and the S/N of $e(t)$ was 2.99[dB].

In the actual room environment, we have to use the fixed threshold values for the selection. Therefore, we have perform several experiments in various environments, and decided as the fixed threshold values to use $M_0 = 0.77$, $F_L = 2$, $T_L = 3$, $T_{th} = 5$. Figure 11 shows the S/N of $e(t)$ when the optimal threshold and the fixed threshold were used for the selection. Those results show that the fixed threshold values are available even in the various situations.

6. CONCLUSION

In this paper, we proposed a new method for a target talker tracking using two microphones. In the method, the FDGSC was used to extract the moving target sound signal, and the three phase data selection was used to estimate the direction of moving source. By numerical experiments, we confirmed that the proposed method was a superior performance, even if the speech signal was used as the noisy source signal.

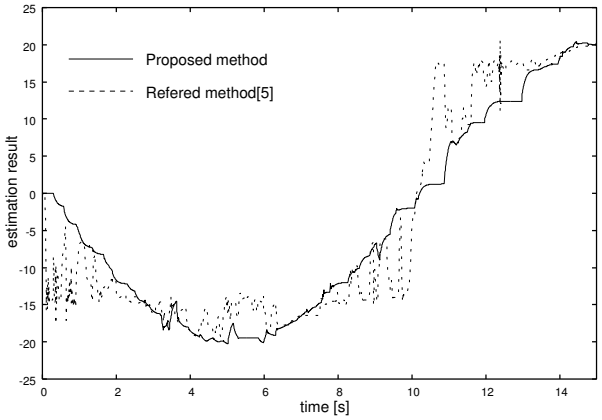

Figure 10: Estimation results by the three phase data selection when the S/N of received sound was -3[dB].

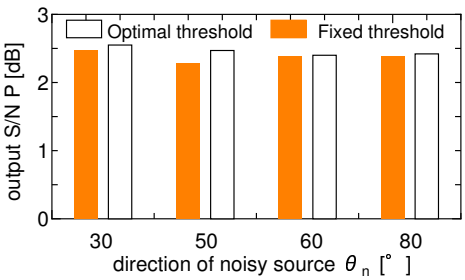

Figure 11: Output S/N when using the optimal threshold and the fixed threshold(the received sound S/N was set to -3[dB]).

REFERENCES

[1] J. V. Berghe : "An adaptive noise canceler for hearing aids using two nearby microphones", J. Acoust. Soc. Am., vol.103, no.6, pp.3621–3626, 1998.

[2] I. Cohen and B. Berdugo : "Two–channel signal detection and speech enhancement based on the transient beam–to–reference ratio", Proc. ICASSP2003, pp.V–233–236, 2003.

[3] Y. Hioka, Y. Koizumi and N. Hamada : "Improvement of DOA estimation using virtually generated multichannel data from two–channel microphone Array", Journal of Signal Processing, vol.7, no.1, pp.105–109, 2003.

[4] B. Widrow, P.E. Mantey, L.J. Griffiths and B.B. Goode, "Adaptive antenna system", Proc. IEEE, vol.55, no.12, pp.2143–2161, 1967.

[5] H. Kawakami, M. Abe and M. Kawamata : "A two–channel microphone array with adaptive target tracking using frequency domain generalized sidelobe cancellers", Proc. ISPACS2002, pp.291–296, 2002.

[6] K. Suyama, K. Takahashi and R. Hirabayashi : "A robust technique for sound source localization in consideration of room capacity", Proc. WASPAA2001, pp.63-66, 2001.