SOUND SOURCE SEPARATION WITH ROBUSTNESS TO VARIATION OF MICROPHONE GAIN CHARACTERISTIC

Shinya Matsui, Katsumasa Nagahama, Makoto Shozakai

Asahi Kasei Corporation, Atsugi AXT Main Tower 22F, Okata 3050,
Atsugi, Kanagawa, 243-0021 JAPAN
Tel: +81-46-230-4951, Fax: +81-46-230-4910
Email: {matsui.sk, nagahama.kc, shozakai.mb}@om.asahi-kasei.co.jp

ABSTRACT

Hands-free voice-activated operation of car navigation and car audio appliances in vehicle is required, not only from convenience but also safety point of view. However, deterioration of voice recognition performance is caused by an ambient conversational voice or an accidental noise. This paper proposes a voice separation system Butterfly Subtraction Array (BSA) based on a microphone array for voice recognition. This method uses a pair of beam formers that has a configuration of complex conjugate. Energies values of two outputs from beam formers are subtracted each other in a frequency domain. Consequently, zone separation of sound source is realized. Advantage of the proposed method is to maintain a stable performance for direction-lag of speaker position and large early-reflected sound by having spread directional pattern in moderation. And, highly important one is robustness on variations of microphone gain characteristics. Experimental results show that this method performs 78% of word recognition accuracy under double-talk situation of driver and passenger while 27% in a single microphone scenario.

1. INTRODUCTION

In recent years, voice recognition systems have become available in various practical applications. They have also been widely adopted in GPS-based automotive navigation systems as a tool for recognizing speech commands. However, the voice recognition system is developed as premises for single talk scene, and hence the recognition performances under double-talk situation become drastically worse. Therefore, on using the voice recognition system, fellow passengers, except a speaker who gives a command, are restricted to perform such activities as making a phone call, making noises, and so on. To solve the above problems, various methods to extract the aimed voice by using more than a single microphone have been proposed in a number of literatures.

As conventional researches, Delay-and-Sum (DS) [1] and Griffith-Jim adaptive array [2] have been suggested. Recent studies also suggested the use of blind source separation (BSS) [3] [4] [5]. However, Delay-and-Sum gives low performance in a few microphones. For making a super directive pattern, many microphones must be needed. Adaptive microphone arrays, for example AMNOR [6], cause significant performance degradation by high correlation between voices and noises. Additionally, VAD (Voice Activity Detection) function will be added for applying it into practical applications. BSS suffers from susceptibility to sound reverberation, and high calculation cost for embedded systems. As a matter of almost all microphone array techniques,"the difference of microphone characteristics" is a serious problem. In general, the lower they cost, the more piece-to-piece variations they have. And some frequency characteristic variations are about ± 3dB in each frequency bin. When DS is used, the piece-to-piece variation just only causes slight low performances. However for the adaptive array, it causes significant performance degradation; especially in low-frequency range under 1 kHz.

Furthermore, large early-reflected sound also poses serious problems. It is specific to an automotive environment. This early-reflected sound may be larger than direct sound in some microphone position. In such a case, performance will be degradation. We must solve the above problem to keep high performance for practical purposes in the automotive environment. However, it is usually difficult to find space for attaching many microphones inside an automotive interior, and to make each microphone share same characteristic with respect to cost issue. For these reasons, a realization of a speech-input system using a couple of cheap microphones is desired for the voice recognition system, which doesn't restrict behavior of fellow passengers.

In this paper, we proposed a sound source separation

method using microphone array for voice recognition system in car cabin, and evaluated the performance of this method. We utilize two beamformers having complex conjugate configuration. The sound source separation is realized by subtracting energy values of two beamformed outputs in each frequency bin. We call this method Butterfly Subtraction Array (BSA). To assess our method's validity, we conducted performance evaluation experiments of voice recognition in an automotive environment. As a result, it was found that recognition performances have been improved substantially under double-talk situation of a driver and a fellow passenger.

