SPEECH RECOGNITION BASED ON SPACE DIVERSITY USING DISTRIBUTED MULTI-MICROPHONE

Yasuhiro Shimizu[†], *Shoji Kajita*[‡], *Kazuya Takeda*[†] *and Fumitada Itakura*[‡]

[†] Graduate School of Engineering, [‡] Center for Information Media Studies
Center for Integrated Acoustic Information Research (CIAIR), Nagoya University
Furo-cho 1, Chikusa-ku, Nagoya 464-8603 JAPAN
kajita@media.nagoya-u.ac.jp

ABSTRACT

This paper proposes space diversity speech recognition technique using distributed multi-microphone in room, as a new paradigm of speech recognition. The key technology to realize the system is (1) distant-talking speech recognition and (2) the integration method of multiple inputs. In this paper, we propose the use of distant speech model for the distant-talking speech recognition, and feature-based and likelihood-based integration methods for multi-microphone distributed in room. The distant speech model is a set of HMMs learned using speech data convolved with the impulse responses measured at several points in room. The experimental results of simulated distant-talking speech recognition show that the proposed space diversity speech recognition system can attain about 80% in accuracy, while the performances of conventional HMM using close-talking microphone are less than 50%. These results indicate that the space diversity approach is promising for robust speech recognition under the real acoustic environment.

1. INTRODUCTION

State-of-art speech recognition systems can attain high recognition accuracy when using close-talking microphone. However, the performance deteriorates significantly as the distance between speaker and microphone becomes large, due to the reverberation in room, background noises, the other competing speakers and so on. In the current main researches of such distant-talking speech recognition, the use of microphone array system as the front-end of speech recognizer have been proposed and shown to be effective[1, 2, 3].

This paper investigates a new paradigm, i.e., space diversity speech recognition system using distributed multi-microphone. The idea of space diversity is the integration of multiple sensory inputs, typically by selecting the most reliable channel, i.e., a microphone, as shown in Figure 1. The advantage of space diversity is that speech recognizer can capture spatial extent acoustic information in room, while microphone array can only capture one directional acoustic information. This view of the acoustic field will make speech recognition system robust against interferes distributed in room. In addition, the use of distributed multi-microphone will make the arrangement of microphone in room easier than the case of using microphone array. Consequently, the system will be easy-to-use and user-friendly, because users do not have to care about where to put microphones in installing and using the system. To realize such system, however, distant-talking speech recognition and the integration method of multiple inputs have to be developed.

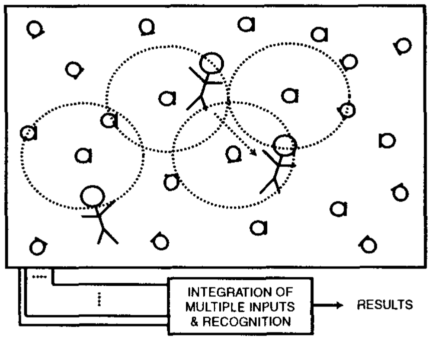

Figure 1: Concept of space diversity speech recognition using distributed multi-microphone in room. The dotted circle stands for the reliable area of the microphone.

This paper proposes the use of distant speech model, and feature-based and likelihood-based integration method. The distant speech model is a set of Hidden Markov Models (HMMs) learned using speech data convolved with the impulse responses measured at several points in room, in expectation of the high ability of modeling provided by HMM.

In this paper, we examine the following three points: (1) the basic performance of distant speech model, (2) the recognition rate in room in the case of using the distant speech model, (3) the feature-based and likelihood-based integration method. Note that the influence of background noises is out of range through this research.

2. DISTANT SPEECH MODEL

The distant speech model is a set of HMMs affected by the room acoustics from a sound source to a microphone. Learning of the set of distant speech model is performed for every microphone using simulated distant-talking speech data convolved with the impulse responses measured at several points in room.

In this research, seven distant speech models are used, which were trained using the simulated distant speech of 41,372 sentences from the Japanese Newspaper Article Sentences (JNAS) and ATR phoneme balance sentences corpus in the Acoustical So-

0-7803-6293-4/00/$10.00 ©2000 IEEE.

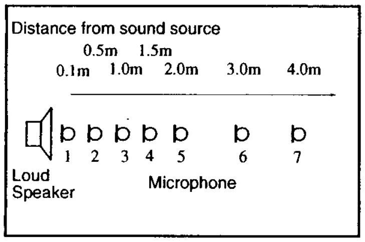

Figure 2: Microphone arrangement 1.

