Development of Infrared Lip Movement Sensor for Spoken Word Recognition

Takahiro YOSHIDA, Seiichiro HANGAI
Department of Electrical Engineering, Tokyo University of Science
1-14-6 Kudan-kita, Chiyoda-ku, Tokyo, 102-0073, JAPAN

ABSTRACT

Lip movement of speaker is very informative for many application of speech signal processing such as multi-modal speech recognition and password authentication without speech signal. However, in collecting multi-modal speech information, we need a video camera, large amount of memory, video interface, and high speed processor to extract lip movement in real time. Such a system tends to be expensive and large. This is one reasons of preventing the use of multi-modal speech processing.

In this study, we have developed a simple infrared lip movement sensor mounted on a headset, and made it possible to acquire lip movement by PDA, mobile phone, and notebook PC. The sensor consists of an infrared LED and an infrared photo transistor, and measures the lip movement by the reflected light from the mouth region.

From experiment, we achieved 66% successfully word recognition rate only by lip movement features. This experimental result shows that our developed sensor can be utilized as a tool for multi-modal speech processing by combining a microphone mounted on the headset.

Keywords: Multi-modal Automatic Speech Recognition, Lip Movement, Infrared Sensor

1. INTRODUCTION

The information of speaker's lip movement is very effective for many application of speech signal processing, because the lip and mouth are moved along utterances of words and sentences.

In the research area of multi-modal automatic speech recognition (ASR), the addition of speaker's lip movement to the speech signal is effective approach for robust recognition against acoustic noise [1][2][3]. The information of lip movement is effective and available for not only in multi-modal ASR but also in password authentication without audio speech, and multi-modal speaker identification [4].

In such a case, however, video camera, large amount of memory, video interface and high speed processor to get the information of lip movement are necessary. In addition, such a system tends to be expensive and large. This is one of reasons to prevent the use of multi-modal speech signal processing. Especially in the mobile use of multi-modal speech signal processing on PDA and mobile phone, it is well expected that their performance is too poor to extract lip movement in real time.

Therefore, we have developed a simple infrared lip movement sensor mounted on a head set, and made it possible to acquire lip movement. The sensor consists of an infrared LED and an infrared photo transistor, and measures the speaker's lip movement by the reflected light from the mouth region. The sensor has also an advantage in small calculation cost for extracting the lip movement features.

In this study, the performance of the developed sensor is evaluated by the spoken word recognition using 50 words only with the lip movement features. In this paper, the details of the developed sensor and the experimental results of spoken word recognition are shown in comparison of previous method using video cameras.

2. RELATED WORKS

There are a few related works of using infrared sensor to get lip movement information. One of works is that of J. Huang, et al. [5]. Their headset had an infrared CCD camera to get moving pictures of speaker's mouth region. Their method needed the large calculation costs and memory to get the information of lip movement as previous method using video camera, because their headset captures moving pictures. In the excellent previous work of Z. Zhang, et al. [6], they developed prototype multi-sensory headsets. One of their prototype headsets had an infrared sensor. The main specialty of their prototype headset was the born conductive microphone. Their infrared sensor is similar to our sensor in the point of the way of lip movement sensing. However, in the paper, the lip movement information from infrared sensor was used only for voice activity detection. In addition, they did not report the details of infrared sensor and its lip movement features, because this sensor and features was not main subject for their work. We could not find any other paper or work in which the lip movement features from infrared sensor were used for ASR or multi-modal ASR.

3. DEVELOPED HEADSET WITH LIP MOVEMENT SENSOR

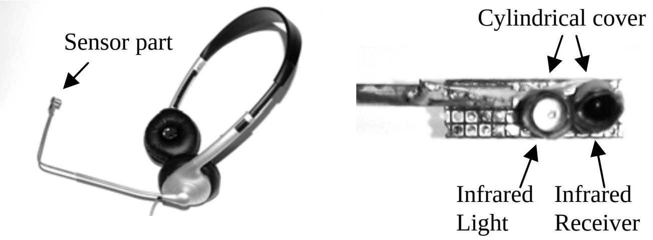
The detail of the lip movement sensor part

The whole view of the developed sensor mounted on a headset is shown in Fig. 1(a). The closed view of lip movement sensor part of the developed headset is shown in Fig. 1(b). The sensor part consists of an infrared LED for radiating infrared light and an infrared photo transistor for measuring the power of reflected infrared light. In this prototype, we use Toshiba TLN103A infrared LED device which has 80 degree half-value angle and Toshiba TPS601A infrared photo transistor device which has 10 degree half-value angle. The infrared light is radiated to the mouth region of speaker continuously, and the reflected light is measured. When the speaker opens mouth widely, the power of the reflected light is slightly reduced. When the speaker closes mouth, the reflected light becomes stronger. This means that the intensity of the sensor output shows the shape of the mouth, which corresponds to the height of the mouth. To improve the dynamic range and S/N of the lip movement sensor, we adjust the directivity of the infrared sensor by adding cylindrical cover to the circumference of the infrared LED and the photo transistor as shown in Fig. 1(b).

