SPEECH ENHANCEMENT BY MULTIPLE BEAMFORMING WITH REFLECTION SIGNAL EQUALIZATION

Takanobu Nishiura[†‡], *Satoshi Nakamura*[†], *and Kiyohiro Shikano*[‡]

[†] ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto, 619-0288 Japan
[‡] Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara, 630-0101 Japan

1. ABSTRACT

In real environments, the presence of room reverberations seriously degrades the quality in sound capture. To solve this problem, multiple beamforming [1], which forms directivity not only in the direction of the desired sound source but also in the direction of reflection images, was proposed by J. Flanagan et al. However, it is difficult to apply this method practically in real environments, since this application requires that the distortion of reflection sound signals by wall impedances be equalized. This paper proposes a new multiple beamforming algorithm that equalizes the amplitude-spectrum and phase-spectrum of reflection signals by a cross-spectrum [2] method. Evaluation experiments are conducted in real environment. In a SDR (Signal to Distortion Ratio) evaluation, the proposed multiple beamformer achieves signal distortion reduction more effectively than the conventional single beamformer and the conventional multiple beamformer without equalization. In addition, in an ASR (Automatic Speech Recognition) evaluation, the equalized multiple beamformer achieves a higher recognition performance than those of the above conventional beamformers.

2. INTRODUCTION

For teleconference systems or voice control systems, the high-quality sound capture of distant talking speech is very important. However, background noise and room reverberations seriously degrade the sound capture quality in real acoustical environments. A microphone array has been applied as one of the promising tools to deal with this problem. With the microphone array, a desired speech signal can be acquired selectively by steering the directivity in the desired speech direction sensitively.

This shows that it is necessary for super directivity to reduce noise signals. Delay-and-sum beamformers [3, 4] and adaptive beamformers [5, 6] were proposed as conventional single beamformers with a microphone array. However, adaptive beamformers with null steering have had a serious problem, i.e., the capturing performance of the desired speech signal is degraded in high reverberant rooms, because of limitations with the amount of null steering. Delay-and-sum beamformers have also had a serious problem i.e., the performance is not enough to capture the desired speech signal with the insufficient number of transducers and the existence of high reverberant rooms.

To solve this problem, we focus on utilizing the reflection signals that cause the reverberations. When a sound signal is reflected by walls, and so on, the quality of the reflection signals is distorted and attenuated by wall impedances. The distorted reflection signals then reach transducers after the direct signal arrival. The quality of some of the signals arriving at the transducers are highly degraded, especially with a lot of wall reflections. However, low-order reflection signals can be highly correlated to the direct signal. Accordingly, low-order reflection signals can be utilized efficiently as the desired signal.

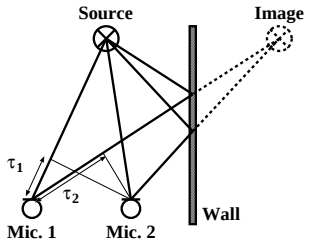

Figure 1: A direct source and a reflection image.

To improve the performance of the delay-and-sum beamformer, a multiple beamformer [1] that utilizes low-order reflection signals was proposed by Flanagan et al. Unlike this beamformer, the conventional multiple beamformer does not consider the degradation of reflection signals by wall impedances. Therefore, although the conventional multiple beamformer is effective in computer simulation experiments without any consideration on signal degradation by wall impedances, this beamformer will not be effective at all in real environments where such wall impedances exist. Furthermore, when the conventional multiple beamformer is applied to ASR, it is necessary to equalize the distorted reflection signals. To cope with this problem, this paper describes a new multiple beamforming algorithm that is effective for ASR in real environments.