Table 1: Analysis conditions

Sampling frequency	16 kHz
Analysis window	Humming window
Analysis frame length	25ms
Frame cycle	10ms
Pre-emphasis	0.97
MFCC	12th order
Δ MFCC	12th order
Δ Power	1st order

ciety of Japan the Continuous Speech Corpus for Research[4, 5]. In measurement of impulse responses, seven omni-directional microphone (Sony ECM-77B) were used. Each microphone was set at 0.1m, 0.5m, 1.0m, 1.5m, 2.0m, 3.0m, and 4.0m from the loud speaker (See Figure 2). The height of each microphone is 1.0m. The size of room is 5.67m × 3.67m. The reverberation time is 320ms. The impulse responses were measured using the Time Stretched Pulse (TSP) signal which is a kind of charp signal and has 1.365 seconds long. Each HMM is the triphone model (gender independent) of 32 mixture per state. The feature parameter and analysis conditions are the same as those of IPA standard acoustic model. Note that CMS was performed for every sentence.

3. EVALUATION OF DISTANT SPEECH MODEL

In this section, we evaluate the basic performance of the distant speech model as a function of the distance from sound source.

3.1. Experimental conditions

The test speech used in evaluation is 50 ATR phoneme balance sentences in the corpus [5] uttered by one male speaker. In recognizing the test speech, we convolved the 50 sentences with each impulse response whose distance from sound source is 0.1m, 0.5m, 1.0m, 1.5m, 2.0m, 3.0m, and 4.0m (Figure 2). The recognition task is grammar-less continuation clause recognition of 311 clauses. The recognition results are measured by the recognition accuracy that takes the insertion error into account.

3.2. Performance of Distant Speech Model

Figure 3 shows the result compared with the following two cases:

1. the distance from sound source and microphone is same in learning and recognizing (Matched HMM),
2. A standard HMM provided from IPA is used, which is not learned any room acoustics (IPA HMM)[4].

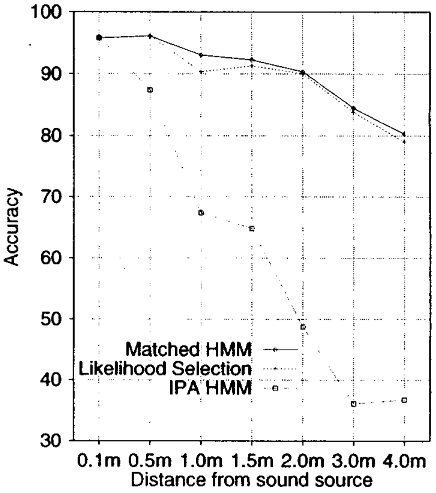

Figure 3: The recognition results as a function of distance.

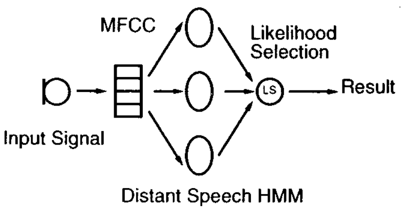

Figure 4: Maximum likelihood-based distant speech model selection.

As shown in the figure, the recognition accuracy of IPA HMM significantly degrades as the distance from sound source is larger, even if CMS is performed. However, the distant speech model learned with the appropriate room impulse response from sound source attains more than 90% even if the distance is 2.0m and below. Comparing with the case of IPA HMM, the error rate of Matched HMM is decreased above one-third. These results indicate that HMM has the ability of modeling for speech affected by the room acoustics is quite high.

From this experiment, it is shown that distant speech model can attain high recognition accuracy if the distance from speaker is below 2.0m.

3.3. Maximum likelihood-based Distant Speech Model Selection

In space diversity recognition system, however, the distance from speaker is unknown. Thus, any criterion to select a reliable model from multiple distant speech models is required. To solve the problem, we also performed a model selection based on maximum likelihood (Likelihood Selection). In the selection, an input speech is independently recognized by two or more distant speech models, and the model of the highest likelihood is selected (Figure 4).

The result of the model selection based on maximum likelihood is the almost same as that of Matched HMM, as shown in Figure 3. This result indicates the maximum likelihood-based selection of the distant speech model is effective when the distance from sound source is unknown.