2. BUTTERFLY SUBTRACTION ARRAY (BSA)

Fig.1 shows a block diagram of the propose method, "Butterfly Subtraction Array" or BSA. Received signals at two microphones positions (Mic.1, Mic.2) are transformed into frequency domain by "Frequency Analysis 1" and "Frequency Analysis 2". The received signals in frequency domain are defined as follows,

$$X(f) = [x_1(f), x_2(f)] \tag{1}$$

In the second stage, the transformed signals are convoluted with the "Beam Former 1" and "Beam Former 2", which yields symmetrically placed null point of energy along center line of two microphones. We define the Beam Former 1 weights in each frequency bin f, as W_{BF1}, and the output signal of Beam Former 1, $S_1(f)$ can be calculated as,

$$S_1(f) = X(f)W_{BF1}(f) \tag{2}$$

Where,

$$W_{BF1}(f) = \begin{bmatrix} \exp[j2\pi f d_1 \sin\theta_1 / c] & \exp[j2\pi f d_2 \sin\theta_1 / c] \\ \exp[j2\pi f d_1 \sin\theta_2 / c] & \exp[j2\pi f d_2 \sin\theta_2 / c] \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{3}$$

c is acoustic velocity, and d_1, d_2 represent the distances from the base point to Mic.1 or Mic.2, respectively. Similarly, the output signal of Beam Former 2, W_{BF2} can be calculated by,

$$S_2(f) = X(f)W_{BF2}(f) \tag{4}$$

,where $W_{BF2}(f)$ represented the Beam Former 2 weights and is defined as,

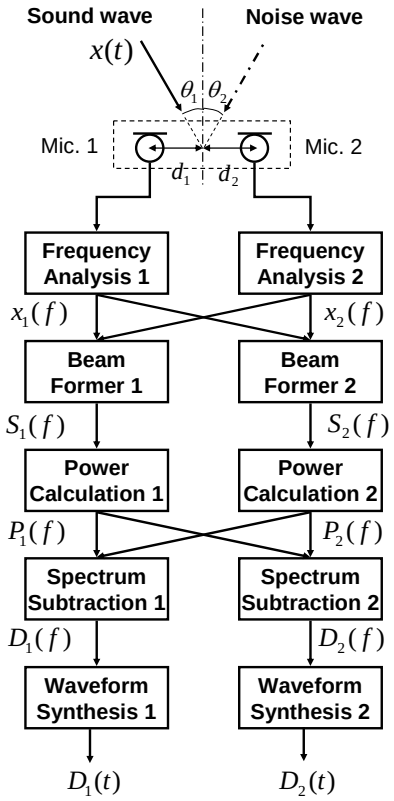


Fig. 1. Block diagram of Butterfly Subtraction Array (BSA)

$$W_{BF2}(f) = \begin{bmatrix} \exp[-j2\pi f d_1 \sin\theta_1 / c] & \exp[-j2\pi f d_2 \sin\theta_1 / c] \\ \exp[-j2\pi f d_1 \sin\theta_2 / c] & \exp[-j2\pi f d_2 \sin\theta_2 / c] \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{5}$$

According to (3) and (5), we capture a following relationship,

$$W_{BF1} = W_{BF2}^* \tag{6}$$

In the next stages, we perform "Power Calculation" and "Spectrum Subtraction". Given the Beam Former weights $W_{BF1}(f)$, $W_{BF2}(f)$ and the received signals $X(f)$, The "Power Calculation 1" output signal $P_1(f)$ and the "Power Calculation 2" output signal $P_2(f)$ are obtained as follows,

$$\begin{aligned} P_1(f) &= \left[X^T W_{BF1}^*\right]\left[X^T W_{BF1}^*\right]^H \\ &= X^T W_{BF1}^* W_{BF1}^T X^* \end{aligned} \tag{7}$$

$$P_2(f) = \left[X^T W_{BF2}^*\right]\left[X^T W_{BF2}^*\right]^H$$
$$= \left[X^T (W_{BF1}^*)^*\right]\left[X^T (W_{BF1}^*)^*\right]^H \quad (8)$$
$$= X^T W_{BF1} W_{BF1}^H X^*$$

