SPEECH RECOGNITION MODULE FOR CSCW USING A MICROPHONE ARRAY

Takashi Endo, Shigeki Nagaya, Masayuki Nakazawa, Kiyoshi Furukawa and Ryuichi Oka

Tsukuba Research Center Real World Computing Partnership

Tsukuba Mitsui Building 13F, 1-6-1 Takezono Tsukuba-shi, Ibaraki 305, JAPAN

Tel. +81 298 53 1687, FAX:+81 298 53 1740, E-mail:enchan@trc.rwcp.or.jp

ABSTRACT

This report proposes a recognition module for use in CSCW that suffers little degradation in recognition performance even when more than one person speaks at the same time and they speak at a distance from a microphone. This is accomplished by controlling directionality using a microphone array and estimating transmission characteristics from speakers to microphones. On the basis of evaluation performed by word spotting from continuous speech, it has been found that this module raises the recognition rate by (1) 30% in an environment where two people are speaking at the same time, and (2) by 15% when people speak at a distance of 160 cm from a microphone.

1. INTRODUCTION

We have proposed a CSCW system with a new format called Open Cooperative Work Space (OCoWS) [1] in which users can perform collaborative work while communicating in a much more natural manner through an interface that inputs speech and gestures. The speech interface used by OCoWS can input a user's voice even from a distance and does not require users to be aware of microphone locations. In this way, OCoWS aims to achieve a natural speech interface for users.

In this kind of speech interface, two problems arise: (1) the effect of noise and speech of people other than the target speaker becomes great compared to the situation when microphones are right next to users; and (2) mismatch occurs with reference data in the recognition module due to distortion in the speech waveform because of reflections by walls.

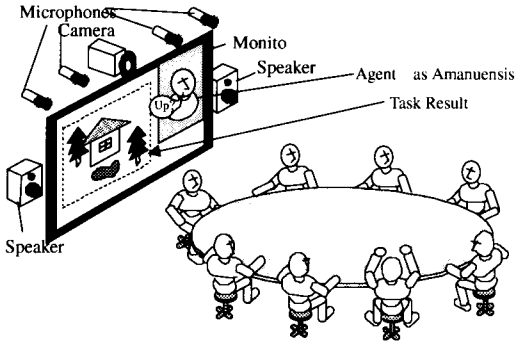

Fig 1 OpenCooperative WorkSpace

In this report, as part of the speech recognition module making up OCoWS, we propose the use of active microphones that feature little drop in recognition rate even in an environment where multiple users are talking at the same time and the target speaker is far from microphones. This is accomplished by performing distortion adaptation through estimation of transfer function using linear predictive analysis and by suppressing interfering sounds through the use of a microphone array [2].

2. ACTIVE MICROPHONE

An active microphone consists of (1) estimation of the target speaker's position through picture recognition [3]; (2) a microphone array for selectively extracting the speech of the target speaker; and (3) processing for estimating transfer function from the target speaker to a microphone and for adapting the recognition system.

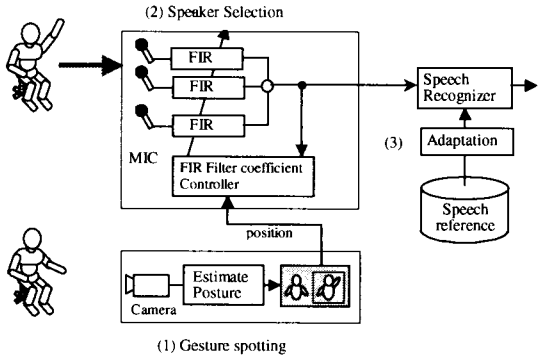

Fig. 2 System diagram

2.1. Gesture recognition

The authors have previously proposed a method [3] for detecting in real time the number of people in a video image and their positions and for classifying their gestures, without the need for operations like removing the area containing people and normalizing size. We use this method to estimate the position of target speakers.

Fig. 3 shows the result of gesture recognition.