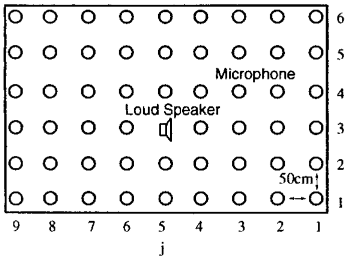

Figure 5: Microphone arrangement 2

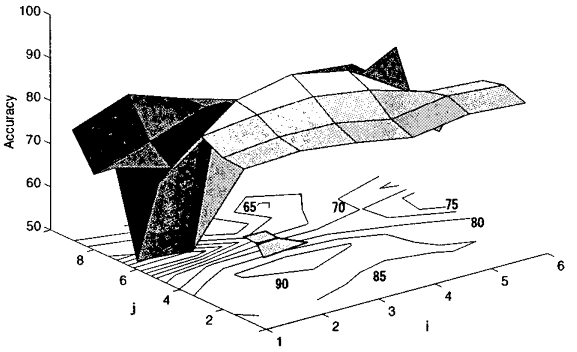

Figure 6: Recognition accuracy in the whole room.

4. RECOGNITION RATE IN ROOM USING DISTANT SPEECH MODEL

Secondly, we investigate whether the maximum likelihood-based HMM selection using the distant speech model introduced in Section 2 is effective or not in any other points in room.

4.1. Experimental Conditions

The microphones were arranged at the height of 1m, and the interval is 50 cm in room, as shown in Figure 5. In the experiment, the test speech data were convolved with the impulse response from the sound source to each microphone, as in the case of Section 3.

The seven distant speech models described in Section 2 were used. The model selection in recognition was performed using the maximum likelihood-based HMM selection described in the previous section. The other experimental conditions are the same as those of Section 3.

4.2. Recognition Results

Figure 6 shows speech recognition accuracy in room. The highest speech accuracy attains about 90% around the position of the distant speech model, i.e., the matched condition in learning and testing. Although the performance still attains about 80% in front of the loud speaker, it degrades behind and at the side of the loud speaker. These results seems to reflect the condition that no distant speech model for the positions behind and at the side of the loud speaker is used.

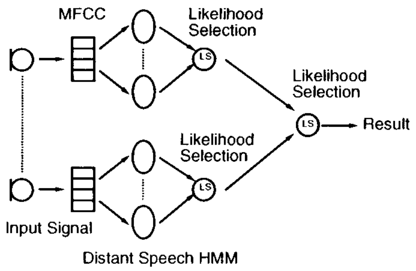

Figure 7: Microphone selection method based on maximum likelihood.

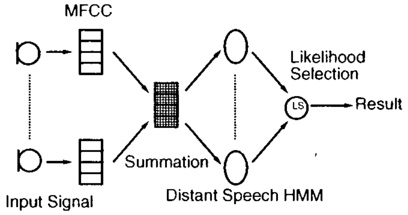

Figure 8: Feature summation method.

We also performed an experiment using the test speech actually recorded by playing back the original test speech data in the room. The difference between the simulated and real acoustic conditions is whether the temporal variation of the room acoustics is taken into consideration or not. From the results, the difference between the simulated and real acoustic conditions was about 3% in accuracy, in front of the loud speaker.

5. FEATURE-BASED AND LIKELIHOOD-BASED INTEGRATION OF MULTIPLE INPUTS

Finally, we will perform space diversity speech recognition using distributed multi-microphone. To perform space diversity, speech recognizer has to integrate multiple inputs from the acoustic environment so that the most reliable result is selected by any criterion, because the positions of speaker and microphones are assumed to be unknown. In this research, we propose the following two techniques:

1. microphone selection method based on maximum likelihood,
2. feature summation method.

5.1. Microphone selection method based on maximum likelihood

In this method, the most reliable distant speech model and acoustic channel, i.e., microphone is selected based on likelihood. At first, the distant speech model selection by maximum likelihood is performed every microphone. Then, the microphone whose likelihood is maximum is selected as the most reliable microphone (Figure 7).

5.2. Feature summation method

Before recognizing input speech by the distant speech models, the feature parameters (MFCC) extracted from all of microphone signals are averaged in every dimension of feature vector. Then, the recognition is performed using the seven distant speech models.