3. CONVENTIONAL MULTIPLE BEAMFORMING

Until now, the reflection signals of the desired signal had been dealt with as “noise signals”. However, since low-order reflection signals are highly correlated to the desired direct signal, a multiple beamformer that utilizes low-order reflection signals was proposed by Flanagan et al. Figure 1 shows an example. There are two transducers and one sound source. In this situation, the first-order reflection image exists on a mirror position of the direct sound source across a wall [7]. If the wall impedances can

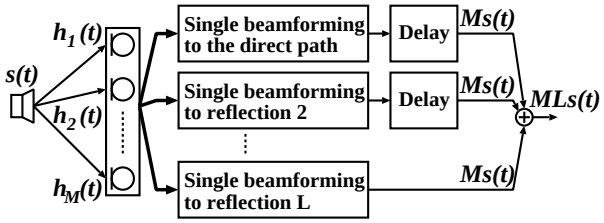

Figure 2: Multiple beamforming.

be ignored, the reflection signals can be considered equal to the direct signal with a short time delay. Therefore, the conventional multiple beamformer creates directivity not only in the direction of the desired sound source but also in the direction of reflection sound images. Figure 2 shows an overview of the conventional multiple beamforming algorithm. Impulse response $h_m(t)$ from the desired direct sound source to the mth transducer is shown as Equation (1) when one direct sound source and $L-1$ reflection sound images are received by the M transducers in the far field without any sound attenuation.

$$h_m(t) = \sum_{l=1}^{L} \delta(t - \tau_{l,m}), \tag{1}$$

where l is the signal number (if $l = 1$, this shows the direct sound signal., if $l = 2, 3, \cdots, L$, this shows a reflection sound signal) and $\tau_{l,m}$ represents the time delay for the lth signal to arrive at the mth transducer. Captured signal $x_m(t)$ by the mth transducer is shown as Equation (2).

$$x_m(t) = s(t) * h_m(t), \tag{2}$$

where $s(t)$ is the desired signal and the symbol $'*'$ represents convolution. Accordingly, output signal $y_l(t)$ by the delay-and-sum beamformer for the lth signal is shown by Equation (3).

$$y_l(t) = \sum_{m=1}^{M} x_m(t) * \delta(t + \tau_{l,m}). \tag{3}$$

Output signal $y(t)$ with multiple beamforming is shown by Equation (4).

$$\begin{aligned} y(t) &= \sum_{l=1}^{L} y_l(t) \\ &= \sum_{l=1}^{L}\sum_{m=1}^{M}\left\{ s(t) * \sum_{l'=1}^{L} \delta(t - \tau_{l',m}) * \delta(t + \tau_{l,m}) \right\}. \end{aligned} \tag{4}$$

Desired sound signal $s(t)$ is LM times as large as signal $x_m(t)$ by Equation (4). On the other hand, noise signal $s(t)$ is not LM times as large as signal $x_m(t)$ because the directions of noise signals are different from the direction of the desired sound signal. As a result, the multiple beamformer performs more effectively than a single beamformer if we do not consider wall impedances. However, this consideration is a necessity in real environments. Therefore, it is indispensable to equalize reflection signals in real environments. To cope with this problem, this paper describes a new multiple beamforming algorithm with the equalization of reflection signals.

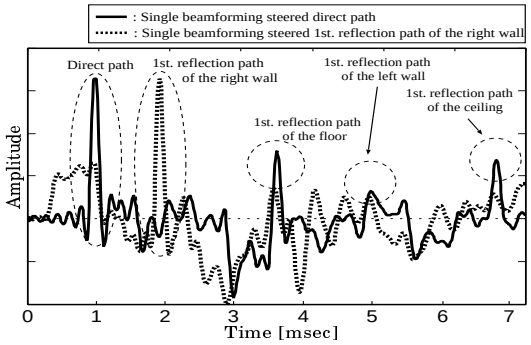

Figure 3: Single beamforming for each signal.

4. PROPOSED MULTIPLE BEAMFORMING

4.1. Equalization of reflection signals

Basically speaking, reflection signals are distorted seriously by wall impedances in real environments. With the conventional multiple beamforming the output signal is also distorted. As a result, a multiple beamformer can not perform more effectively than a single beamformer. However, reflection signals can be utilized as the desired signal if their amplitude-spectrum and phase-spectrum can be equalized.

