Noisy Speech Recognition Based on Robust End-point Detection and Model Adaptation

Zhipeng Zhang[1] and Sadaoki Furui[2]

[1] Multimedia Laboratories, NTT DoCoMo
3-5 Hikari-no-oka, Yokosuka, Kanagawa, 239-8536 Japan
zzp@mml.yrp.nttdocomo.co.jp

[2]Department of Computer Science, Tokyo Institute of Technology
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
furui@cs.titech.ac.jp

Abstract

How to detect speech periods in noisy speech and how to cope with the temporal variation of noise characteristics are challenging problems. This paper proposes a new robust noisy speech recognition method based on robust end-point detection and online model adaptation using tree-structured noisy speech HMMs. The basic algorithm consists of 1) blind speech segmentation, 2) best matching GMM selection, 3) recognizing the speech with the HMM that corresponds to the GMM, 4) end-point detection based on the recognition results, 5) HMM adaptation based on the recognition results, and 6) re-recognition using the adapted HMM. The processes of 1) through 6) are repeated by shifting the blind segmentation window until the end of the sequence of utterances is detected. The proposed method is evaluated by noisy speech collected by a Japanese dialogue system. Experimental results show that the proposed method is effective in recognizing noisy speech under various noise conditions.

1. Introduction

Most speech recognition experiments are conducted with the premise that end-points of input speech are known in advance, even in noisy speech recognition conditions. However, in reality, the signal input to the recognition system is continuous without any decisive information about end-points. This is especially true for noisy speech. Techniques for automatically recognizing continuous speech signal are necessary.

One of the difficult barriers to noisy speech recognition is that the level and spectral characteristics of noise is always changing. Therefore, it is crucial to implement a technique that can adaptively subtract noise or modify speech models in an online process.

Techniques for recognizing continuous input speech with no sentence boundaries can be classified into two categories. One is the sequential approach, in which an input flow is segmented using a signal processing technique and then each segment is recognized [1,2]. The other is the simultaneous approach in which recognition and segmentation are conducted simultaneously [3,4]. The simultaneous approach uses acoustical models for speech and non-speech events. The input flow is recognized by these models and the segmentation decision is made based on the changing point of speech and non-speech models. Both methods work well in the clean condition. Although the simultaneous approach usually performs better than the sequential approach, its performance degrades significantly in noisy environments because the acoustical models do not match the noisy environments.

We recently proposed the piecewise linear-transformation (PLT) method for noisy speech recognition [5]. The PLT method is performed in two steps: noisy speech HMM selection from clustered noisy HMMs and linear transformation of the selected HMM; both are based on the likelihood maximization criterion. The proposed method was confirmed to be effective in recognizing noisy speech under various noise conditions.

This paper proposes the application of the PLT method to speech input having no sentence boundaries. In order to reduce the delay in obtaining the recognition results, a fixed length of the speech input is extracted and the processes of recognition, sentence segmentation, model adaptation, and re-recognition are performed simultaneously. After processing the extracted sentence units, the next fixed-length speech input is extracted and processed in the same way. Therefore, continuous speech recognition is performed with small delay.

The remainder of the paper is organized as follows. Section 2 details the proposed method. Section 3 describes the evaluation task. Section 4 describes the speech recognition experiments conducted to evaluate the proposed method. The paper concludes with a general discussion and issues related to future research.

2. PLT-based Noisy Speech Recognition

Figure 1 shows a flow diagram of the proposed method. This method consists of two parts: model construction in the training process (tree-structured noisy GMM/HMM and language model modification) and continuous speech recognition.

2.1 Language model having no sentence boundaries

A new language model was built to allow the recognition of continuous speech with no sentence boundaries. All texts in the training corpora that contain start and end marks in each sentence are connected to form a single sentence. Each pair of end and start marks is modified to suit the long pause (LP) model. The text data are then used to construct a language model. Consequently, a phone model is made for this LP model in the acoustic model.

2.2 Tree-structured noisy speech HMM construction

A tree-structured noisy speech HMM is constructed to cover various conditions of SNR as well as noise types. The root node of the tree models all noise samples and each leaf node corresponds to only one noise condition. An HMM in the leaf layer is selected if the input noise speech is close to one of the noise conditions used for training, while a model in one of the upper layers should be selected if the input noise speech differs from the noises used for training. The tree-structured hierarchical clustering method makes it easy to select the optimum model.

2.3 Continuous speech recognition

Since no sentence boundaries are given, a fixed length of the speech input is extracted. In order to reduce the delay in obtaining the recognition results, the processes of recognition, sentence segmentation, model adaptation, and re-recognition are performed simultaneously as described below. Once the extracted speech unit is processed, the next fixed length of speech input starting from the end-point of the segmented sentence is extracted and processed in the same way. Therefore, continuous speech recognition is performed with small delay.