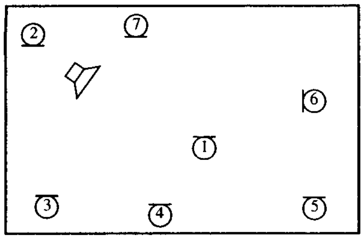

Figure 9: Distributed multi-microphone arrangement.

Finally, the model whose likelihood is maximum is selected as the most reliable model (Figure 8). Note that the distant speech models are not re-trained using the averaged feature vectors.

5.3. Experimental conditions

The seven distant speech models described in Section 2 were used. Each microphone was arranged at the arbitrary position in room as shown in Figure 9. The room is the same used in the previous sections. The test speech data used in the previous sections were convolved with the impulse responses between sound source and each microphone. The other experimental conditions are the same as those in Section 2.

5.4. Experimental results

The speech recognition results are shown in Figure 10. The results of the four methods are compared:

1. The microphone selection method based on maximum likelihood (LS of Mic),
2. The feature summation method (FS),
3. The distant speech model selection based on maximum likelihood (LS of HMM),
4. IPA HMM described in Section 2 (IPA HMM).

The experimental results show that the proposed space diversity speech recognition system can attain about 80% in accuracy, while the performances of conventional HMM are less than 50%. The microphone selection based on maximum likelihood outperforms the feature summation about 3%. The recognition rate for the microphone behind the loud speaker degrades because no distant speech model behind the roud speaker is used. However, the microphone selection method based on maximum likelihood and the feature summation method are not influenced by the microphone of the low speech recognition rate, and keep the high speech recognition rate.

6. CONCLUSIONS

In this paper, we investigated space diversity speech recognition using distributed multi-microphone. In special, we examined the following three points: (1) the basic performance of distant speech model, (2) the recognition rate in room in the case of using distant speech model, (3) the feature-based and likelihood-based integration method. From experimental results, we clarified that the proposed space diversity speech recognition system can attain about 80% in accuracy, while the performances of conventional HMM that is not learned any room acoustics are less than 50%.

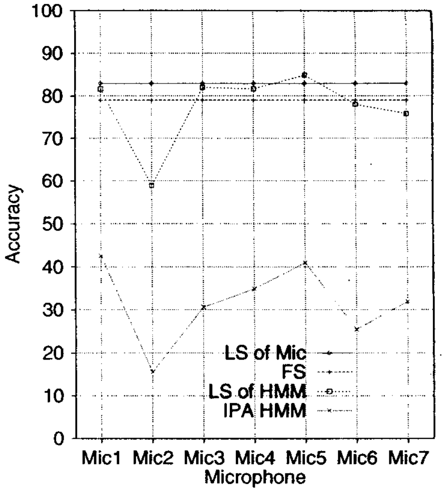

Figure 10: Results of the space diversity recognition using distributed multi-microphone.

However, in order to develop in real-world applications, the following additional works must be done: (1) the incorporation of the distant speech model behind and at the side of speaker, (2) the development of any criterion in arranging distributed multi-microphone in real situations, (3) the evaluation for background noises and the compensation method.

ACKNOWLEDGMENTS

This work is partially supported by Grant-in-Aid for COE Research of Ministry of Education, Science, Sports and Culeune in Japan.

7. REFERENCES

[1] Q. Lin, C.-W. Che, B. de Vries, J. Pearson and J. Flanagan: "Experiments on distant-talking speech recognition", Proceedings of the Spoken Language Systems Technology Workshop, pp. 187–192 (1995).

[2] T. B. Hughes, H.-S. Kim, J. H. Dibiase and H. F. Silverman: "Using a real-time, tracking microphone array as input to an HMM speech recognizer", Proc. of ICASSP, Vol. 1, pp. 249–252 (1998).

[3] T. Yamada, S. Nakamura and K. Shikano: "Hands-free speech recognition with talker localization by a microphone array", Transactions of Information Processing Society of Japan, **39**, 5, pp. 1275–1284 (1998). in Japanese.

[4] T. Kawahara, T. Kobayashi, K. Takeda, N. Minematsu, K. Itou, M. Yamamoto, A. Yamada, T. Utsuro and K. Shikano: "Sharable software repository for Japanese large vocabulary continuous speech recognition", Proc. of ICSLP, Vol. 7, pp. 3257–3260 (1998).

[5] T. Kobayashi, S. Itahashi, S. Hayamizu and T. Takezawa: "ASJ continuous speech corpus for research", J. Acoust. Soc. Jpn., **48**, pp. 888–893 (1992).