In this paper, we consider multiple beamforming that utilizes reflection signals after equalizing distorted reflection signals by wall impedances. For the reflection signal equalization, equalization filters are designed for the distorted reflection signals. These filters equalize the reflection signals. As a result, the amplitude-spectrum and phase-spectrum of the reflection signals can be adjusted to the direct signal.

4.2. Equalization filter

Next, we describe the method of designing the equalization filters. In real environments, impulse responses are measured by time-stretched pulse signals (TSP) [8, 9]. Then, the directivity of the beamformer is steered to the direct sound source direction and the reflection sound image direction as the desired signal with single beamforming. Figure 3 shows an example. In the figure, the first-order reflection signal by the right wall is focused to be utilized as the reflection signal.

The direct signal must not be distorted by wall impedances. Therefore, we try to equalize distorted reflection signals with the direct signal. The filters for equalizing the distorted reflection signals are designed by estimating the transfer function between output signals, with single beamforming for the direct sound source and for reflection sound images. However, if we assume output signals with single beamforming for the direct sound as the target signal, it is difficult to design the equalization filters accurately. This is because the target signal can also include undesired signals besides the direct sound signal. Therefore, as shown in Figure 3, we extract a $0.5msec.$ windowed signal surrounding the signal peak. Then, the equalization filters are designed by estimating the transfer function between the cut output signals of the single beamforming for the direct sound and for reflection sound images.

In this paper, the transfer function is estimated by a cross-spectrum [2] method. The spectrum $Y(k)$ is defined as the discrete Fourier transform (DFT) of the output signal with single beam-

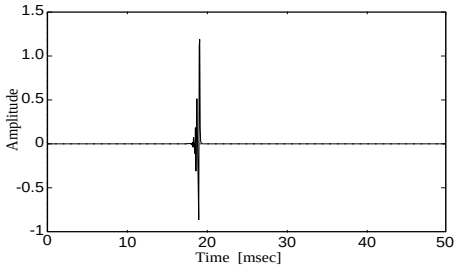

Figure 4: Equalization filter.

forming for the direct sound. Similarly, the spectrum $X(k)$ is defined as the DFT of the output signal with single beamforming for the reflection images. The equalizing filter (the transfer function) $f(n)$ is calculated by Equation (5).

$$f(n) = DFT^{-1}\left(\frac{Y(k)\cdot X(k)^*}{X(k)\cdot X(k)^*}\right), \qquad (5)$$

where the symbol $'*'$ represents the complex conjugate. Figure 4 shows an example of the equalization filter coefficient $f(n)$ in the time domain. The equalization filter is designed to add a $20msec.$ delay to each output signal for causality in Figure 4. A cross-spectrum method is one of the most effective methods for estimating the transfer function. The amplitude-spectrum and phase-spectrum of reflection signals can be equalized accurately by this technique. This equalization technique can not only equalize the amplitude-spectrum and phase-spectrum of reflection signals but also the time delay between the direct signal and reflection signals. As a result, the synchronization of each single beamformer can be carried out easily. We try to achieve an effective multiple beamformer with this algorithm in real environments.

5. EVALUATION EXPERIMENT

5.1. Experimental conditions

An experiment on the proposed multiple beamformer is conducted to evaluate its signal capturing performance and ASR (Automatic Speech Recognition) performance in a real room ($[T_{60}] = 0.64$). Table 1 shows the recording conditions, microphone array type, and ASR experimental conditions. The experimental room is shown as Figure 5. A multiple beamformer is utilized for first-order reflection sound images by the right wall. The talker is placed at points A$\sim$F.

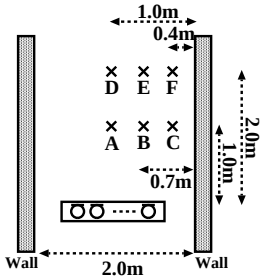

Figure 5: Experiment room.

Table 1: Evaluation experiment conditions.

