VECIMS 2003 - International Symposium on
Virtual Environments, Human-Computer Interfaces, and Measurement Systems
Lugano, Switzerland, 27-29 July 2003

# Realtime Detection and Identification of Plural Speakers Using a Microphone Array

Minoru OHKADO and Hideyuki SAWADA
Department of Intelligent Mechanical Systems Engineering, Faculty of Engineering, Kagawa University
Address: 2217-20, Hayashi-cho, Takamatsu-city, Kagawa, 761-0396, JAPAN
Phone: +81-(0)87-864-2324, Fax: +81-(0)87-864-2369, Email: sawada@eng.kagawa-u.ac.jp

***Abstract*** *– Voices are used as primary media in human communication. A human voice is a sound generated by the complex movements of the vocal organs, and is the most important media employed for the communication in the day life to logical discussions. A human is able to detect the position of a source sound in 3D space by perceiving the time difference reaching to the both ears. Furthermore we can selectively listen to an objective voice in the crowds. This paper presents a realtime detection and identification of a particular person among plural speakers using a microphone array. The system identifies the position of a particular speaker, and enhances the voice signal selectively.*

## I. INTRODUCTION

Human voice is used as primary media in the human communication. It is employed not only in simple daily communication, but also for the logical discussions. Human is able to exchange information smoothly using voice under different situations such as noise circumstances in a crowd and existence of plural speakers. Human is able to detect the position of a source sound in 3D space, extract a particular sound from mixed sounds, and recognize who is talking. If this mechanism is realized in a computer, it will be possible to record a sound with high quality by reducing noise, present a clarified sound, and realize a microphone-free speech recognition by extracting particular sound.

Various techniques for the detection and identification of a particular sound from sound signal have been proposed so far, which are classified into two approaches; one is to develop the technique of sound source separation from a monophonic input[1]-[3], and the other uses sound source information based on the stereo inputs [4]-[6]. Since most of the techniques require sound source model or assume certain special conditions and restrictions, the computational costs become large, which causes the difficulties of the realtime processing.

This paper introduces a realtime detection and identification of a particular person among plural speakers using a microphone array based on the location of a speaker and the individual voice characteristics.

## II. SYSTEM CONFIGURATION

Figure 1 shows the system configuration, which consists of a microphone array, a low-pass filter (LPF), A-D board, a stereo speaker and a computer. The microphone array consists of 13 microphones arranged inline as shown in Figure 2. The microphone array is connected to the computer via the A/D board. The cut-off frequency of the LPF, which is placed in front of the A/D board, was set to 20 kHz. The sampling conditions are shown in Table 1.

The system identifies the position of a particular sound source among plural sources by inputting sound data in parallel from the microphone array, and selectively enhances the particular sound signal to output from a stereo speaker.

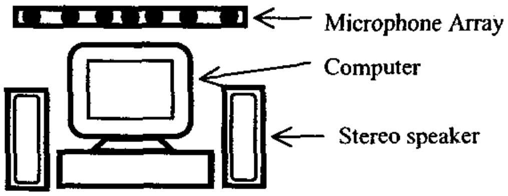


Fig. 1. System configuration.

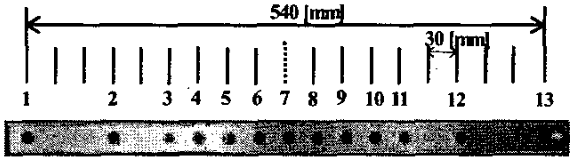


Fig. 2. Microphone array.

Table 1. Sampling condition.

| | |
|---|---|
| Sampling frequency | 44 [kHz] |
| Sampling number | 512 |
| Window function | Hamming Window |
| Cut-off frequency | 20 [kHz] |

0-7803-7785-0/03/$17.00 ©2003 IEEE

## III. ESTIMATION OF SOUND SOURCE DIRECTION

Here we assume that the source sound is a plane wave, and travels straight. The direction of a speaker $\theta$ can be estimated by measuring the time delay between two microphones as described in Figure 3. The direction can be estimated by the calculation of CSP (Cross-Power Spectrum phase analysis) coefficients as

$$CSP_{1,2}(k) = DFT^{-1}\left[\frac{DFT[x_1(n)]DFT[x_2(n)]^*}{|DFT[x_1(n)]||DFT[x_2(n)]|}\right]$$

$$CSP_{1,2,M} = \sum_{n=1}^{M} CSP_{1,2,n}(k), \qquad (1)$$

where $x_1(n)$ and $x_2(n)$ are the sampled signals of the microphone No1 and No2, respectively. DFT represents the calculation of the discrete Fourier transform, * represents the complex conjugate, and k corresponds to the time difference between the two signals [7],[8].