1) Blind segmentation

From the continuous input utterance stream, a speech unit with fixed length (T) is extracted from the beginning. Each block of fixed length is processed by steps 2) through 6).

2) HMM selection by using GMM

For each block, the noise-cluster HMM that best fits the speech unit is selected by tracing the tree downward from the top (root). The selected model best fits the speech unit in terms of the noise type as well as SNR.

Since it needs a huge amount of computation to calculate the likelihood values if HMMs are used directly, GMMs are made using the same noise-added speech employed to construct the HMMs and used for model selection. The noise-adapted HMM corresponding to the selected noise-adapted GMM that yielded the largest likelihood for the speech unit is taken as the best model.

3) First pass recognition

The selected HMM and language model made described in Subsection 2.2 are used to recognize the speech unit.

4) End-point detection

According to the recognition result, end-points are determined as follows: if an LP model is found in the output result, the end-point is set at the LP position. From this position, the next blind segmentation (Step 1) is performed. If no LP model is found in the output result, the next segmentation is commences from the end of this speech unit.

5) HMM adaptation

To use a more accurate model for recognition, we apply the MLLR [6] adaptation technique. Gaussian mean parameters of the selected noise-cluster HMM are adapted to the input speech. Transform sharing over Gaussian distributions can allow all distributions in a system to be updated using the detected sentence utterance.

6) Second pass recognition

The adapted model is then used to re-recognize the sentence utterance which yields the final recognition result.

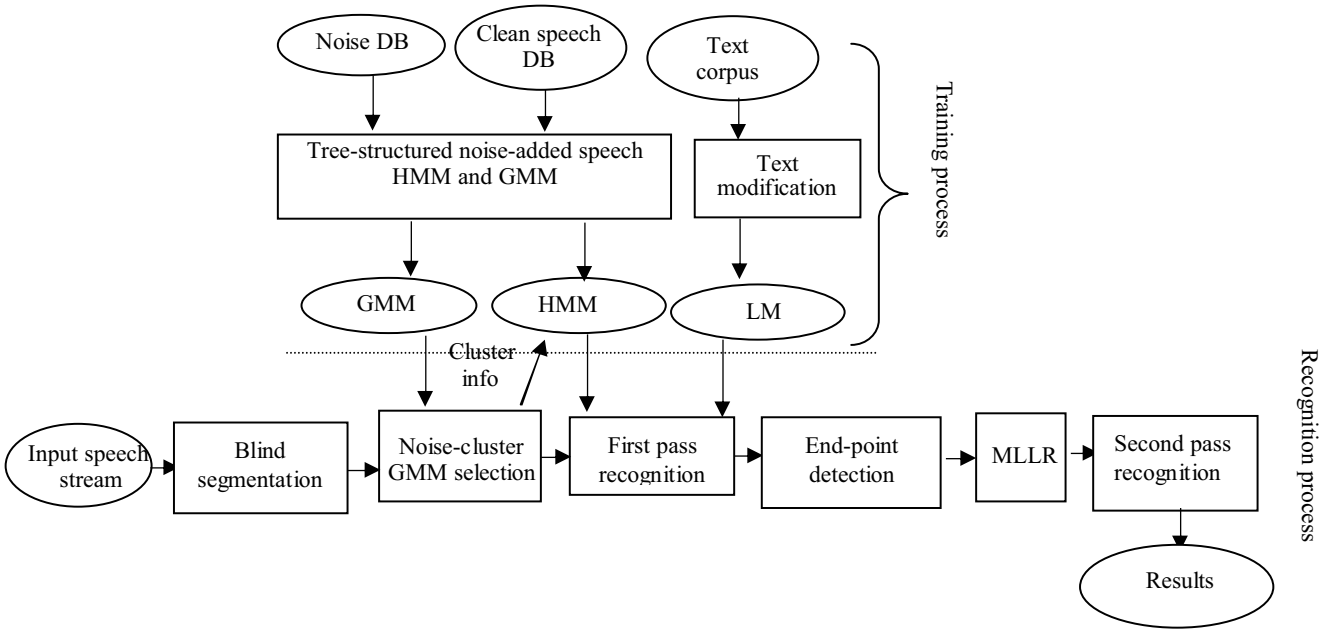

Fig. 1. System flow of the proposed method.

3. Evaluation Task

3.1 Task

The task of the recognition system was to retrieve information about restaurants and food stores. To narrow down the retrieval candidates, the subject uttered one kind of food, a station name, and conditions. A database of restaurants and food stores open to the Internet was used. The database consists of 80 business categories and holds data on about 4,091 food stores and restaurants.

