A MICROPHONE ARRAY SYSTEM FOR SPEECH RECOGNITION

Kenji Kiyohara, Yutaka Kaneda,* Satoshi Takahashi,** Hiroaki Nomura,* and Junji Kojima**

NTT Human Interface Laboratories
* 3-9-11, Midori-cho, Musashino-shi, Tokyo, 180 Japan
** 1-2356, Take, Yokosuka-shi, Kanagawa, 238-03 Japan
kiyohara@splab.hil.ntt.co.jp

ABSTRACT

This paper proposes a microphone array system which realizes the following important functions for speech recognition: i) SNR improvement, ii) flat spectrum response for an arbitrary speaker position, and iii) speech period detection in noisy speech. This microphone array system features time delay estimation using prewhitening signal processing, an optimally weighted delay-and-sum array, and speech period detection (called MLD) based on the level difference between signals before and after array processing. Word recognition experiments performed in the presence of crowd noise demonstrate that the proposed system has great robustness against noise than does the system with a conventional directional microphone and a speech period detection method.

1. INTRODUCTION

One major problem with putting speech-recognition systems into practical use is contending with environmental noise, which degrades the recognition rate. Directional microphone and noise-reduction signal-processing have been used to improve the signal-to-noise ratio (SNR) of the acquired speech. However, these conventional techniques cannot improve the SNR sufficiently. Recently, microphone arrays have been investigated [1], [2]. Microphone array improves the SNR much more than conventional directional microphones, so is considered to be a possible solution to the problem.

This paper proposes a new microphone array system for speech recognition. This system uses array processing to improve SNR and for speech period detection.

2. SYSTEM CONFIGURATION

Figure 1 shows a block diagram of the system. The microphone array consists of M elements. A Time Delay Estimator (TDE) estimates the time delay differences between the signal acquired by microphone M_0 and those acquired by M_i. A Time Delay Compensator (TDC) compensates for the estimated time delay differences. The compensated signals are weighted optimally and summed (OW). The summed signal and output of microphone M_0 are fed to the Speech Period Detector (SPD) so that the speech period can be determined. The detected speech signal is sent to the Speech Recognizer (SR).

3. ESTIMATION OF TIME DELAY DIFFERENCE

TDE estimates the time delay (time difference) between direct sounds received by the i-th microphone M_i and the reference one M_0. There are several available estimation methods based on cross-correlation peak (CC) [3], cross-correlation of prewhitened signals (PW) [4],

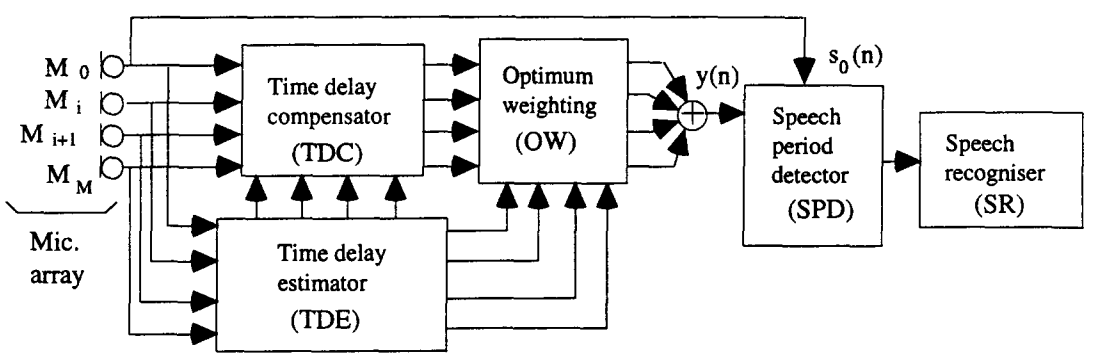

Figure 1. Block diagram of system.

and crosspower spectrum phase (CSP) [2].

In the CC method, the time delay is estimated such that it gives the maximum value of the cross-correlation function of objective and reference signals. In the PW method, acquired signals are first passed through prewhitening filters. These prewhitening filters are low-order linear prediction filters, and have the same characteristics for all channels. Time delays are estimated based on the cross-correlation peaks using these prewhitened signals (called residual signals).

