Noise-Robust Speech Recognition Technologies in Mobile Environments

Mobile environments are highly influenced by ambient noise, which may cause a significant deterioration of speech recognition performance. We conducted research related to "advancement of multi-modal speech recognition" and on "noise processing technologies" for the purpose of improving speech recognition in the presence of noise in mobile environments. This research was conducted jointly with the Furui laboratory (Professor Sadaoki Furui), the Department of Computer Science, Graduate School of Information Science and Engineering, Tokyo Institute of Technology.

Zhipeng Zhang, Tomoyuki Ohya and Toshiaki Sugimura

1. Introduction

In mobile environments, user input and operation via voice are simple and effective. However, the speech recognition performance is highly influenced by ambient noise in the mobile environments, which may cause a significant deterioration of the performance, and there is a strong demand for improvement of the performance. In order to resolve these issues, we conducted research in two directions: "advancement of multi-modal speech recognition" and "noise processing technologies."

1) Research Related to Advancement of Multi-Modal Speech Recognition

a) Multi-modal speech recognition using side-face moving image data

We propose a multi-modal speech recognition method that uses side face and lip moving image data as part of a noise-robust speech recognition method in mobile environments. This method uses side face image data and allows the user to input voice in a natural posture. **Figure 1** shows a configuration of the multi-modal speech recognition proposed. In this method, acoustic and image data are merged using a multi-stream Hidden Markov Model (HMM), to improve the recognition performance.

b) Examination of stream-weight optimization method in multi-modal speech recognition

In the multi-modal speech recognition of acoustic and image information using the aforementioned multi-stream HMM, we propose to optimize a normalized likelihood criterion, and confirm that the error ratio can be reduced by approximately 40% when the volume of sample data is small compared to the conventional likelihood-ratio maximization method.

2) Research on Noise Processing Technologies for Speech Recognition

The signal input to the recognition system is continuous without any decisive information about end-points. Techniques for automatically recognizing continuous speech signal are necessary. For this reason, we propose a methodology to automatically and robustly detect utterance intervals under conditions where the Signal to Noise Ratio (SNR) changes over time, based on a tree-structured noise overlay speech model, and confirm that the proposed method does improve the speech recognition performance.

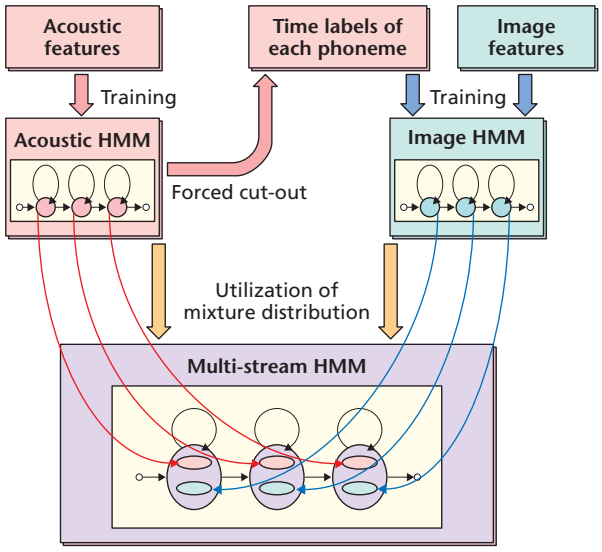

Figure 1 Configuration of multi-modal speech recognition

2. Multi-Modal Speech Recognition Using Side Face Moving Images

The multi-modal speech recognition system, which utilizes data obtained from moving images of lips at utterance recorded together with the speech data itself, is attracting attention as one of the speech recognition methodologies that are not sensitive to acoustic noise, and research has recently been carried out in this field [1]. Conventionally, in these research activities, lip images taken from the front have been utilized. When using this approach in mobile environments, however, various problems arise. For instance, in order to acquire speech and image data using a mobile terminal equipped with a camera, the user is required to take photos, holding the mobile terminal in front of his/her face while talking. This is highly inconvenient for the user and he or she cannot provide natural voice input. Moreover, it is necessary to keep some distance between the mobile terminal and the mouth to take the photos, which lowers the SNR of the speech data.