An experiment was held by placing source sounds in front of the microphone array at intervals of 5 degrees on the circumference with a 150 cm radius, and the direction was estimated in a open 3D space. The experimental result is shown in Figure 4. The result presented that fair estimation was assumed in the direction between 30 and 150 degrees in front of the microphone array.

## IV. LOCATION ESTIMATION OF A SPEAKER IN 3D SPACE

The direction of a sound source can be estimated by using a pair of microphones placed apart. By using two pairs, the location of a sound source in 3D space can be calculated based on the measurement of the triangulation as shown in Figure 5.

We held an experiment of source position estimation. Sound sources were placed at 28 positions ( a to $\beta$ ) with the interval of 300mm as shown in Figure 6. A human voice was output from an acoustic speaker in the alphabetic order, and its position was estimated by using the algorithm mentioned above.

An experimental result of location estimation is shown in Figure 7. Estimation errors are described as vectors, where the origin of a vector shows the actual position of a source sound, and the end point indicates the estimated source position. The shorter an arrow length is, the more accurate the estimation was achieved. The result proved a satisfactory estimation in the frontal area of the microphone array.

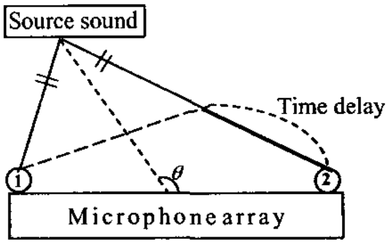


Fig. 3. Time difference reaching two microphones.

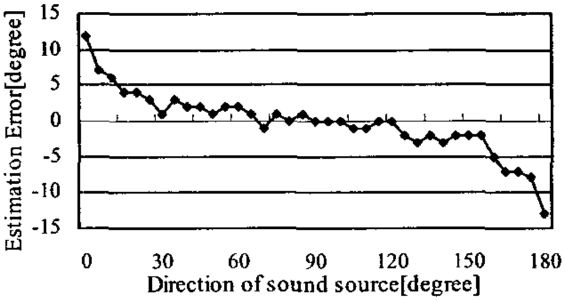


Fig. 4. Experimental result of direction estimation.

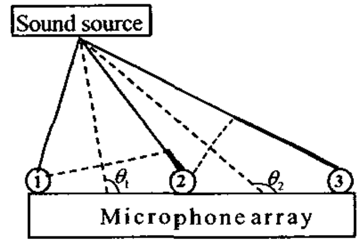


Fig. 5. Location estimation using two pairs of microphones.

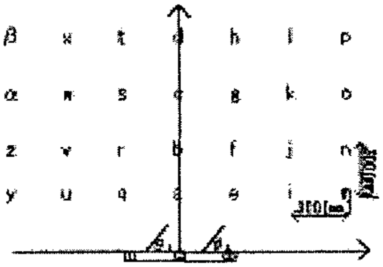


Fig. 6. Source positions.

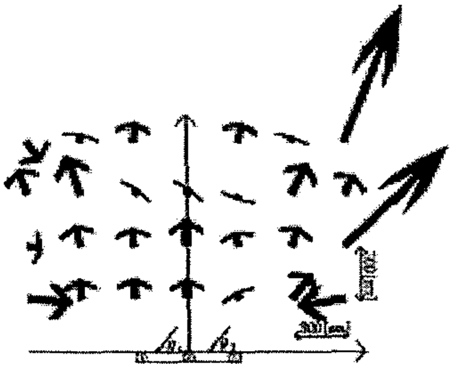


Fig. 7. Experimental result of location estimation.

## V. SPEECH ENHANCEMENT OF A PARTICULAR SPEAKER

A sound source located in the direction $\theta$ is enhanced by the calculation of the delay-and-sum beam forming[3] as

$$y(t)=\sum_{i=1}^{M} x_i(t)exp\left\{j2\bullet f(i-1)\frac{d\cos\theta}{c}\right\}, \tag{2}$$

where $M$ is the total number of microphones, $d$ is the distance between two microphones, $c$ is the sound speed, and $i$ is the microphone number.