CSP is calculated as [2]

$$\phi_{ik}(f) = \frac{S_i(f)S_k(f)^*}{|S_i(f)S_k(f)|} , \quad (1)$$

where $S_i(f)$ and $S_k(f)$ are Fourier Transforms of signals acquired through the i-th and k-th microphones. Time delay is estimated from the peak of the inverse Fourier Transform of $\phi_{ik}(f)$. The CSP method is regarded as that it whitens acquired signals in the frequency domain.

Later, this paper compares these methods using sound data in a real environment.

4. DELAY-AND-SUM ARRAY

4.1. Array configuration

In conventional systems, linear arrays of microphones have been widely used because they are easy to design and do not require many elements. However, they do not achieve sufficient spatial selectivity, since their directivity is axisymmetric. A two-dimensional array [5] can avoid this axisymmetric directivity. So, a two-dimensional array attached to a wall is adopted in our system.

Figure 2 shows a block diagram of the delay-and-sum array processing. Time delay differences among desired signals are compensated by delays $D_1,\cdots,D_M$. Time aligned signals are summed to improve the SNR. At this point, proper signal weighting by g_i can further improve the SNR.

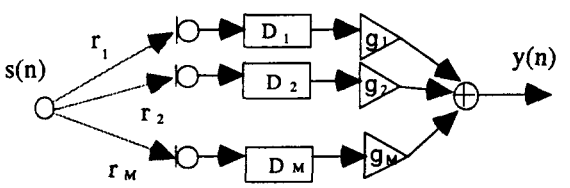

Figure 2. Delay-and-sum array.

4.2. Optimum weighting coefficients

Assuming that acquired noises are uncorrelated among array elements and have equal power, the SNR of the array output can be expressed as follows [6]:

$$SNR \propto \frac{(g^T a)^2}{g^T g} \quad (2)$$

$g = [g_1, g_2, \cdots, g_M]^T$
$a = [1/r_1, 1/r_2, \cdots, 1/r_M]^T$,
where r_i are the distances of the microphones from the source position.

Partially differentiating (2) with respect to g and setting it equal to zero gives the coefficient vector g'_0 that maximizes SNR as $g'_0 = k_0\, a$ (k_0: scalar).

To ensure equal sensitivity for all source positions, optimum gain g_0 is derived by normalizing g'_0 as

$$g_0 = \frac{k_0 a}{\sqrt{(a^T a)^2 + \frac{a^T a}{r_C^2}}} , \quad (3)$$

where r_C represents the reverberation distance (critical distance) at which direct and reverberant sound energy densities are equal [7]. When the distance r_i is unknown, the root power ratio of the acquired signals is used instead.

5. SPEECH PERIOD DETECTION

Even when the noise is reduced by the delay-and-sum array processing, speech period detection is still a problem especially when the noise is nonstationary. To solve the problem, a speech period detection method based on the level difference between before and after array processing (called MLD: Multi-channel Level Difference) [8] is adopted.

The process has two steps.

1. Detection of speech period candidates: The candidates are detected in the array output as the periods with high power.
2. Testing the candidates by the MLD criterion: First, the short-time power levels of an input and the output of the array $s_0(n)$ and $y(n)$ (see Figure 1) called P_s and P_y are calculated. Since the array processing suppresses undesired noise, but not desired speech, the period is considered to be a noise period when $P_s - P_y > P_{th}$ (P_{th} : a certain threshold value). The period is accepted as a speech period when $P_s \approx P_y$.

Since the directivity of the array is narrower at high frequencies, speech periods are better detected with high-pass filtered signals.

6. SPEECH RECOGNITION EXPERIMENTS

Recognition experiments were performed using one hundred words (Japanese city names). A two-dimensional microphone array composed of 15 omni-directional microphones, as shown in Figure 3, was used. It was attached to one wall of the experimental room, as shown in Figure 4. The reverberation time T_{60} of the room was about 0.3 s. Words to be recognized were uttered through a mouth simulator (MS) 0.5 m away from the array. The frequency characteristics of the MS were compensated by filtering the words with the inverse characteristics of the MS in advance. The words of one female speaker were used. Recorded crowd noise (nonstationary noise) was emitted through four loudspeakers located at and facing into the four corners of the room.