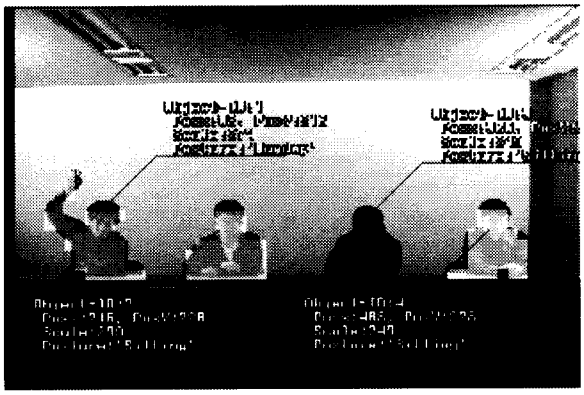

Fig. 3 Gesture Recognition [3]

2.2. Extracting the target speaker's speech

Directional control for suppressing interfering sounds from other than the target speaker can be achieved by a microphone array connected to a FIR filter. The AMNOR [2] algorithm is used in coefficient estimation of the FIR filter.

2.3. Adaptation for distortion

Fig. 4 shows the model for transfer distortion in this system.

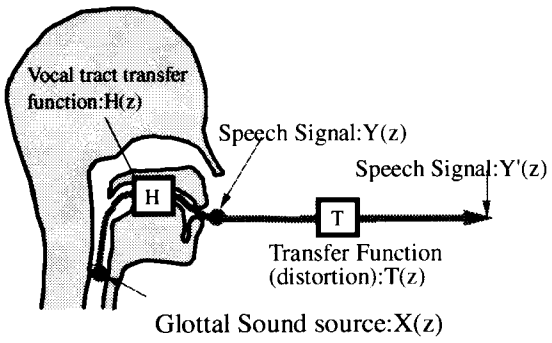

Fig. 4 Distortion Model

The human voice can be modeled as the sum of the wave form from a glottal sound source and of the vocal tract transfer function. The vocal tract transfer function in the case of a glottal sound source is modeled as a totally polar system, and the transfer function can be estimated by linear predictive analysis. Here, if the transfer function can also be modeled as a totally polar system, then is totally polar as well and it can be determined by linear predictive analysis from the microphone input signal . We therefore obtain the following equation:

$$T(z)=\frac{T(z)H(z)}{H(z)}=\frac{1+\sum_{i=1}^{p}\alpha_i z^{-i}}{1+\sum_{j=1}^{q}\beta_j z^{-j}}=1+\sum_{k=1}^{\infty}h_k z^{-k} \qquad (1)$$

A FIR filter can be configured here by truncating the impulse response in the above equation at the order in which the value of h_k becomes sufficiently small. Then, by combining with reference speech of the speech recognition system, adaptation processing can be performed with regard to transfer distortion.

3. EVALUATION EXPERIMENTS

We evaluated ability of our method about speech separation and adaptation for distortion by word spotting from continuous speech using Continuous DP[4].

3.1 Extract a specific speaker's speech

For the situation of two male speakers talking at the same time, an experiment was performed to see if the speech of the target speaker could be isolated and recognized.

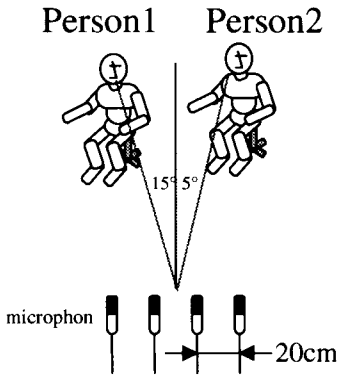

Fig. 5 Speaker isolation experiment

3.1.1. Experimental condition

The positional relationship between the two speakers and the microphone layout are shown in Fig. 5. Person 1 in the figure is taken to be the target speaker for recognition. For experimental samples, we used the speech of two males independently collected beforehand and synthesized on computer. In this experiment, the position of Person 1 was given as known. The number of FIR filter taps was 30 in the case of 2 and 4 microphones and 66 for 8 microphones. To estimate filter coefficients, we used a speaking interval of 0.5 sec from Person 2 only. The experiment was conducted completely off line.

3.1.2. Result

Experimental results are shown in Fig. 6. Although recognition rate improved from 30% to 60% for 2 microphones, the S/N ratio increased for more than this number. In this range, it was confirmed that speech could be clearly understood if listened to by people, but improvement in the recognition rate of the recognition system could not be observed. In addition, processing speed was about twice that of real time. Directivity characteristics and frequency characteristics in the case of 8 microphones are shown in Figs. 7 and 8, respectively.