An experiment of speech enhancement was held by placing a sound source at the direction of 45 degrees and 135 degrees. The results of the enhancement are shown in Figure 8, in which sounds travelled from the direction of 45, 90 and 135 degrees were enhanced. Figure (a) and (b) show the successful enhancements of sounds from the direction of 45 and 135 degrees, respectively, where the attention to the other directions attenuated the sound.

## VI. RECOGNITION OF PARTICULAR SPEAKER

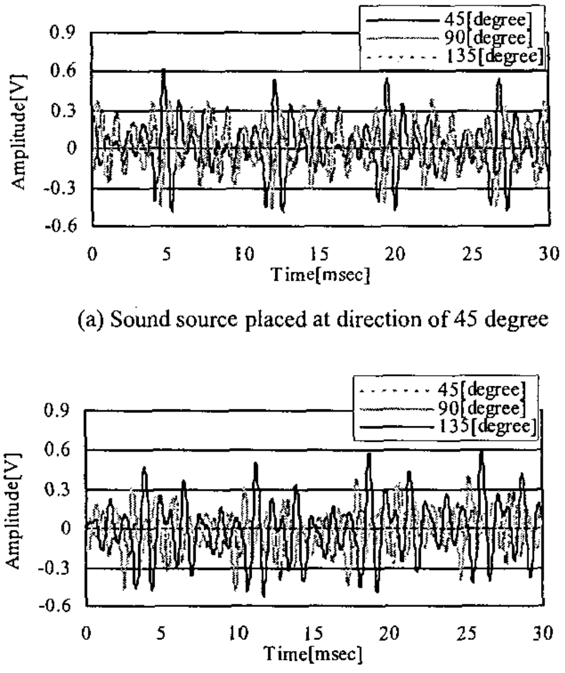
Human is able to identify and segregate a particular person from plural speakers by using information such as the location of the source and the personal information of voice characteristics[9]. In this regard, computerized sound segregation can be realized by reproducing this procedure. As presented in the chapter above, the active selection of a particular sound source in 3D space was achieved with the microphone array.

Here we paid attention to the extraction of characteristics from a particular sound to be segregated from plural different sounds, and examined the employment of cepstrum coefficients. Figure 9 shows examples of cepstrum coefficients extracted from Japanese /a/, /i/ and /u/ vowels vocalized by two different persons. The cepstrum coefficients are often used for voice recognition, since the coefficient values differ from each vowel. We have further noted the difference between persons who spoke same vowels, and tried to extract the individuality of voice characteristics. In this study, the cepstrum coefficients from the 2nd to 20th orders were employed as individual voice templates for the identification of a person. Each template was averaged by repeating the acquisition 10 times. Figure 9 also presents the examples of templates.

In the recognition, each time a sound from unknown source is inputted, cepstrum coefficients from 2nd to 20th orders are extracted and absolute errors against each template are calculated. The speaker whose error is the minimum is selected as the recognition result.

An experiment of the speaker recognition was examined. All of the 10 speakers' voices were recorded in advance, and cepstrum coefficients for all vowel sounds were extracted and stored as individual voice templates.

