Spot-forming method by using two shotgun microphones

Motoyuki Suzuki and Takeshi Honjo
Faculty of Information Science and Technology, Osaka Institute of Technology, Japan
E-mail: moto@m.ieice.org

***Abstract*—Shotgun microphones and microphone arrays have super-cardioid polar patterns with a main lobe called "beam." This pattern suppresses noise located outside the beam, but it cannot suppress noise originating from the same direction as the signal.**

In this paper, we have proposed a "spot-forming" method using two shotgun microphones. Two shotgun microphones are first located in parallel, and then one microphone is rotated in order to make the beams intersect. This system suppresses all but the sound signals originating from the cross-point, by using the delay-and-sum beamformer.

The experimental results showed that the proposed method could suppress noise located in front of or behind the target speaker, especially in the lower frequency band.

I. Introduction

Noise reduction technology is one of the most important issues for spoken dialog systems. For instance, autonomous robots are used in places where headsets and handheld microphones are not allowed. As distance between a user and a microphone increases, the Signal to Noise Ratio (SNR) becomes lower.

Many kinds of noise reduction methods have been proposed, such as noiseless speech recording, signal processing methods for speech enhancement, noise-robust feature extraction, and statistical modeling using speech data with noise. Noiseless speech recording technologies are easy to use in spoken dialog systems, because no modification is needed to the system. We only needed to exchange the microphone of the spoken dialog system for a noiseless recording system.

Microphone array methods (e.g. [1], [2], [3]) are frequently used for noiseless speech recording. In these systems, many microphones are used, and directivity is found by processing the many speech signals recorded by each microphone (the technique is called "beam-forming"). All noise signals located outside of the beam can be suppressed. Moreover, a speaker can move around while speaking because the direction of the beam can be changed easily by changing some parameters in the calculation. Microphone array methods are some of the most effective methods for noiseless recording.

However, they have the following disadvantages;

- Many microphones are needed. It is difficult to mount many microphones on small robots.
- Rich computational power is needed.
- Several of the methods are vulnerable to changes in the acoustic environment.

Another way to record noiseless speech uses directional microphones[4], [5]. Especially, a shotgun microphone is highly directional and easy to use. Compared to the microphone array system, it is difficult to trace the user's direction. However, this problem is not critical for communication robots, because it can be expected that a user stands in front of the robot begin spoken to. In particular, most robots in the RoboCup @ Home league[6] employed a shotgun microphone as an input device for spoken dialog system.

Microphone arrays and shotgun microphones are both highly directional. However, they have one critical problem. They cannot suppress a noise located behind the speaker. For example, if a television is located behind the targeted speaker, then signal has noise when it is input to the spoken dialog system, even though a shotgun microphone is employed. This causes deterioration of recognition performance.

In order to solve this problem, a new "spot-forming" method using two shotgun microphones has been proposed in this paper. This method makes the directivity a "spot," which means a region like a sphere. Beam-forming methods normally make directivity like a tube, but spot-forming method makes directivity like a sphere. Any noises behind the targeted speaker can be suppressed by adjusting the location of the "spot" to the mouth of the speaker. The new method requires slightly increased computational power in order to be used by small autonomous robots.

II. Spot-forming method

A. Basic idea

The basic idea of this method is very simple. A shotgun microphone has a directivity beam like a tube. Two shotgun microphones M_a and M_b are located in parallel, and then one is rotated in order for their directivity beams to intersect (Fig.1). When a sound source is located at the cross-point

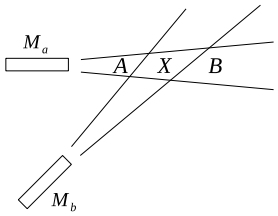

Fig. 1. Schematic view of Spot-forming

of the two directivity beams (region "X" in Fig.1), then both shotgun microphones will record it. On the other hand, when a sound source is located in any other region, then at least one microphone will not record it. Therefore, the sound signal located in the cross-point can be enhanced by adding the two recorded signals, while sound signals located in any other points are relatively suppressed. This idea is the same as a delay-and-sum beamformer[7]. Of course, the two recorded signals should be added with an appropriate delay time, because distances between the sound source and microphones are different.