For this reason, we propose a multi-modal speech recognition method that uses side face moving image data, rather than front face images. In this method, it is assumed that a micro-camera is mounted at the microphone part of the mobile terminal or other mobile device to acquire the lip moving image data. Using image data of the lip movement seen from the side in the speech recognition allows the user to give voice in put in a natural posture, and the speech can be recognized without being disturbed by noise in the mobile environments. Moreover, a new feature extraction method by lip line extraction was proposed using a model of the lip outline seen from the side of the face. In the following, lip line extraction is explained, along with the multi-modal speech recognition methodology, evaluation conditions and test results.

2.1 Lip Line Extraction

In the context of our study, lip lines refer to "two lines drawn to include the largest area of the upper and lower lip, respectively, from the datum point, i.e., the leftmost point of the lips seen from the side" (**Figure 2**). Note that this study uses lip images taken from the right side of the speaker's face, which are extracted via the following three steps [2].

1) Lip Area Extraction

The lip area is extracted in the following procedure: i) determine the left and right edges, and then the upper and lower edges within the initial image; ii) generate an image for area extraction; and iii) remove all pixels other than the lips from the hue image; and iv) determine the left and right edges, and then the upper and lower edges within the hue image.

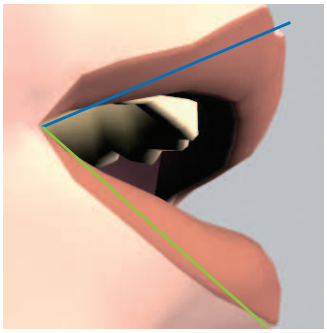

Figure 2 Example of lip lines

2) Datum Point Extraction

The datum point of the lip lines is determined from the lip area image obtained in 1). First, the internal parts of the lips are extracted from the original image. The internal parts of the lips constitute the darkest part of the image; this area is thus extracted by binarizing the image based on the brightness values. The white area within the lips is the obtained area.

3) Determination of the Upper/Lower Lip Line

From the lip area and datum point obtained above, a search start line is placed in the hue image and rotated clockwise to find the line position where the number of pixels is the largest; this is set as the lower lip line. The upper lip line is determined in a similar manner.

2.2 Speech Recognition Experiment Using Side Face Moving Images

1) Features

Voice data were acquired using a PC, at a low sampling frequency of 16 kHz. Then, the data were separated into utterance intervals with a frame length of 25 ms and a frame rate of 100 Hz, and Mel-Frequency Cepstrum Coefficients (MFCC) and normalized logarithmic power were extracted. Each image was a 24-bit full-color image of 720 (width)×480 (height) pixels taken by a digital video camera. It was first scaled down to a resolution of 180×120 pixels, and the lip lines were extracted from this shrunk image. Two features, the angle formed by the lip lines and the accompanying rate of change, were then calculated for each frame image to obtain a two-dimensional set of image features.

In this methodology, the acoustic and image features are merged in the parameter phase to generate acoustic-visual features to be used in recognition. In this experiment, the frame rate of the acoustic features was 100 Hz, while that of the image features was 30 Hz. It was thus not possible to merge the

features as is. For this reason, the image features were interpolated in time using cubic-spline functions to obtain inter-sample values at the same rate as the acoustic features, i.e., 100 Hz, and these features were then merged frame by frame.

2) Construction of Multi-Stream HMM

In this study, first acoustic and then image HMMs were trained separately and then merged to construct the multi-stream HMM.

3) Experimental Results

In the experiment, data of 11 male speakers reading out continuous numbers were recorded in a noise-free environment and used for both acoustic and image training data. Utterance of numbers containing two to six digits by each speaker, recorded in a car driving on a highway, was utilized as test data. Acoustic noises observed in the test data were engine noise, wind noise, direction indicator noise etc., and the SNR was approximately 10 to 15 dB.

Looking at the results of using the upper/lower lip line features, the absolute value was improved by 8.0% compared to the result obtained using acoustic data only; it was thus confirmed that the proposed image features have a positive impact on the speech recognition performance.

3. Stream-Weight Optimization Method in Multi-Modal Speech Recognition

In stream-weighted multi-stream HMMs, it is necessary to set weight coefficients appropriately for the acoustic and image streams according to the noise conditions, in order to obtain better recognition performance. For this reason, a set of weight coefficients is estimated adaptively according to a likelihood-ratio. We propose a stream-weight optimization method using a normalized likelihood criterion.