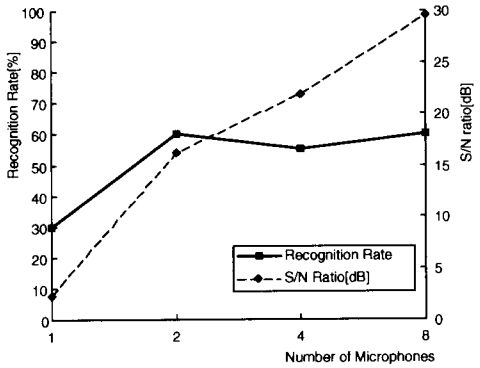

Fig. 6 Recognition rate and S/N ratio

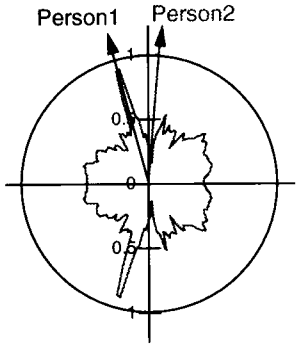

Fig. 7 Directivity characteristics

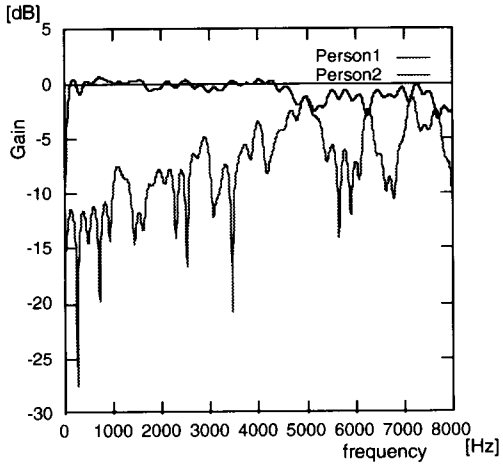

Fig. 8 Frequency characteristics

3.2. Adaptation for transfer distortion

For the case in which recognition performance drops when a speaker and microphones are far apart, an experiment was performed to see how far recognition could be improved by transfer distortion adaptation.

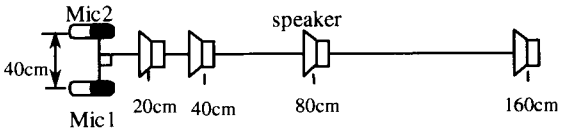

Fig. 9 Positional relationships between speaker and microphones

3.2.1. Result

The positional relationships between a speaker and microphones are shown in Fig. 9. In this experiment, speech recorded earlier in a soundproof room was presented from an electronic speaker, and reference speech used a specific speaker. A recognition experiment was conducted under the following three conditions.

(Condition1) Ordinary

Recognition of the recorded speech is attempted using microphone 1 in Fig. 9.

(Condition2) Extraction of target speaker (directivity control)

Using the two microphones in Fig. 9, recognition of the recorded speech is attempted while controlling directivity in the direction of the speaker.

(Condition3) Condition 2 plus transfer distortion adaptation (adaptation)

For the speech in Condition 2, its transfer function is also estimated, and recognition is attempted by combining with reference speech of the speech recognition system. Transfer characteristics were estimated from two seconds of speech.

3.2.2. Result

Experimental results are shown in Fig. 10. For condition 1, recognition rate deteriorated as the speaker and microphones became further apart. By controlling directivity in Condition 2, recognition rate could be improved for S/N of from 1 to 2 dB and deterioration in recognition rate with distance could be eased. Finally, in Condition 3, where transfer distortion adaptation was combined with directivity control, it was found that recognition rate could be improved over all distances, thus confirming the effectiveness of transfer distortion adaptation.

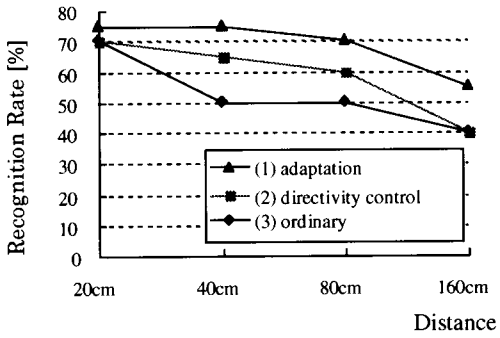

Fig. 10 Recognition rate

4. IMPLEMENT ON DSP

We implemented the system for selectivity extracting the speech on DSP board plugged in the PC/AT. Table. 1 shows the specification of the system. The microphones of this system are omnidirectional microphones by their selves. Fig.12 shows the microphone array of the system.