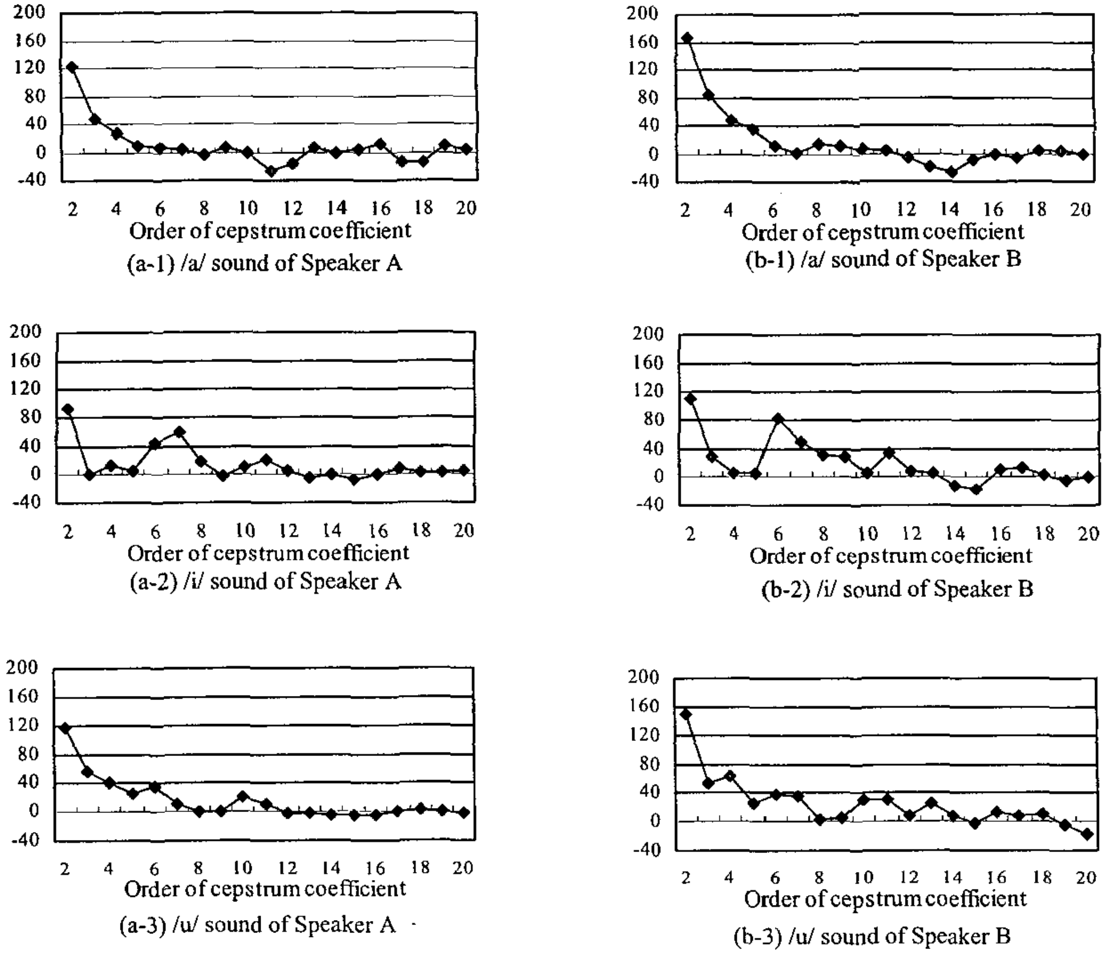
Experimental results of speaker recognition are shown in Figure 10. In the figures, for example, (a) presents an example of the recognition result which answered speaker A correctly. 98% recognition rate was achieved for 10 persons.



(a) Sound source placed at direction of 45 degree

(b) Sound source placed at direction of 135 degree

Fig. 8. Enhancement of particular sound source.