If only M_a is used, then all sound signals included in the directivity beam of M_a (region "X," "A," and "B") are recorded. However, only sound signals from region "X" are also recorded by M_b. Therefore, sound signals from regions "A" and "B" can be suppressed. "Spot-forming" can be realized with this very simple idea.

Note that one of the most important issue is setting of two microphones. If directivity of a microphone becomes off to the side, then the cross-point disappears. Two microphones should be fixed firmly.

B. Measurement of delay-time

Delay-time depends on the difference of distances between the cross-point and each microphone. It can be calculated by measuring two distances, but measurement should be strict. For example, when the sampling frequency is set to 48 kHz, then a time lag of one sample ($\approx 20.8\mu$s) corresponds to only 7 mm ($= \frac{340\times1{,}000}{48{,}000}$). In general, a shotgun microphone consists of an acoustic tube and a microphone element. The distance cannot be measured strictly because the microphone element is included in an acoustic tube, and we cannot see it. Therefore, an estimation method of delay-time is proposed.

A loudspeaker is located in the cross-point, and then white noise is played through the loudspeaker. The parallel mounted shotgun microphones record the signal, and the delay-time is estimated by searching for the maximum point of the cross-correlation function between two signals.

Let $x_a(t)$ and $x_b(t)$ denote white noise recorded by the shotgun microphone M_a and M_b respectively, and $C_{x,y}(\tau)$ denotes cross-correlation function of functions $x(t)$ and $y(t)$. The delay-time $\hat{\tau}$ can be estimated by using the following equation:

$$\hat{\tau} = \mathrm{argmax}_{\tau} C_{x_a,x_b}(\tau) \tag{1}$$

If any other noise signal corrupts the recorded signals, then the estimation accuracy is decreased. Therefore, measurement of delay-time should be carried out in a silent environment.

At the same time, we also calculate the ratio of average power. Average power of the signal recorded by microphone M_a is different from that by M_b because of the different distances to the cross-point. The ratio of two average powers r is calculated by using Eq.2 in order to compensate power difference of recorded signals.

$$r = \frac{\frac{1}{N}\sum_{t}^{N}\{x_a(t)\}^2}{\frac{1}{N}\sum_{t}^{N}\{x_b(t)\}^2} \tag{2}$$

After estimating both $\hat{\tau}$ and r, the speech enhancing step can be carried out. In other words, the estimation step should be carried out in advance of recording any information. The enhanced signal $s(t)$ can be calculated by;

$$s(t) = \frac{1}{2}\left\{s_a(t) + \sqrt{r}\cdot s_b(t-\hat{\tau})\right\} \tag{3}$$

Note that $s_i(t)$ denotes the signal recorded by the shotgun microphone M_i.

In order to measure more exactly, a high sampling frequency (e.g. 96 kHz) is used. It is also used during the enhancing of speech step, and down-sampling is carried out after adding the two recorded signals.

III. Experiments

In order to investigate the effectiveness of the "spot-forming" method, several experiments were carried out.

A. Experimental setup

The shotgun microphone CS-3e (Sanken) was used in the experiments. Figure 2 shows the polar pattern of the microphone. It can be seen from this figure that the microphone has tight directivity for higher frequencies, but it has wider directivity for lower frequencies.