Table 1 Specification of the system

DSP	Motolola 56002 x 4
Microphone	Earthworks OM-1
Channels	8 channel
D/A	16 bit / 16 KHz
Taps	128 taps
PC/AT	Pentium 166

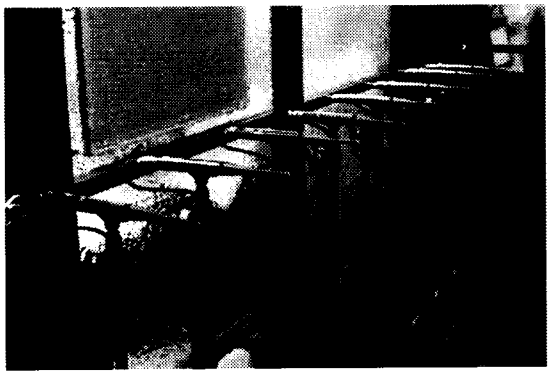

Fig. 11 Mic array of Active Microphone

4.1. Experiment

We evaluated ability of our system about speech separation in the real environment. For the situation that two male speakers talk at the same time in front of the microphones. Fig. 12 shows the layout of speakers and microphones. Microphones are placed at intervals of 26 cm. Distance between microphones and speakers is 180cm, and distance between speakers is 60cm.

In the experiment, FIR filters are adapted to extract speaker 2's voice. The position of speaker-2 is given by delay time calculated from cross correlation between microphones. Gain at each point is expressed by height in fig. 12.

As shown in Fig. 12, a gain at a speaker-1 is suppressed while a gain at the speaker-2 keeps high level. After extracting speaker-2's voice, S/N ratio of extracted speech wave has improved by 13 dB.

(1) Original speech [sound A0174S01.WAV]
(2) Extracted speech[sound A0174S02.WAV]

5. CONCLUSION

This paper has proposed active microphones for use in speech recognition modules applied to CSCW. Evaluation experiments were also performed, and it was found that (1) recognition rate could be improved by 30% in an environment where two people are speaking at the same time, and (2) recognition rate could be improved by 15% for a speaker-to-microphone distance of 160 cm. These results demonstrated the effectiveness of the active microphone technique.

REFERENCES

[1] Nagaya et al. : *"A Proposal of Novel Information Integration Architecture - Open Cooperative Work Space"*, Proc. of MFI-96, pp.425-432 (1996 Washington D.C).

[2] Kaneda : *"Adaptive Microphone-Array System for Noise Reduction"*, IEEE Trans. ASSP, Vol. ASSP-34, No. 6, pp. 1391-1400 (1986-12).

[3] Nagaya et al. : *"Interaction Control of the CSCW System using Posture Recognition"*, Proc. of ICMI96, pp. 299-304 (1996-10).

[4] Oka : *"Phonemic Recognition of Each Frame by Partial Matching Method Based on Continuous Dynamic Programming"*, Tran. of the Institute of Electronics, Information and Communication Engineers, D, J70-D, No.5, pp. 917-924(1987-5).

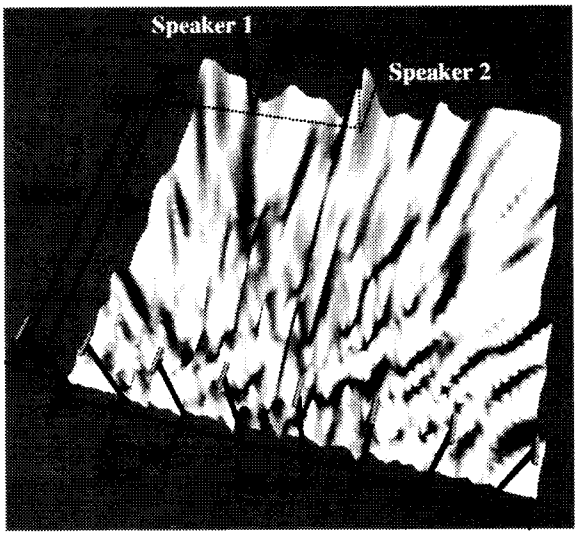

Fig. 12 gain at each points