The details of the I/F part and extraction process

In developing the sensor, we designed a special interface (I/F), which enables to connect the sensor to PDA or notebook PC via audio input port of sound card. By this, no special interface on PC or PDA is required. This is done by amplitude modulation of the sensor output. For the application of multi-modal ASR, this sensor and a microphone mounted on a headset will become an excellent tool, because PC or PDA gets two data via standard 2 channel audio input port simultaneously.

The block diagram of process in the developed I/F and PC is shown in Fig.2. The sensor signal from photo transistor is amplified, and the amplitude modulated sensor signal is made by multiplying the sensor signal and 1 kHz sine wave, which is generated by an oscillator. The frequency of oscillator, i.e. the frequency of carrier wave, decides the maximum transfer rate of a lip movement feature. In this case, the maximum transfer rate is 1000fps. After the modulated sensor signal is captured by PC or PDA, the lip movement feature is extracted by peak-picking calculation on PC. The down sampling is easily done by a little additional calculation, if an adaptation of the frame rate for ASR engine or multi-modal ASR engine is necessary.



(a) Whole view (b) Lip movement sensor

Fig. 1 The view of the developed sensor mounted on headset

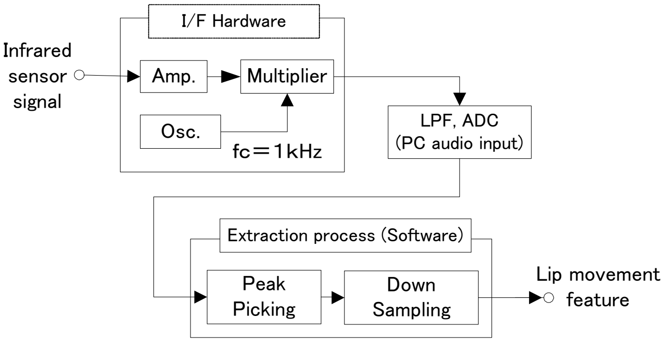


Fig.2 The block diagram of extraction process

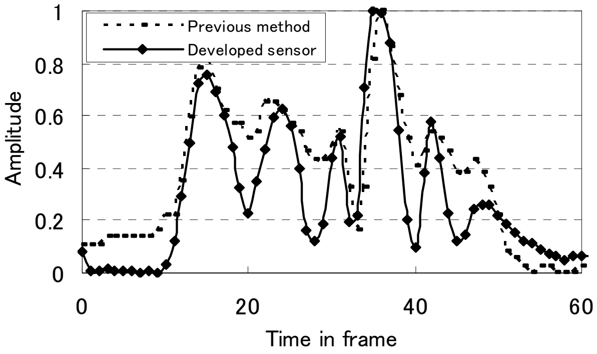


Fig.3 The example of temporal features of lip movement from the developed sensor output and lip height features extracted by previous method using video camera. (Utterance is "Tokyoto-meguroku")

Example of the sensor signal

Fig. 3 shows an example of the sensor output with solid line, when Japanese word "Tokyoto-Meguroku" is uttered. The lip height feature, which is extracted by previous method using video camera and a post image processing, is also shown with dotted line in Fig. 3. In this figure, in order to certify the correctness of the sensor output, we convert the sampling rate from 1000fps to 90fps, this is because the sampling rate in previous method is done by from 30fps to 90fps. The amplitude of each feature is normalized between 0 and 1. From this figure, it is found that the correlation between the sensor output and lip height feature which is extracted by previous method is extremely high.

4. SPOKEN WORD RECOGNITION ONLY WITH THE LIP MOVEMENT FEATURES EXTRACTED BY THE DEVELOPED SENSOR

Experimental Condition

To evaluate the effectiveness and performance of the lip movement information from the developed sensor, we examined isolated word recognition of 50 words only with the lip movement features from the sensor.

The experimental condition is shown in Table 1. For multi-modal test database, we record the sensor output signal as right-channel audio signal, VGA 30fps face image sequence from DV video camera, and speech signal as left-channel audio signal simultaneously. This database consists of 5 speakers and 50 Japanese words for car navigation system. Each speaker utters each word 10 times. We use the first 8 times, i.e. 2000 utterances, for training of HMMs, and the latter 2 times, i.e. 500 utterances, was used for recognition. Four-state Left-to-Right HMMs are prepared for each of the 33 Japanese phonemes. Each HMM model has single Gaussian mixture.