3.1 Stream-Weight Optimization Method

1) Multi-Stream HMM Using Acoustic/Image Features

This study uses a multi-stream HMM consisting of acoustic stream and image stream data as the underlying model for speech recognition. In multi-stream HMMs, the logarithmic likelihood $b_w(O_t)$ of word w given acoustic/image feature O_t is defined as in equation (1):

$$b_w(O_t)=b_{Aw}(O_{At})^{\lambda_{Aw}} \times b_{Vw}(O_{Vt})^{\lambda_{Vw}} \quad (1)$$

Where t denotes time, $b_{Aw}(O_{At})$ is the logarithmic likelihood of word w given acoustic feature O_{At}, $b_{Vw}(O_{Vt})$ is the logarithmic likelihood of word w given image feature O_{Vt}, and λ_{Aw} and λ_{Vw} are the weights of the acoustic/image stream of word w used in the HMM. The following constraints must be satisfied.

$$\lambda_{Aw}+\lambda_{Vw}=1, 0\le\lambda_{Aw}, \lambda_{Vw}\le 1 \quad (2)$$

2) Optimization using a Normalized Likelihood Criterion

In noisy environments, bias and variation occur in the likelihood of each word output from the model due to differences between the model training environment and the actual speech recognition environment, and recognition errors related to particular words tend to occur. For this reason, we examined a method to improve the recognition performance by adjusting the stream weights so that the average of the likelihood output by each model becomes constant in a certain period of time, thus suppressing the bias and variation in the output likelihood. Specifically, the weight of the acoustic stream with respect to word $d \in W$ is estimated using equation (3) [3].

$$\lambda_{Ad}=\frac{\frac{1}{NT}\sum_{t=1}^{T}\sum_{w\in W} b_{Aw}(O_{At})}{\frac{1}{T}\sum_{t=1}^{T} b_{Ad}(O_{At})} \quad (3)$$

This normalized likelihood criterion has the advantages that it does not require repeated calculation so that the calculation amount and time can be reduced. Moreover, unless the noise conditions change significantly, a highly reliable mean logarithmic likelihood can be obtained from relatively small data sets.

3.2 Recognition Experiment

1) Experiment Conditions

We used unsupervised learning for the stream-weight optimization. The evaluation database was divided into 36 data sets for individual speakers. For each data set, the stream weights were optimized using the first n utterances and the corresponding recognition hypotheses, and the speech utterances within the set were recognized using the obtained stream weight.

2) Experimental Results

In case of the normalized likelihood method, the recognition ratio improved even with small amounts of training data. The larger the data sets, the greater the improvement of recognition performance; we consistently obtained far better performance than with the conventional methods. It was confirmed that the

error ratio can be reduced by approximately 40% compared to cases where the conventional likelihood-ratio optimization method is used.

4. Noise Processing Technologies for Speech Recognition

As mentioned above, there is a demand for technologies for automatically recognizing speech with unknown utterance intervals, especially continuous input of voice data under noisy conditions. This chapter presents a method for adapting a phoneme model to noise in semi-real time in order to detect utterance intervals automatically and robustly under conditions where noise characteristics and SNR change over time and improve the speech recognition performance.

4.1 Creation of Tree-Structured Noise Overlay Speech Model

Under conditions where various types of noise and varying SNR are present, a noise overlay speech model is trained and the noise overlay speech models are combined to create a single tree structure. By tracing this tree structure from the root to the leaves and selecting the optimal model, it is possible to select the optimal noise interval space for a given input voice data set [4]. **Figure 3** shows the concept of the tree-structured model. By expressing the noise characteristics in a tree structure, the general noise characteristics are obtained in the upper layer of the tree structure and specific noise characteristics in the lower layers.

4.2 Continuous Speech Recognition

No sentence delimiting data is given explicitly in continuous input voice data. It is thus necessary to extract speech segments of a fixed length (blocks) first and perform the processing on each block. The tree structure is traversed from top to bottom for each input block and the optimal model is selected, thereby selecting the optimal HMM for the input voice data, which is then recognized. Based on the recognition result, the utterance interval is detected as follows: if there are any delimiters in the recognition result, the position where the first delimiter is encountered is set as the start of the utterance interval; and if there is no delimiter in the recognition result, the end point of this speech block is set as the end of the utterance interval. Moreover, the selected model is adapted using Maximum Likelihood Linear Regression (MLLR)* [5] and the final result is output.