Reverberation time $[T_{60}]$	0.64 sec.
Ambient noise level	49.3 dBA
Sampling frequency	12 kHz
Temperature	$19.5^{\circ}C$
Microphone array	Linear type, 14 transducers
Frame length (shift)	32 msec. (8 msec.)
Feature vectors	$MFCC, \Delta MFCC, \Delta power$
HMM	Tied-mixture
Number of models	54 phoneme models
Training data	Speaker dependent 5240 words
Test data	Isolated 216 words

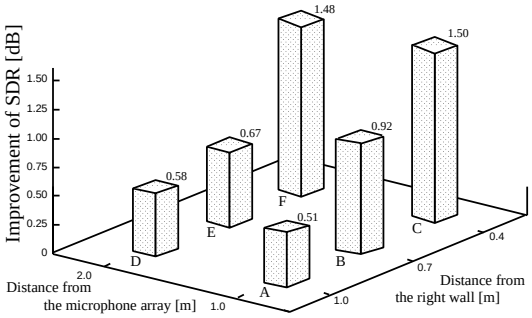

Figure 6: $SDR_{(ipr)}$ with multiple beamforming.

5.2. Evaluation of signal capturing performance

The SDR of the each signal for evaluation is calculated as shown in Equation (6).

$$SDR = 10log_{10}\frac{\sum_n (s(n))^2}{\sum_n (s(n)-\beta\hat{s}(n))^2}\ \ [dB], \qquad (6)$$

where $s(n)$ is the original signal, $\hat{s}(n)$ is the evaluation signal, and β represents the coefficient to equalize the amplitude between $s(n)$ and $\hat{s}(n)$. The SDR improvement $SDR_{(ipr)}$ is calculated by Equation (7) after Equation (6).

$$SDR_{(ipr)} = SDR_{(MBF)} - SDR_{(SBF)}\ \ [dB], \qquad (7)$$

where $SDR_{(MBF)}$ and $SDR_{(SBF)}$ represent the SDR of the output signal with multiple beamforming and single beamforming for direct sound, respectively. The signal capturing performance of the proposed multiple beamforming is evaluated by the SDR improvement, $SDR_{(ipr)}$, shown by Equation (7). As a result of this evaluation experiment, Figure 6 shows $SDR_{(ipr)}$ when white Gaussian noise is generated from each talker position. The proposed multiple beamforming performs more effectively than the single beamforming for direct sound because $SDR_{(ipr)}$ shows positive values in Figure 6. However, the best $SDR_{(ipr)}$ is only about $1.5dB$ in this evaluation experiment.

To investigate these results, the impulse response is measured by TSP in talker position B. Figure 7 shows the output signal with each beamforming for the desired impulse signal. As shown in this figure, single beamforming and multiple beamforming can suppress the reflection signals by the right and left walls although the signal of the single transducer can not. We can also confirm that undesired signals exist $0.8 \sim 1.3msec.$ before the desired impulse signal with multiple beamforming in Figure 7. These undesired

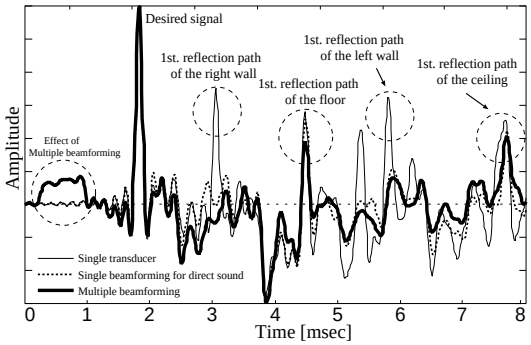

Figure 7: Multiple beamforming with impulse responses.