and, the outputs, $D_1(f)$, $D_2(f)$ of "Spectrum Subtraction" can be calculated as follows,

$$D_1(f) = P_1(f) - P_2(f)$$
$$= X^T (W_{BF1}^* W_{BF1}^T - W_{BF1} W_{BF1}^H) X^* \quad (9)$$
$$= 2\,\mathrm{Re}\left[w_1^* w_2 x_1 x_2^*\right] - 2\,\mathrm{Re}\left[w_1 w_2^* x_1 x_2^*\right]$$

$$D_2(f) = P_2(f) - P_1(f)$$
$$= X^T (W_{BF1} W_{BF1}^H - W_{BF1}^* W_{BF1}^T) X^* \quad (10)$$
$$= 2\,\mathrm{Re}\left[w_2^* w_1 x_2 x_1^*\right] - 2\,\mathrm{Re}\left[w_2 w_1^* x_2 x_1^*\right]$$

We further introduce a parameter to represent variations of microphone gain characteristics. Making the one microphone gain α times as large as the other one is equivalence to making the one weight α times as large as the other one. And, we assume that the weight w_2 of Mic. 2 is α times as large as original weight w_{org} , while the weight w_1 of Mic. 1 is the same as the original weight w_{org} . Therefore,

$$w_1 = w_{org}$$
$$w_2 = \alpha\, w_{org} \quad (11)$$

By substituting (11) into (9), we get

$$D_1(f) = 2\,\mathrm{Re}\left[w_{org}^* (\alpha w_{org}) x_1 x_2^*\right] - 2\,\mathrm{Re}\left[w_{org} (\alpha w_{org})^* x_1 x_2^*\right]$$
$$= 2\alpha\left(\mathrm{Re}\left[w_{org}^* w_{org} x_1 x_2^*\right] - \mathrm{Re}\left[w_{org} w_{org}^* x_1 x_2^*\right]\right) \quad (12)$$

On the other hand,

$$D_1(f) = 2\alpha\left(\mathrm{Re}\left[w_{org}^* w_{org} x_2 x_1^*\right] - \mathrm{Re}\left[w_{org} w_{org}^* x_2 x_1^*\right]\right) \quad (13)$$

By varying the value of α , it only results in the change of the amplitudes of the outputs $D_1(f)$ and $D_2(f)$ in all frequency bins, but the directional patterns are not affected.

Fig. 2, and Fig. 3 are directional patterns of an example of Subtraction Beam Former (SBF) and BSA, respectively ,where (a), (b) are the case where microphone gains are the same. (c), (d) are the case where one microphone gain is 3dB larger than the other one. It is assumed that sampling frequency is 11025 [Hz], a number of microphone is 2, microphone distance is 30 [mm]. If microphone gains are the same, subtractive beam former accurately makes null point, and maintains desired performance. But if any gain differences among two microphones exist, the directional pattern's shape distorted to a remarkable degree. Briefly speaking, this method has no robustness to variation of microphone gain characteristic. In contrast, the BSA's directional pattern does not change even if two microphone gains are different. As is clear from Fig. 3, the incoming wave from $\theta = 0^\circ - 180^\circ$ is suppressed, while the incoming wave from $\theta = 180^\circ - 360^\circ$ is only extracted.

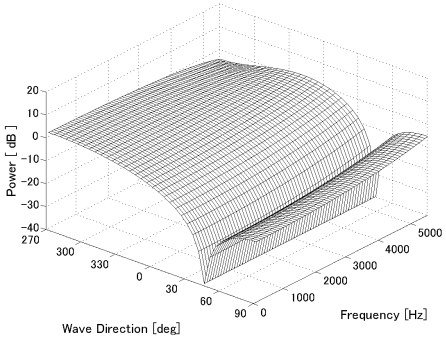
3. EXPERIMENTAL RESULT

To show the effectiveness of our method, we conducted a voice recognition experiment under an automotive environment. The experimental condition is shown in Table 1. In this experiment, we applied three methods to the double-talk situation of a driver and a fellow passenger, and then obtained voice recognition rate to each voices.

