Acoust. Sci. & Tech. **24**, 6 (2003)

ACOUSTICAL LETTER

Hands-free speech recognition in real environments using microphone array and 2-levels MLLR adaptation as a front-end system for conversational TV

Masakiyo Fujimoto*, Yasuo Ariki and Shuji Doshita

Department of Electronics and Informatics, Faculty of Science and Technology, Ryukoku University, 1–5, Yokotani, Oe-cho, Seta, Otsu, 520–2194 Japan

(*Received 24 February 2003, Accepted for publication 13 March 2003*)

Keywords: Conversational TV, Hands-free ASR, Microphone array, 2-levels MLLR adaptation
PACS number: 43.72.Kb, 43.72.Ne [DOI: 10.1250/ast.24.379]

1. Introduction

Recently, many TV news programs are broadcast throughout the world. They are broadcast in one-way from broadcasting stations to the viewer (user). In this situation, the user cannot obtain the details of the interesting information when it appears in the TV news. Even when retrieving the information through the internet, the user has to activate the internet clients for the retrieval. To retrieve the interesting information without the inconveniences, a conversational TV is desired which can retrieve the details of the interesting information through man-machine interaction [1].

In enquiring the details of the interesting information to the TV, input devices such as keyboard or mouse are not suitable for the conversational TV, because we would like to enjoy the content, as opposed to personal computers. Therefore in our system, speech recognition is employed instead of keyboard and mouse, because speech is a strong tool to communicate with human or machines. Here, using a microphone is cumbersome because we have to grasp and bring the microphone to near the mouth before speaking. Wearing a headset microphone is unnatural, because we would like to feel easy when watching the TV. From this viewpoint, we propose the hands-free speech recognition using a microphone array as a front-end system of the conversational TV.

In the proposed hands-free speech recognition system, the DOA (Direction Of Arrival) of a target speech signal (user utterance) is estimated by using a microphone array and the target signal is enhanced by beam forming. Then, time section of the user utterance is detected automatically from continuously observed signals based on the time stability of the DOA. Furthermore, by applying a 2-levels MLLR (Maximum Likelihood Linear Regression) [2] based acoustic model adaptation, the enhanced speech signal is recognized accurately.

2. Hands-free speech recognition

Figure 1 shows the overview of our proposed hands-free speech recognition as a front-end system of the conversational TV. In the figure, the DOA of target speech signal is estimated by using a microphone array and the target speech signal is enhanced by beam forming. Then, the time section of user utterance is detected automatically from the continuously observed signal. Furthermore, by applying a 2-levels MLLR based noise adaptation, the enhanced speech signal is recognized accurately.

2.1. Direction estimation of arriving signal and beam forming

To capture the target speech signal with high quality, it is required for a microphone array to form the directivity to the target speech signal and suppress the other signals. In this paper, a delay and sum beam former [3] is employed to capture the target speech signal with high quality. Then the DOA is estimated by using a CSP(Cross power Spectrum Phase analysis) method [4].

2.2. Detection of user utterance

In the conversational TV, it is supposed that the user inquires to the TV about the interesting information under TV news sound. Therefore, to recognize the user utterance in a hands-free mode, a speech input interface is required to barge into TV news sounds. Generally, this speech input interface is realized by detecting the time section of user utterance from continuously observed signal. We propose here the user utterance detector based on time stability of the DOA.

In this paper, it is supposed that the TV news sounds arrive from the back of the microphone array, as shown in Fig. 2. In this case, the TV news sounds are captured with various reflections at the microphones. Therefore, the DOA of the TV news sounds is not stable in the time sequence. On the other hand, the DOA of the user utterance is stable because it arrives from the front of the microphone array. Moreover, it is assumed that the user utterance section continues on more than 1 second, because comparatively long sentences such as "what is ..." or "please tell me about ..." are usually spoken in the conversational TV. Under these assumptions, time section with more than 1 second DOA stability is detected as the user utterance section as shown in Fig. 3.

2.3. 2-levels MLLR adaptation

The beam forming can reduce sounds other than the target speech. However, the background noise is still superimposed on the waveform after beam forming. To cope with the residual background noise, we propose here a 2-levels MLLR adaptation. The first level of the 2-levels MLLR is the adaptation of HMMs to the residual noise after the beam forming using 15 sentences spoken by 5 males (noise adaptation). The second level of the 2-levels MLLR is the adaptation of HMMs to an individual speaker using 3 sentences spoken by each user after the first level noise adaptation (speaker adaptation).

*e-mail: masa@arikilab.elec.ryukoku.ac.jp

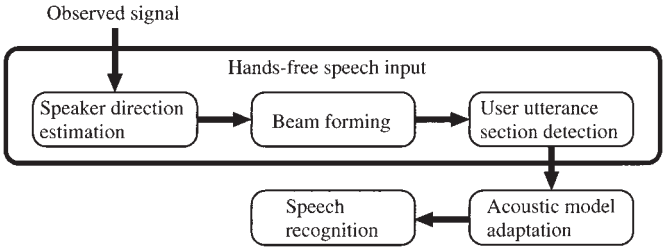

Fig. 1 Processing diagram of hands-free speech recognition.