All experiments were carried out in a soundproof chamber. One microphone (M_a) is located at 1 m height, with horizontal directivity. The other microphone (M_b) is located at 0.2 m height, just under M_a. "Spot" was set to 1.0 m height (the same height as M_a) and 1.0 m away from M_a. Directivity

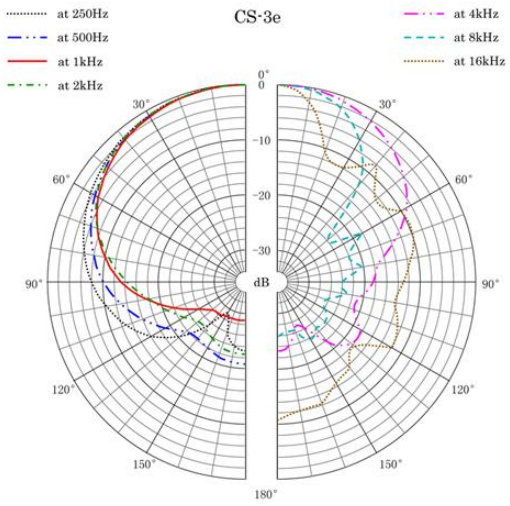

Fig. 2. Polar pattern of shotgun microphone (CS-3e)[8]

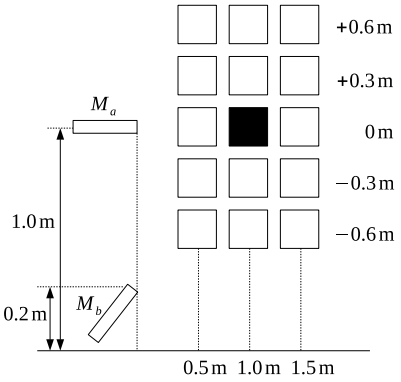

Fig. 3. Measurement positions

of M_b was pointed at the "spot." A loudspeaker MSP3 (YAMAHA) was used, and Quad-Capture UA-55 (Roland) was used as an A/D converter. Sampling rate was set to 96 kHz.

B. Polar patterns of the method

The polar pattern of the "spot-forming" method was measured. White noise was played through a loudspeaker that was relocated to 15 different positions. Figure 3 shows the layout of these positions. In this figure, the black square indicates the targeted "spot." After recordings were made, the power spectrum was calculated, and the value of each position was normalized by the value of "spot" position for several frequencies.

Figure 4 shows normalized power value (dB) for 375 Hz $\sim$ 3 kHz. From these figures, it can be seen that the sound signal from the "spot" was enhanced. Also, the position located at 0.5 m on the x-axis and 0 m on the y-axis (left side, center height) was highly suppressed in every frequency, even though this position is the closest to the microphone M_a. This means that the signal from that position was recorded loudly by M_a. However, the position was outside of the directivity of M_b, and delay-time was bigger (the position is close to M_a, but relatively far from M_b). As a result, the signal was not enhanced.

On the other hand, the position closed to M_b (0.5 m on the x-axis, -0.3 m on the y-axis) was not suppressed. In this position, M_b recorded a sound signal loudly, but M_a did not. This is the same situation as the expression above. However, in general, power ratio r is bigger than 1.0. This means that any sound signal recorded by M_b is enhanced by a compensation of signal power, causing enhancement of the sound signal from this position.

C. Noise suppression performance compared with one shotgun microphone

In this section, we have checked the noise suppression performance compared with one shotgun microphone. First, the polar pattern of the microphone M_a was calculated by the same way described in Sec.III-B. After that, each power value was subtracted from the corresponding power value given by the microphone array (shown in Fig.4) in order to show the difference in noise suppression performance.

Figure 5 shows the results. In these figures, a negative value means that "spot-forming" can suppress noise more effectively than "beam-forming." From these figures, it can be seen that "spot-forming" can suppress noise effectively near the microphone (0.5 m on the x-axis). "Spot-forming" only records sound signals in the "spot." In other words, sound signals originating in front of and behind the "spot" are not recorded. On the other hand, a shotgun microphone records all sound signals located in the directivity beam. Thus, the sound signal located in front of M_a is recorded in the "beam-forming," but not recorded in the "spot-forming." This results in very effective suppression in the near region of the microphone.