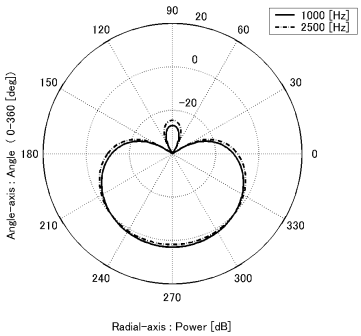
circumstance	automotive (idling & driving)
speaker	HATS
speaker position	driver – 45 [deg] fellow passenger – 315 [deg]
recognition voice	double-talk (driver & fellow passenger)
number of mic	2
mic distance	30 [mm]
mic position	near a map lamp
sampling rate	11025 [Hz]
method	Single microphone SBF (null : 45 or 315 [deg]) BSA (Proposed Method)

Table 1 Experimental condition

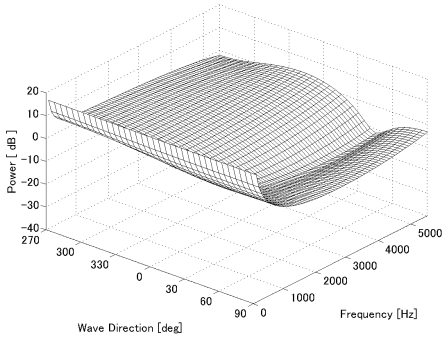
Fig. 4(a) shows one example of the original sound recorded in this experiment, which is composed of the drive's and the passenger's voices. Fig. 4(b) and Fig. 4(c) respectively show the driver's voice and the fellow passenger's voice separated by BSA. The experimental results are shown in Fig. 5 and Table 2. Fig. 5 shows voice recognition rate in idling car. Fig. 5(a) shows driver's voice recognition rate that extracted by each method, and Fig. 5 (b) is fellow passenger's voice recognition rate by each method. In addition, Table 2 is an average rate of voice recognition during idling and driving, respectively.



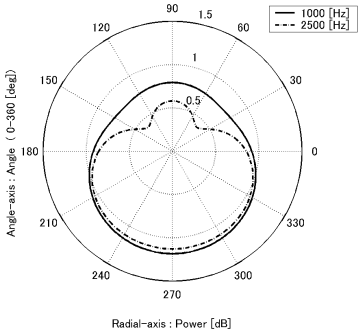
(a) directional pattern – 0 [dB]



(b) polar pattern – 0 [dB]

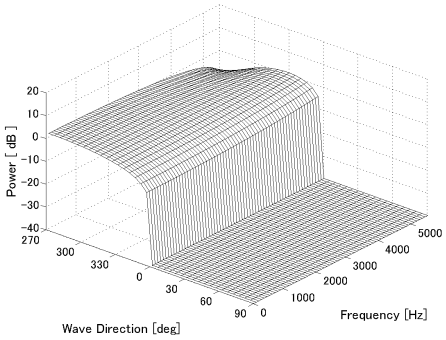


(c) directional pattern – 3 [dB]

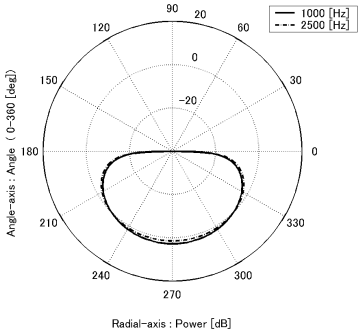


(d) polar pattern – 3 [dB]

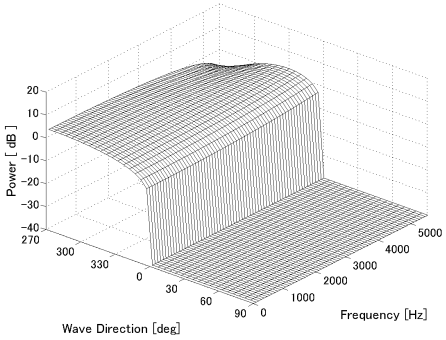
Fig. 2. Simulation at Subtraction Beam Former (SBF)