3.2 Language models

Language models consisting of class bigrams and reverse class trigrams with backing-off were used. The models were trained using text corpora that were prepared separately for each dialogue content (topic) category. Some training texts were transcribed from real dialogue utterances, and other texts were manually entered by human subjects on the assumption that they were actually using the dialogue system. Several sets of words, such as numbers, store names, fillers, and prices, were grouped to make the class language models. Words belonging to each class were given an equal word occurrence probability [7].

3.3 Acoustic models

A tied-mixture triphone HMM with 2,000 states and 16 Gaussian mixtures in each state was used as the acoustic model. Utterances from 338 presentations in the "Corpus of Spontaneous Japanese (CSJ)" [8] produced by male speakers (approximately 59 hours) were used for training.

4. Experiments

4.1 Evaluation data

The following two test data were used to evaluate the proposed method:

- **Test 1**: 50 query utterances by male speakers were recorded under quiet environments. Utterance duration was distributed over 1.0~4.5 seconds and the average duration was 1.54 seconds. All 50 utterances were concatenated to yield one test utterance. Two noises, "Station" and "Hall" noises, recorded at a station concourse and a department store elevator hall, respectively, were numerically added to the copies of the test utterances at three SNR levels: 5, 10, and 15dB. Experiments were therefore performed under 6 different conditions (2 noises x 3 SNRs).

- **Test 2**: 540 query utterances from 12 speakers (45 per speaker) were recorded over three days in real noisy environments ("Station" and "Office"). Test utterance duration was distributed over 2.5~7.5 seconds and the average duration of the test utterances was 4.86 seconds ("Station" noise-added speech) and 4.89 seconds ("Office" noise-added speech). All 540 sentences and noises were concatenated to yield one test utterance. The average SNRs were 10dB ("Station" noisy speech) and 12dB ("Office" noisy speech). This task was relatively difficult, since the noise was non-stationary.

4.2 Evaluation on Test 1

a) Overall results

Recognition experiments using Test 1 data under the following three conditions were conducted:

(1) Baseline: Clean HMM was used instead of the tree-structured noise-added speech HMMs
(2) Proposed method
(3) Proposed method but sentence end-points were given (A best matching noise-added HMM was selected and further adapted to each sentence utterance according to the given end-points)

The length of the speech unit extracted for blind segmentation was set at 10 seconds.

Tables 1 and 2 show the word accuracy on the two kinds of noise added speech at the three SNR levels, 5, 10, and 15dB. These results indicate that the proposed method achieves significantly better performance than the baseline and achieves performance very close to that achieved when correct end-points are given.

Table 1. Recognition accuracies with Test 1 data for 3 conditions: the baseline, the proposed method but end-points are given, and the proposed method (Station noise-added speech)

	Baseline	Given end-point	Proposed method
5dB	23.7	40.8	38.2
10dB	57.9	67.1	67.1
15dB	67.1	76.3	73.7

Table 2. Recognition accuracies with Test 1 data for 3 conditions: the baseline, the proposed method but end-points are given, and the proposed method (Exhibition hall noise-added speech)

	Baseline	Given end-point	Proposed method
5dB	40.8	47.4	43.4
10dB	61.8	76.3	75.0
15dB	71.1	77.6	77.6

b) Impact blind segmentation unit length

Recognition experiments were performed using Test 1 data to investigate the effects of changing the length of the blind segmentation unit. In these experiments, lengths of 3,5,8,10 seconds were examined.

Figures 2 and 3 show the results for the two kinds of noise at three SNR values, 5, 10, and 15dB. These results indicate that the performance basically saturates at lengths of 5 seconds and beyond; the performance degrades at 3 seconds. The critical length coincides with the maximum length of the input query utterances.

c) Effectiveness of MLLR adaptation

Another experiment was performed using Test 1 data to investigate the effectiveness of MLLR model adaptation. In this experiment, the length of the blind segmentation unit was set at 10 seconds.

Tables 3 and 4 show the word accuracies when the MLLR adaptation is not applied. Comparing these results with that with MLLR adaptation shown in Tables 1 and 2, it can be concluded that using the MLLR adaptation after selecting the best-matching noise-added speech HMM is very effective in further adapting to the noise.