Unfortunately, "spot-forming" could not suppress noise in upper and lower regions of the microphone, especially for

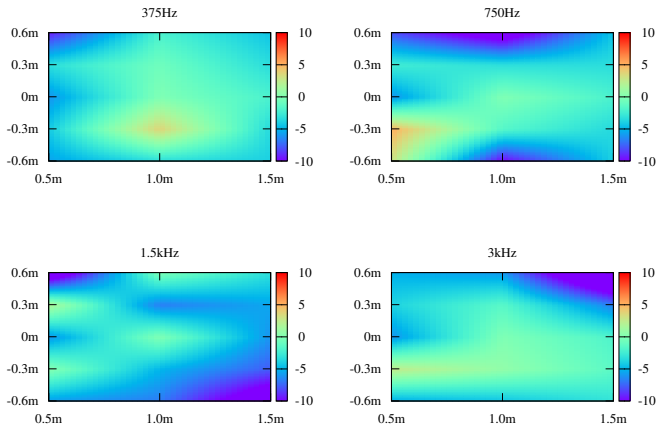

Fig. 4. Polar pattern

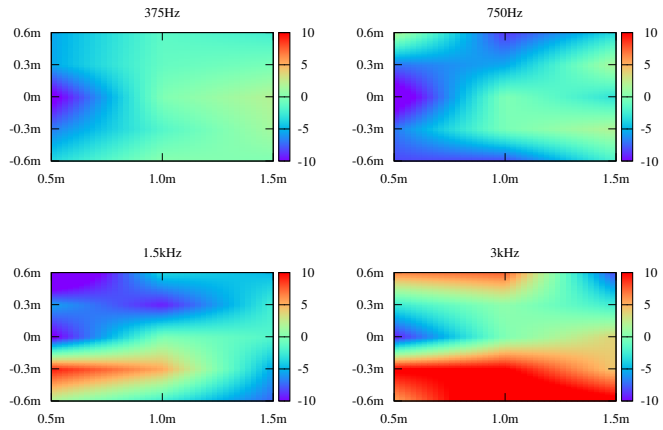

Fig. 5. Noise suppression performance

higher frequencies. The reasons are:

- As described in Sec.III-B, sound signals in front of M_b are recorded loudly
- For frequencies of which (multiples of) period time is the same as the delay-time, the signals are added even if they are located outside of the "spot." This means that a delay-and-sum beamformer will also enhance it. On the other hand, a shotgun microphone has sharp directivity for higher frequencies.

In conclusion, the proposed method can make "spot-forming," but decreases noise suppression performance compared with a shotgun microphone for higher frequencies.

D. SNR improvement for speech

In this section, SNR was measured using a speech signal. The two microphones were the same as the previous experiment, including location.

Two loudspeakers were used. One was located at "spot" position, and played the signal speech. The other was used as a noise signal, and located in four positions ("a," "b," "c" and "d" in Fig.6). In this figure, a black square indicates the loudspeaker for the signal speech, and a white square indicates the loudspeaker for noise. The x-axis of the noise was set to 1.5 m, and the y-axis was set to 0 m (just behind the "spot") and ± 0.3 m, and shift to 0.7 m right position (x=1.5 m, y=0 m, and z=0.7 m). A lecture speech[9], [10] was used as the target signal, and noise recorded in an exhibition hall[11] was used as the noise signal. This noise signal consists of a number of voices and high reverberation. The volume of the noise signal was set to three different levels, and SNR was calculated for each.

Average SNR and improvement are shown in Tbl.I. From this table, the proposed method improved SNR in all positions. This proves that "spot-forming" can improve SNR in "just behind" position. It means that the method is a success. On the other hand, improvement in the upper position was lower than in other positions, because the noise loudspeaker was located into the directivity of M_b.