(a-1) /a/ sound of Speaker A

(b-1) /a/ sound of Speaker B

(a-2) /i/ sound of Speaker A

(b-2) /i/ sound of Speaker B

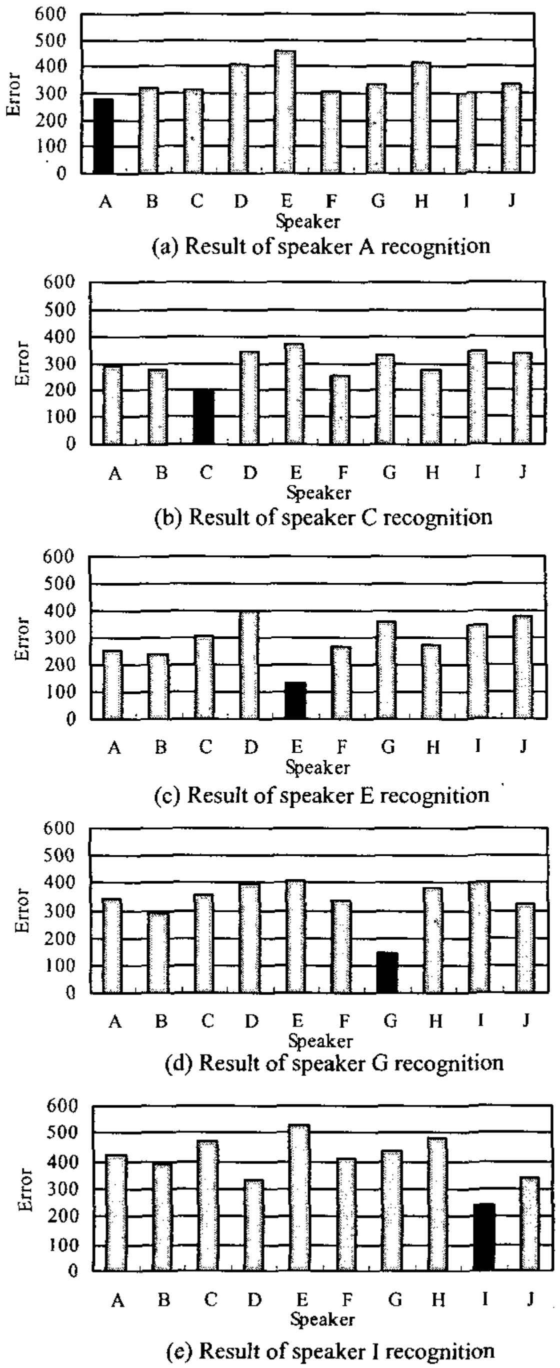
(a-3) /u/ sound of Speaker A

(b-3) /u/ sound of Speaker B

Fig. 9. Cepstrum coefficients of vowel sounds.

Table 2. Recognition results.

| | |
|---|---|
| Recognized both speakers correctly | 26 |
| Recognized as same speaker<br>( Recognized both speakers correctly by considering 2nd candidate ) | 14<br>(4/14) |
| Recognized one speaker correctly | 5 |
| Mis-recognized both speakers | 0 |



(a) Result of speaker A recognition

(b) Result of speaker C recognition

(c) Result of speaker E recognition

(d) Result of speaker G recognition

(e) Result of speaker I recognition

Fig. 10. Examples of single speaker recognition.

## VII. IDENTIFICATION OF A PARTICULAR SPEAKER AMONG PLURAL SPEAKERS

This chapter introduces the identification of a particular person among plural speakers who talk simultaneously. The system segregates a speaker by actively using the clue of his location and the individual characteristics of voice.

First, the system estimates the location of each sound source, whose sound wave is enhanced one by one. Then, each sound wave is used for the recognition of individuals. Each voice is selectively outputted from a stereo speaker, according to the listener's intention.

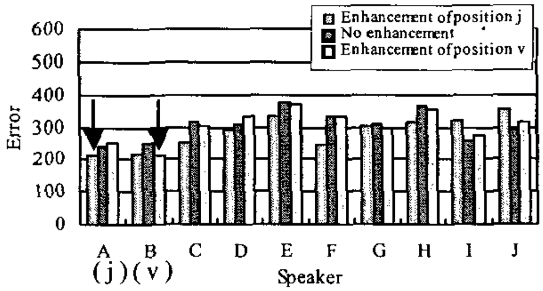
In the experiment, the sampling frequency was set to 44[kHz], and the frame length was fixed 25.6[ms]. The calculation was executed by using sound data after multiplying a humming window.

Two sound sources were placed at the position j and v as presented in Figure 6. An experiment was conducted by locating 10 speakers (A to J) at the two positions. The recognition ability of the proposed method was examined by enhancing sounds from the positions j and v.

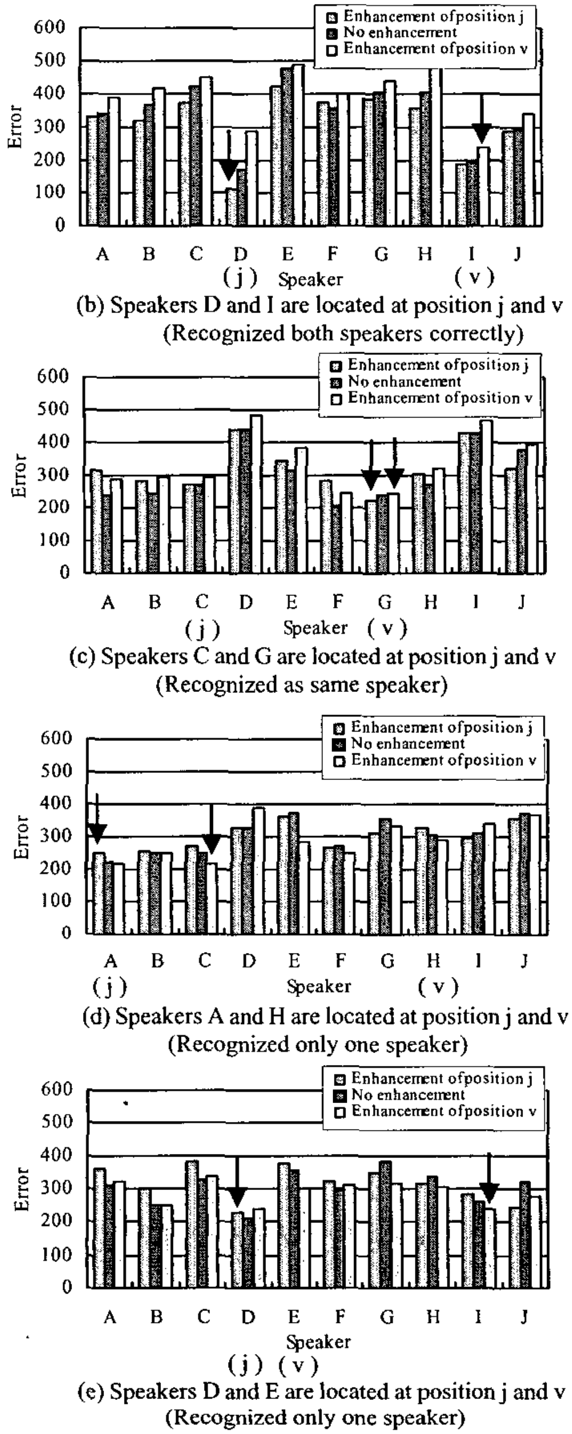
Examples of the estimation of two speakers are shown in Figure 11. In the figures, (a) and (b) recognized both speakers correctly, (c) answered same speaker as the first candidate and the true speaker was ranked as 2nd candidate. (d) and (e) recognized only one speaker correctly.