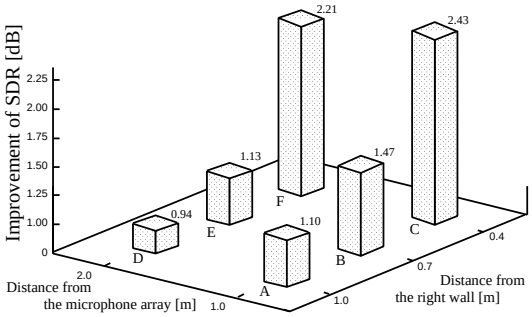

Figure 8: $SDR_{(ipr)}$ for the output signal excluding residual signals with multiple beamforming.

signals are the residual signals by the multiple beamforming. In particular, these signal are direct signals unable to be suppressed when the directivity of the single beamformer is steered to a reflection signal. Therefore, $SDR_{(ipr)}$ is re-calculated after cutting off this residual signal. As a result, the best $SDR_{(ipr)}$ is about $2.4dB$ in this evaluation experiment as shown in Figure 8. Therefore, it is necessary to form a sharper directivity by increasing the number of transducers in order to suppress undesired signals.

5.3. Evaluation of ASR performance

To evaluate the ASR performance with the proposed multiple beamforming, a Japanese isolated word recognition experiment is conducted. The word recognition rate (WRR) is 98.7% for clean data. Next, an ASR experiment is conducted with the speech uttered at talker position B, which is 0.7 meters from the right wall. Figure 9 shows the results. In Figure 9, when one transducer is used, WRR first degrades from 98.7% to 69.4% by the effect of reverberations, and so on. Then, WRR improves from 69.4% to 74.5% with single beamforming for the direct sound. On the other hand, WRR degrades in comparison with the single beamforming for the direct sound from 74.5% to 71.8% with the conventional multiple beamforming (without equalization). However, WRR of the proposed multiple beamforming (with equalization) improves in comparison with the single beamforming for the direct sound from 74.5% to 79.2%. At talker position E, the ASR performance is also improved identical to the proposed multiple beamforming. We could therefore confirm that the ASR performance is improved by the proposed multiple beamforming utilizing reflection signal equalization.

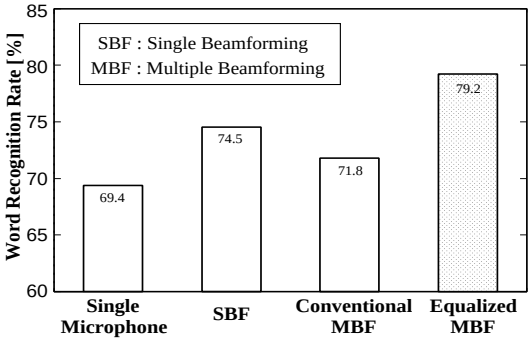

Figure 9: Word recognition results (Talker position B).

6. CONCLUSION

This paper describes a new multiple beamforming algorithm for ASR. The proposed multiple beamformer utilizes reflection signals by wall impedances after equalizing the signal. As evaluation experimental results, we confirm that the proposed multiple beamformer can improve SDR and WRR compared with conventional beamformers. In future work, it will be necessary to utilize not only first-order reflection signals but also second- and third-order reflection signals.

7. REFERENCES

[1] J.L. Flanagan, A.C. Surendran, and E.E. Jan, "Spatially selective sound capture for speech and audio processing," Speech Communication, Vol. 13, pp. 207–222, 1993.

[2] S. Lawrence Mariple Jr., "Digital Spectral Analysis with Applications," Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1987.

[3] 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.

[4] S.U. Pillai, "Array Signal Processing," Springer-Verlag, New York, 1989.

[5] L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beam-forming," IEEE Trans. AP, Vol. 30, No. 1, pp. 27–34, 1982.

[6] Y. Kaneda and J. Ohga, "Adaptive microphone-array system for noise reduction," IEEE Trans. ASSP, Vol. 34, No.6, pp. 1391–1400, Dec. 1986.

[7] J.B. Allen and D.A. Berkley, "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am., Vol. 65, No. 4, pp. 943–950, 1979.

[8] A.J. Berkhout, D. de vries, and M.M. Boone, "A new method to acquire impulse responses in concert halls," J. Acoust. Soc. Am., Vol. 68, pp. 179–183, 1980.

[9] Y. Suzuki, F. Asano, H.Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," J. Acoust. Soc. Am., Vol. 97, No. 2, pp. 1119-1123, 1995.