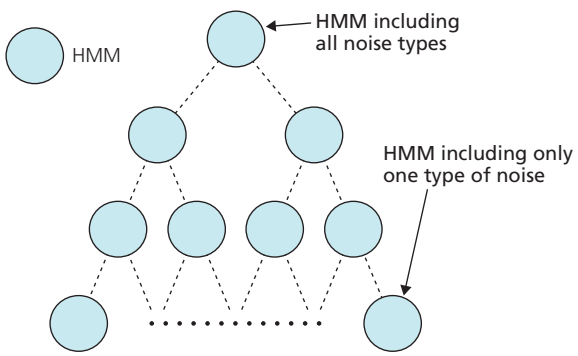

Figure 3 Concept of tree-structured model

4.3 Recognition Experiment

1) Experiment Conditions

The speech recognition experiment was carried out using a task of searching for restaurants via voice input into a dialog system. The acoustic model used was a triphone HMM with 2,000 states and 16 Gaussian mixtures in each state. As test data, we prepared data sets consisting of noisy sound recordings obtained at a station, which were not used in the training, were overlaid at three levels of SNR (SNR=5, 10, 15 dB) on 50 dialog speech data samples uttered by 10 speakers.

2) Experimental Results

The speech experiments were evaluated using the following performance measures: "recognition rate using HMM trained by clean speech" (clean HMM method), "recognition rate upon specifying correct sentence break data" (specified break method) and "recognition rate upon detecting utterance interval by the proposed method" (proposed method). **Figure 4** shows the recognition performance (Acc %) under these three conditions. The proposed method shows significantly better performance than the clean HMM method under all noise conditions. Moreover, it can be seen that a performance close to that of the specified break method, which is the target performance, is obtained with the proposed method.

5. Conclusion

This study proposed and examined "multi-modal speech recognition using side face moving images" and "weight optimization using a normalized likelihood criterion" as approaches to advancement of multi-modal speech recognition. It was confirmed that they function effectively, especially in case of small data sets. The future issues include application of multi-modal

* MLLR: Linear model parameter adaptation technique based on likelihood maximization

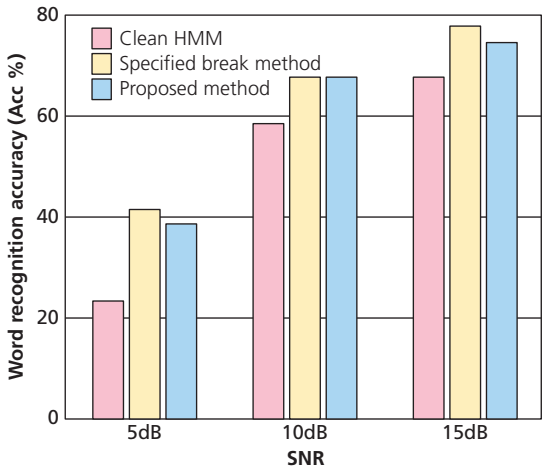

Figure 4 Recognition accuracy for three types of performance measures (Acc %)

speech recognition to large-vocabulary continuous speech recognition and construction of real-time multi-modal speech recognition systems.

Moreover, as a part of our research on noise processing technologies for speech recognition, we proposed a method for "speech recognition under noisy conditions based on robust interval detection and model adaptation," which confirmed to achieve a high recognition performance with short delay. Future work includes refinement of the method for real-time recognition application and other related issues.

REFERENCES

[1] C. Bregler and Y. Konig: " 'Eigenlips' for robust speech recognition," Proc. ICASSP 94, Vol. 2, pp. 669–672, 1994.

[2] T. Yoshinaga, S. Tamura, K. Iwano and S. Furui: "Audio-visual speech recognition using lip movement extracted from side-face images," Proc. AVSP 2003, pp. 117–120, 2003.

[3] S. Tamura, K. Iwano and S. Furui: "A stream-weight optimization method for multi-stream HMMs based on likelihood value normalization," Proc. ICASSP 2005, pp. 469–472, 2005.

[4] Z. P. Zhang and S. Furui: "Noisy speech recognition based on robust end-point detection and model adaptation," Proc. ICASSP, pp. 981–984, 2005.

[5] C. J. Leggetter and P. C. Woodland: "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, Vol. 9, No. 2, pp. 171–185, 1995.

ABBREVIATIONS

HMM: Hidden Markov Model
MFCC: Mel-Frequency Cepstrum Coefficient
MLLR: Maximum Likelihood Linear Regression
SNR: Signal to Noise Ratio