The speech recogniser used speaker-independent context-dependent HMMs. The models were trained using 118,720 utterances from 84 speakers recorded in a sound absorbent room. All speech was sampled at 12 kHz. The frame size for analysis was 32 ms and frames were analyzed every 8 ms. The feature parameters were 16 LPC cepstral coefficients, 16 delta cepstral coefficients, and delta log power.

The speech and noise level to determine SNR was measured by the speech voltmeter [9]. Noise-free results are also shown for reference. MS was set at two positions: directly in front of the array and 0.4 m to one side, as shown in Figure 4. The recognition results were compared with those obtained using other types of conventional microphone: an omni- and a uni-directional microphone.

7. RESULTS

7.1. Time delay estimation

The peak position of the cross-correlation function of the acquired speech signals was strongly affected by additive noise. This caused incorrect delay estimation. On the other hand, prewhitening the signals made the correlation peak sharper. Therefore, the prewhitening (PW) method gave much better results than the simple cross-correlation (CC) method. However, the PW and CSP methods did not show much difference under these experimental conditions. PW was used in the experiments below.

7.2. Array response to signal and noise

Figure 5 shows the frequency responses of the delay-and-sum array. Lines (a) and (b) are respectively the responses to the desired signal before and after processing when MS is located directly in front of the array. These responses shows that the array response to the desired sound is almost flat, and the desired sound level is almost unchanged before and after processing.

Lines (c) and (d) are the responses to the noise before and after processing. The figure shows that noise is suppressed by about 10 dB above the 500 Hz range. Similar results were also obtained for the MS located 0.4 m to one side.

7.3. Effect of SNR improvement

To test the effect of SNR improvement achieved by the array, a speech recognition experiment was conducted.

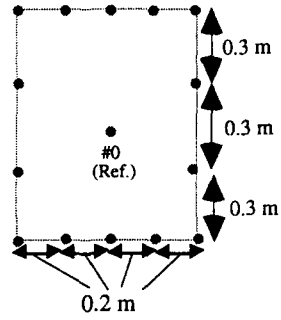

Figure 3. Geometry of array.

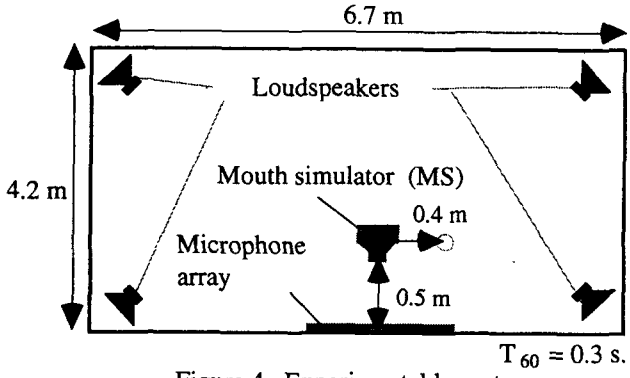

Figure 4. Experimental layout.

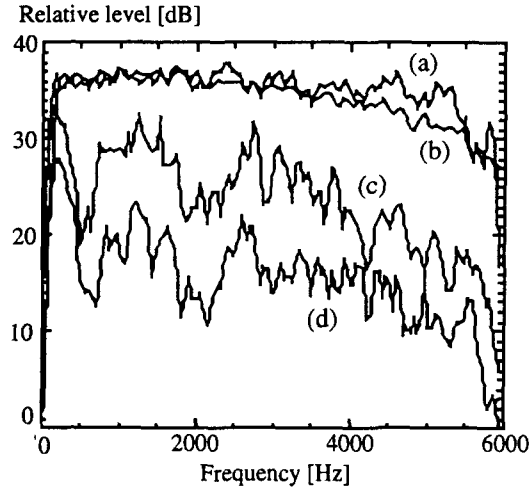

Figure 5. Frequency response of the array.

One hundred words with 11.5-dB SNR were acquired by omni- and uni-directional microphones and by our array system. To evaluate only the effect of SNR improvement, speech periods were assumed to be known (detected with noise-free speech) in this experiment. Table 1 shows the experimental results. Under the condition that SNR was 11.5 dB and MS was located directly in front of the array (0 m), the rate derived by the proposed array was 89%, which was about 10% and 20% (in absolute value) higher than the rates achieved with unidirectional and omnidirectional microphones. This improvement was mainly due to the SNR improvement of the array. When MS was located 0.4 m to the side, the array had almost the same rate as for the MS located directly in front, while the rate for the unidirectional microphone fell by 6% in absolute value. This indicates the robustness of our array against changes of speaker position.