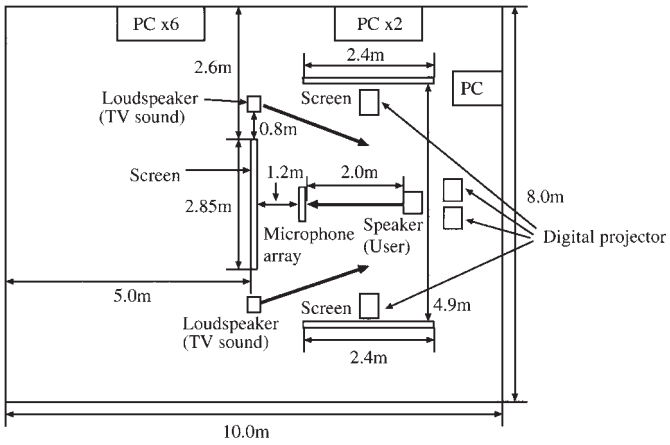

Fig. 2 Experimental room environment.

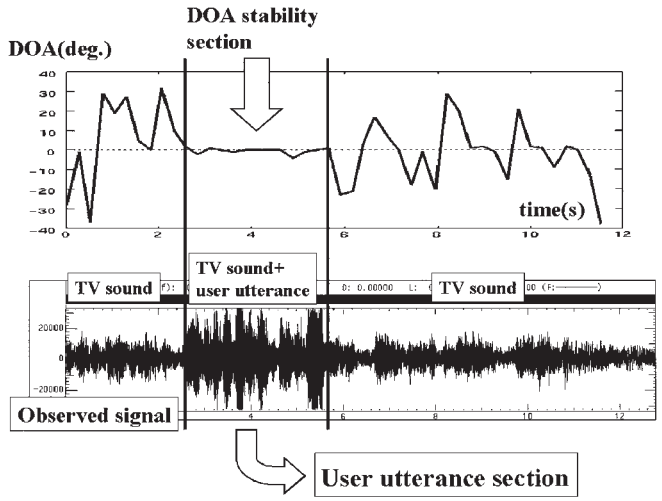

Fig. 3 An example of detection of user utterance.

3. Experiments

We evaluated the hands-free speech recognition in the conversational TV environment.

3.1. Experimental setup

The experimental materials are 100 sentences spoken by 5 Japanese male subjects and include 20 keywords appeared in the TV news. Each utterance is spoken by the subject in front of a microphone array. The distance from the subject to the microphone array (linear type, 16 microphones and 2 cm intervals) is 2 m and sampling rate is 16 kHz. By using these materials, we evaluated the hands-free speech recognition by sub-word model based keyword spotting. The noise sources are the TV news sounds and the fan noise generated from 4 digital projectors and 9 PCs (Noise level is about 55 dB(A)). The TV news is NHK TV news broadcast at 12:00 on November 30 in 2001.

The acoustic models used in the speech recognition are the speaker independent monophone HMMs (3 states and 12 mixtures). They were trained using 21,782 sentences spoken by 137 Japanese males. These speech data were taken from the JNAS (Japanese Newspaper Article Sentences) database. The feature parameters are composed of 39 MFCCs with 12 MFCCs, Log-energy and their first and second order derivatives (frame length: 20 ms and frame shift: 10 ms).

3.2. Experimental results

In the experiments of time section detection of the user utterance and the DOA estimation, 89 user utterance sections were correctly detected and the DOA estimation rate to the correctly detected sections was about 95.5%. Here, the DOA estimation rate allows the estimation error of ±5 degrees.

Tables 1 and 2 show the keyword extraction rate of the correctly detected sections. Here, the baseline keyword extraction rate (without adaptation) was 48.3%.

In Table 1, by using only supervised noise adaptation, the keyword extraction rate was about 80.9%. On the other hand, in Table 2, by using supervised noise and speaker adaptation, the best keyword extraction rate 83.2% was obtained.

However, in Table 2, the results by unsupervised speaker adaptation after noise adaptation were inferior to the results by only noise adaptation as shown in Table 1. The reason of these degradation will be explained as follows. The accurate phoneme labels for adaptation data were not obtained because

Table 1 Keyword extraction rate by conventional MLLR adaptation (1-level MLLR adaptation) (%).

	Unsupervised adaptation	Supervised adaptation
Noise adaptation	68.5(61/89)	**80.9(72/89)**
Speaker adaptation	67.4(60/89)	69.7(62/89)

Table 2 Keyword extraction rate by 2-levels MLLR adaptation (%).

	Unsupervised speaker adaptation (2nd level)	Supervised speaker adaptation (2nd level)
Unsupervised noise adaptation (1st level)	55.1(49/89)	78.7(70/89)
Supervised noise adaptation (1st level)	71.9(64/89)	**83.2(74/89)**

the speech recognition results includes some errors and is used for the unsupervised adaptation. From this fact, it is required for the unsupervised adaptation to be robust for the errors of phoneme labels.

4. Conclusions

In this paper, we proposed a hands-free speech recognition using microphone array and 2-levels MLLR adaptation as a front-end system of conversational TV. As the experimental results, the proposed method showed 89.0% of the time section detection rate of the user utterance, 95.5% of the DOA estimation rate and 83.2% of the keyword extraction rate. In future, to improve the speech recognition rate in hands-free environments, we are planning to develop the more robust acoustic model adaptation method.

References

[1] M. Fujimoto and Y. Ariki, "Noise robust hands-free speech recognition using microphone array and kalman filter as front-end system of conversational TV," *CD-ROM Proc. MMSP2002* (2002).

[2] C. L. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech Lang.*, **9**, 171–185 (1995).

[3] J. L. Flanagan, J. D. Jhonston, R. Zhan and G. W. Elko, "Computer-steered microphone arrays for sound transduction in large rooms," *J. Acoust. Soc. Am.*, **78**, 1508–1518 (1985).

[4] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," *Proc. ICASSP '94*, Vol. 1, 273–276 (1994).