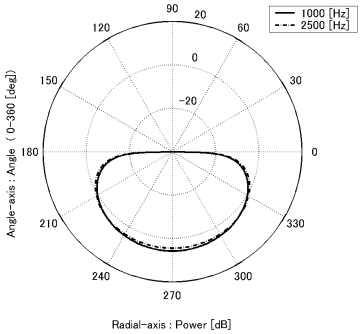
(a) directional pattern – 0 [dB]



(b) Polar pattern – 0 [dB]



(c) directional pattern – 3 [dB]



(d) Polar pattern – 3 [db]

Fig. 3. Simulation at BSA (Proposed Method

(a) Source Sound in Microphone

(b) detected voice of driver

(c) detected voice of passenger

Fig. 4. Example of BSA's results with experimental circumstance

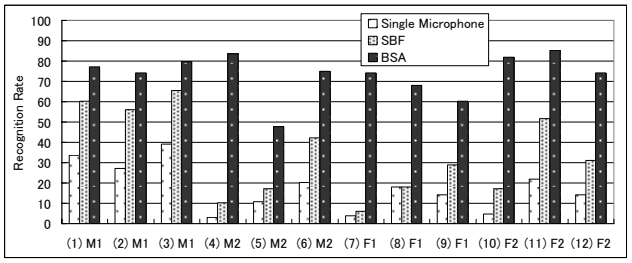
	Idling	Driving
Single Microphone	28.93 %	26.99 %
SBF	42.84 %	43.49 %
BSA (Proposed Method)	78.35 %	78.26 %

Table 2 Recognition rate under double-talk situation

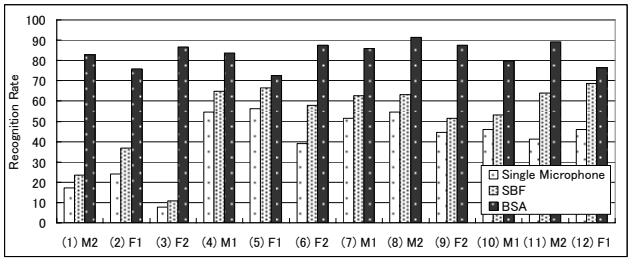
In conventional methods of single microphone and SBF, complete sound source separation was difficult under double-talk situation of a driver and a fellow passenger even if no aimed voices could be suppressed in part. Especially automotive environment, large early-reflected sound makes it more difficult. As our method matches automotive environment, complete sound source separation of the driver and the fellow passenger is possible. As a result, voice recognition rate of double-talk situation is dramatically improved for both idling and driving. As described above, our method is very effective in pre-processing stage of voice recognition system. However voice distortion arises as a result of adopting this method, we still have problem for hands-free conversation system.

4. CONCLUSIONS

In this paper, we proposed the sound source separation method which uses a pair of beam formers that has a configuration of complex conjugate. This method could maintain a stable performance for direction-lag of speaker position and large early-reflected



(a) for driver's seat



(b) for passenger's seat

Fig. 5. Recognition rate for each method under Table 1

sound by having spread directional pattern in moderation. We showed robustness to variation of microphone gain characteristic by simulation. And, by the result of voice recognition experiment in automotive environment, we showed the usefulness of our method.

5. REFERENCES

[1] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, Vol.78, No. 5, pp. 1508-1518, Nov. 1985.

[2] L. J. Griffiths, C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE. Trans.Antennas Propag.*, vol. AP-23, no. 1, pp. 27-34, Jan 1982.

[3] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, Vol.22, pp. 21-34, 1998.

[4] H. Saruwatari, T. Kawamura, and K. Shikano, "Blind source separation for speech based on fast-convergence algorithm with ICA and beamforming" *Proc. Eurospeech2001*, pp. 2603-2606, Sept. 2001.

[5] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency-domain blind source separation" *IEICE Trans. Fundamentals*, vol. E86-A, no. 3, pp. 590-595, March. 2003.

[6] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," *IEEE Trans. Acoust. Speech Signal Process.*, ASSP-34, pp. 1391-1400, 1986