7.4. Speech period detection

The proposed detection method (MLD) was compared with the conventional method based on the temporal power pattern (TPP) [10]. The MS was positioned directly in front of the array, and SNR was 6.5 dB.

The following three methods were compared.

A) unidirectional microphone + TPP
B) the array + TPP
C) the array + MLD

Table 2 presents the recognition rates for each speech period detection method. Method A could not detect most of speech periods correctly, and the recognition rate was 54%. Method B recognized 69% of the words correctly. This improvement was mainly due to SNR improvement of the array. Method C recognized 76% of the words correctly. This improvement was mainly due to the proper detection of speech periods by MLD for nonstationary crowd noise. Thus, the efficiency of the proposed speech period detection method is demonstrated.

8. CONCLUSIONS

The array system described here can improve the SNR of acquired sound, pick up desired speech with a flat spectrum response at an arbitrary speaker position, and detect the speech period in a noisy speech signal better than conventional methods. Its performance was assessed by word recognition experiments in a real sound field. This system achieves a lower error rate of speech recognition than the system with a typical conventional directional microphone and speech period detection method.

ACKNOWLEDGMENTS

We would like to thank Dr. N. Kitawaki and Dr. S. Furui for their helpful suggestions.

Table 1. Recognition rates: effectiveness of the array.

Noise	noise free	SNR = 11.5 dB	
MS location	0 m	0 m	0.4 m
Omni.	99%	70%	67%
Uni.	100%	79%	73%
Array	100%	89%	88%

Table 2. Recognition rates: effectiveness of MLD speech period detection method compared with conventional TPP method. SNR = 6.5 dB.

Methods	Rates
A) Uni.+TPP	54%
B) Array+TPP	69%
C) Array+ MLD	76%

REFERENCES

[1] Q. Lin, C. W. Che, B. Vries, J. Pearson, and J. Flanagan, " Experiments on distant-talking speech recognition.", Proc. ARPA Spoken Language Systems Technology Workshop, pp.187-192 (Jan. 1995).

[2] D. Giuliani and M. Omologo and P. Svaizer, " Talker localization and speech recognition using a microphone array and a cross-powerspectrum phase analysis", Proc. ICSLP, Yokohama, Vol.3, pp.1243-1246 (Sep. 1994).

[3] H. F. Silverman and S. E. Kirtman, " A two-stage algorithm for determining talker location from linear microphone array data ", Computer Speech and Language, Vol.6, pp.129-152 (1992).

[4] I. Yamada and N. Hayashi, " Improvement of the performance of cross correlation method for identifying aircraft noise with pre-whitening of signals", J. Acoust. Soc. Jpn. (E), Vol.13, No.4, pp.241-252 (1992).

[5] J. L. Flanagan, D. A. Berkley, G. W. Elko, J. E. West, M. M. Sondhi, " Autodirective Microphone Systems", Acustica, Vol.73, pp.58-71 (1991).

[6] H. Nomura, Y. Kaneda, and J. Kojima, " Optimum gains of a delay-and-sum microphone array for near sound field", Proc. Acoust. Soc. Am. and Acoust. Soc. Jpn. Third Joint Meeting, Honolulu, Hawaii, 3aSPa9, pp.1291-1296 (Dec. 1996).

[7] H. Kuttruff, " Room Acoustics (Third Edition)", Elsevier Applied Science, pp.100-132 (1991).

[8] Y. Kaneda, " Speech period detection using an adaptive microphone array", Proc. IEEE Workshop on Application of Signal Processing to Audio and Acoustics, Session 1-8 (Oct. 1989).

[9] ITU-T Software Tool Library, release 1996, Chapter 10, SVP56: The Speech Voltmeter.

[10] L. R. Rabiner and M. R. Sambur, " An Algorithm for Determining the Endpoints of Isolated Utterances", Bell Syst. Tech. J., Vol. 54, No. 2, pp.297-315, (Feb. 1975).