Table 2 shows the experimental results, and 58% of speaker recognition was achieved for the simultaneous speeches by 45 pairs. For the 42% rest, at least one person was correctly recognized, and the other speaker could be narrowed and re-estimated by considering the other candidates.

The algorithm should be further improved for the recognition and identification of persons speaking simultaneously.



(a) Speakers A and B are located at position j and v
(Recognized both speakers correctly)



(b) Speakers D and I are located at position j and v (Recognized both speakers correctly)

(c) Speakers C and G are located at position j and v (Recognized as same speaker)

(d) Speakers A and H are located at position j and v (Recognized only one speaker)

(e) Speakers D and E are located at position j and v (Recognized only one speaker)

Fig. 11. Recognition of two speakers located at two different positions.

## VIII. CONCLUSIONS

In this paper, a realtime detection and identification of a particular person among plural speakers was presented by using a microphone array and individual voice characteristics. First the location of a speaker among plural persons was estimated, and the voice was enhanced according to the location. Then a particular person was identified to extract only his speech.

We have achieved fair results of speaker identification, without total misrecognition of both speakers. We have to further work for the improvement of the recognition algorithm to be applied to realize the recognition of unknown speakers who is not included in the voice templates in advance.

We are now working to construct a hands-free speech tele-communication system with a speech recognition to present a new multimodal communication.

## REFERENCES

[1] Masashi Unoki and Masato Akagi, "A Method of Signal Extraction from Noise-Added Signal", IEICE, Vol. J80-A, No.3, pp.444-453, 1997

[2] Shoji Hayakawa, Kazuya Takeda and Fumitada Itakura, "Speaker Recognition Using the Harmonic Structure of Linear Prediction Residual Spectrum", IEICE, Vol. J80-A, No.9, pp.1360-1367, 1997

[3] Nehorai, A. and Porat, B. "Adaptive Comb Filtering for Harmonic Signal Enhancement" IEEE trans. ASSP, Vol.34, No.5, pp.1124-1138(1986)

[4] Takeshi Yamada, Satoshi Nakamura and Kiyohiro Shikano, "Hands-free Speech Recognition with Talker Localization by a Microphone Array", Information Processing Society of Japan, Vol.39, No.5, pp. 1275-1284, 1998.

[5] Futoshi Asano, Satoru Hayamizu and Toshihiro Matsui, "A Realtime Noise Reduction System using Delay-and-Sum Beamformer and its Application to Speech Recognition", Electrotechnical Laboratory, 1996

[6] 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.

[7] Takanobu Nishiura, Takeshi Yamada, Satoshi Nakamura and Kiyohiro Shikano, "Localization of Multiple Sound Sources Based on CSP Analysis with a Microphone Array", IEICE, D-II, Vol. J83-D-II, No.8, 2000.

[8] C.H. Knapp and G..G. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust., Speech & Signal Processing, Vol.24, No.4, pp.320-327, 1976.

[9] Tetsuo Funada and Takashi Tsuzuki, "Feature extraction based on spectral slope for speech recognition", IEICE, D-II, Vol. J82-D-II, No.11, pp.2184-2187, 1999.