IV. Conclusion

Either a shotgun microphone or a microphone array can suppress all noises located outside the directivity, but they cannot suppress noises originating from the same direction as the target signal. In this paper, we have proposed a "spot-forming" method by using two shotgun microphones.

Two shotgun microphones are mounted in parallel, and one microphone is rotated in order to make an intersection of the two beams. After recording, one of the recorded signals is shifted in time-dimension and average power is normalized in order to equalize the signals, and then the signals are added. This step is the same as the delay-and-sum beamformer. A shotgun microphone has "beam" directivity, and the proposed method enhances a sound signal located in the cross-point of two "beams."

The experimental results showed that the proposed method could suppress noise located near the upper microphone better than a shotgun microphone. On the other hand, noise signal near the lower microphone was not suppressed well. Moreover, suppression performance was decreased for higher frequencies.

We also evaluated performance using a speech signal. A target signal was located at "spot" and a noise signal, which was recorded in an exhibition hall, was played behind the target signal. SNR of the proposed method improved 1.15 dB (about 10.3%) compared with SNR of a shotgun microphone.

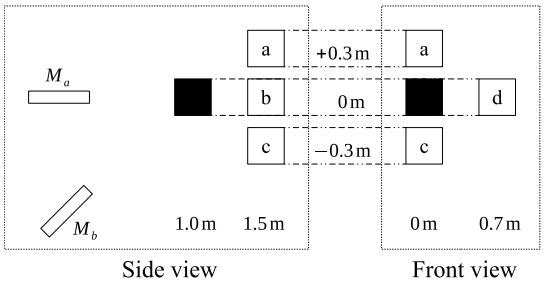

Fig. 6. Speaker settings

TABLE I
Average SNR and improvement

Noise position	a	b	c	d
Beam-forming	11.26 dB	11.14 dB	13.06 dB	13.05 dB
Spot-forming	12.08 dB	12.29 dB	14.24 dB	14.12 dB
Improvement (ratio)	+0.82 dB (+7.28%)	+1.15 dB (+10.3%)	+1.18 dB (+9.04%)	+1.07 dB (+8.20%)

References

[1] M. Omologo, M. Matassoni, P. Svaizer, and D. Giuliani, "Microphone array based speech recognition with different talker-array positions," in *Proc. ICASSP'97*, 1997, pp. 227–230.

[2] J. Adcock, Y. Gotoh, D. Mashao, and H. Silverman, "Microphone-array speech recognition via incremental MAP training," in *Proc. ICASSP'96*, 1996, pp. 897–900.

[3] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas and Propagation*, vol. AP-30, no. 1, pp. 27–34, 1982.

[4] J. Gustafson, N. Lindberg, and M. Lundeberg, "The august spoken dialogue system," in *Proc. EUROSPEECH*, vol. 3, 1999, pp. 1151–1154.

[5] K. Kayama, A. Kobayashi, E. Mizukami, T. Misu, H. Koshioka, H. Kawai, and S. Nakamura, "Spoken dialog system on plasma display panel estimating users' interest by image processing," in *Proc. 6th International Conference on Intelligent Environments*, 2010, pp. 4–13.

[6] The RoboCup Federation, "RoboCup," 1997, http://www.robocup.org/.

[7] J. L. Flanagan, J. D. Johnston, R. Zahn, and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *The Journal of the Acoustical Society of America*, vol. 78, no. 5, pp. 1508–1518, 1985.

[8] SANKEN MICROPHONE Co., LTD, "Polar pattern (CS-3e)." [Online]. Available: http://www.sanken-mic.com/product/freqpola.cfm/8.5001500

[9] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proc. Second International Conference on Language Resources and Evaluation (LREC)*, 2000, pp. 947–952.

[10] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, 2003.

[11] S. Itahashi, "A noise database and Japanese common speech data corpus," *The Journal of the Acousitc Society of Japan*, vol. 47, no. 12, pp. 951–953, 1991, (in Japanese).