Table 3. Recognition accuracies with Test 1 data when MLLR adaptation is not applied (Station noise-added speech)

	Baseline	Given end-point	Proposed method
5dB	15.8	36.8	35.5
10dB	47.4	64.7	60.5
15dB	63.2	75.0	65.8

Table 4. Recognition accuracies with Test 1 data when MLLR adaptation is not applied (Exhibition hall noise-added speech)

	Baseline	Given end-point	Proposed method
5dB	30.3	47.4	40.8
10dB	51.3	70.3	69.7
15dB	67.1	75.0	72.4

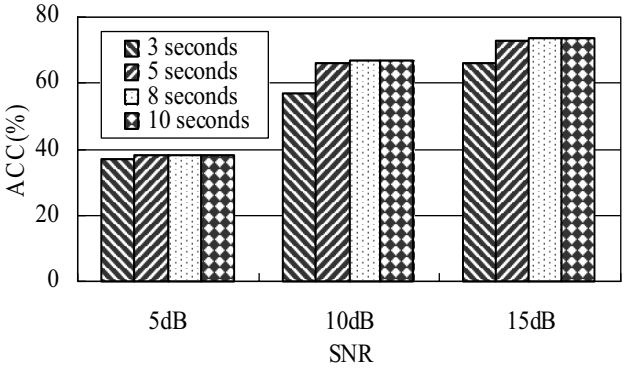

Fig. 2. Recognition accuracies on Test 1 for various blind segmentation lengths (Station noise-added speech)

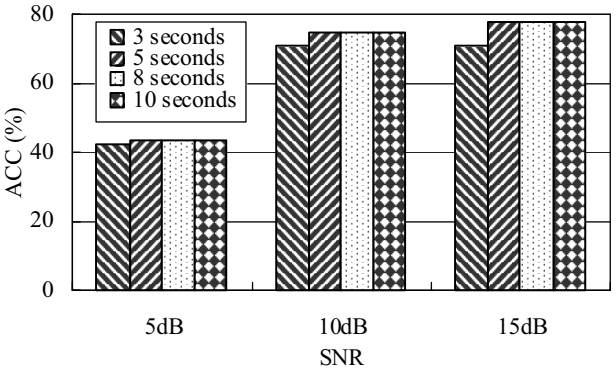

Fig. 3. Recognition accuracies on Test 1 for various blind segmentation lengths (Exhibition hall noise-added speech)

4.3 Evaluation on Test 2 Data

Recognition experiments were performed using Test 2 data. The length of the blind segmentation unit was set at 10 seconds.

Table 5 shows the recognition accuracies for the following three conditions: baseline (clean HMM was used instead of the tree-structured noise-added speech HMMs), the proposed method but correct end-points were given, and the proposed method using the estimated end-points. These results indicate that the proposed method achieves significantly better performance than the baseline and achieves performance very close to that achieved when correct end-points are given. This confirms the effectiveness of the proposed method for real-world noisy speech.

Table 5. Recognition accuracies with Test 2 data for 3 conditions: the baseline, the proposed method but end-points are given, and the proposed method

	Baseline	Given end-point	Proposed method
Station noise added speech	40.8	56.6	55.8
Office noise added speech	55.2	73.4	72.5

5. Conclusion

This paper proposes a new robust noisy speech recognition method based on robust end-point detection and online model adaptation using tree-structured noisy speech HMMs; it focuses on recognizing real-environment speech with no explicit sentence boundaries. The process consists of blind segmentation, model selection, sentence segmentation, model adaptation, and recognition. Since the process is repeated by shifting the blind segmentation unit and recognition can be completed with only a few seconds, there is no limit on the length of input speech. The proposed method has the significant advantage that it can adapt to slowly changing noise environments. The proposed method was evaluated by noisy speech collected by a Japanese dialogue system. Experimental results show that the proposed method is effective in recognizing noisy speech under various noise conditions, including that recorded under real noisy environments.

Future research includes increasing the variation of noises in both training and testing, and evaluation with other recognition tasks.

References

[1] L.R.Rabiner : "An algorithm for determining the endpoints of isolated utterances", The Bell System Technical Journal, pp. 297-315 (1975)

[2] http://www.3gpp.org

[3] A. Acero: "Robust HMM-based end-point detector", Proc. Eurospeech, pp. 1551-1554 (1993)

[4] J.G.Wilpon et al.: "Application of hidden Markov model to automatic speech end-point detection", Computer Speech and Language, pp. 321-341 (1987)

[5] Z.P. Zhang et al.: "A tree-structured clustering method intergrating noise and SNR for piecewise-linear transformation-based noise adaptation", Proc. ICASSP, pp. 981-984 (2004)

[6] C. J. Leggetter et al.: "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models", Computer Speech and Language, pp. 171-185 (1995)

[7] R.Taguma et al.: "Parallel computing-based architecture for mixed-initiative spoken dialogue", Proc. ICMI, pp. 53-58 (2002)

[8] S. Furui: "Recent advances in spontaneous speech recognition and understanding", Proc. ISCA&IEEE SSPR Workshop, pp. 1-6 (2003)