The lip movement features from the developed sensor are extracted automatically. As the lip movement features for spoken word recognition, we use S, i.e. the demodulated sensor output signal, dS, i.e. the temporal difference of S, and aS, i.e. the temporal difference of dS. These three features are weighted to 1:3:2 in recognition process, because its ratio is the best performance in our pre-experiment.

Lip shape features of previous method are extracted automatically from speaker's face bitmap image sequences [7]. As lip shape features, we use R, i.e. the ratio of height Y to width X (=Y/X), A, i.e. the area (=XY), dX, i.e. temporal differences of lip width X, and dY, i.e. the temporal differences of lip height Y. These four features are weighted to 1:1:1:1 in recognition process.

Experimental Results

Table 1 Experimental condition of the isolated word recognition of 50 words

HMM	4 states Left-to-Right, 33 phoneme models, HTK (HTK Tool Kit) V.3.2
Multi-modal database	Vocabulary: 50 words for car-navigation 5 persons (Man), Each 10 times (Total: 2500 utterances) 8 times for training HMMs (Total: 2000 utterances) 2 times for recognition experiment (Total: 500 utterances)
Features of developed sensor	S (infrared sensor output S), dS (temporal difference of S), aS (temporal difference of dS)
Features of previous method	R (ratio of height Y to width X, Y/X), A (area, XY), dX (temporal difference of lip width X), dY (temporal difference of lip height Y)

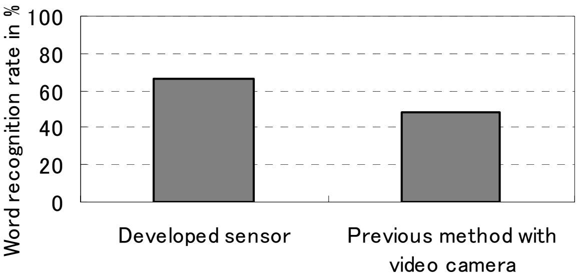


Fig.4 Comparison of the word recognition rate between the lip movement features from the developed sensor and the lip shapes features by previous method.

The word recognition rate using only lip movement features from the sensor or the lip shapes features by previous method are shown in Fig. 4. We achieved 66% word recognition rate by the lip movement features from the developed sensor without audio speech information. It's 17% higher than that of the previous method.

The reason is that the temporal resolution of the lip movement features from the sensor is high. When the speaker utters the complex word, in which phonemes change fast, the lip moves quickly and it is difficult to track the movement with normal video camera. The frame rate of the developed sensor has high frame rate up to 1000fps, so we expected further improvement for speech recognition in the future.

5. CONCLUSION

In this study, we have developed the infrared lip movement sensor. This sensor has advantages for sensing the lip movement with low hardware cost, simple structure, small calculation, and connectivity. These advantages are effective for mobile computing using PDA, mobile phone, and notebook PC. We achieved 66% word recognition rate only by the lip movement features. This experimental result shows that the developed sensor can track the speaker’s lip movement and extract the lip movement feature successfully.

In future work, we will improve a normalization method of changing lighting condition around speaker. In addition, we will add more infrared sensor to the other part of the headset for improving recognition rate.

6. REFERENCES

[1] D. Thambiratnam, et al., “Speech Recognition in Adverse Environments using Lip Information,” *IEEE TENCON ’97*, Vol.1, 1997

[2] T. Yoshida, T. Hamamoto, S. Hangai, “A Study on Multi-modal Word Recognition System for Car Navigation,” *Proc. of URSI ISSSE ’01*, pp.452-455, 2001

[3] G. Potamianos, C. Neti, G. Gravier, A. Garg, A. W. Senior, “Recent Advances in the Automatic Recognition of Audio-Visual Speech,” *Proc. of the IEEE*, vol. 91, no. 9, Sep. 2003

[4] T. Wark, S. Sridharan, V. Chandran, “The Use of Temporal Speech and Lip Information for Multi-Modal Speaker Identification via Multi-Stream HMM’S,” *Proc. of ICASSP 2000*, Vol. 6, pp.2389-2392, 2000

[5] J. Huang, G. Potamianos, C. Neti, “Improving Audio-Visual Speech Recognition with an Infrared Headset,” *Proc. of AVSP*, pp. 175-178, 2003

[6] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. Huang, Y. Zheng, “Multi-Sensory Microphones for Robust Speech Detection, Enhancement and Recognition,” *IEEE ICASSP*, 2004

[7] M. Omata, H. Machida, T. Hamamoto, S. Hangai, “Extraction of Lip Shape for Recognition Phonetics under Noisy Environment,” *Proc. of National Conf. of IEICE*, D-12-65, pp235, 2000