A Tree-Structured Clustering Method Integrating Noise and SNR for Piecewise Linear-Transformation-Based Noise Adaptation

Zhipeng Zhang[1],Toshiaki Sugimura [1] and Sadaoki Furui[2]

[1] Multimedia Laboratories, NTT DoCoMo
3-5 Hikari-no-oka, Yokosuka, Kanagawa, 239-8536 Japan
{zzp,sugimura}@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

This paper proposes the application of a tree-structured clustering method that integrates the effects of noise as well as SNR variation in the framework of piecewise-linear transformation (PLT)-based noise adaptation for robust speech recognition. According to the clustering results, a noisy speech HMM is made for each node of the tree structure. An HMM that best matches the input speech is selected based on the likelihood maximization criterion by tracing the tree downward from the top (root), and the selected HMM is further adapted by linear transformation. The proposed method is evaluated by applying it to a Japanese dialogue recognition system. Experimental results confirm that the proposed method is effective in recognizing numerically noise-added speech and actual noisy speech uttered by a wide range of speakers under various noise conditions.

1. Introduction

Increasing the robustness of speech HMMs (hidden Markov models) to additive noise is one of the most important issues in state-of-the-art speech recognition. A noise added speech sample, $\hat{s}$ is modeled by:

$$\hat{s} = F(s, n, SNR) \qquad (1)$$

where *s, n* and *SNR* represent clean speech signal, noise and speech-to-noise ratio, respectively. *F* represents a non-linear function in the cepstral domain. Since the noise spectrum and SNR usually vary over time, it is crucial to build a model adaptation method that can handle the non-linear effect as well as the noise variation.

Likelihood maximization is a common criterion used in model construction and model adaptation for speech recognition. Minami and Furui [1] proposed extending the PMC (Parallel Model Combination) method to variable noise by using the maximum likelihood (ML) estimation criterion. Experimental results confirmed that this method greatly improved the recognition rate when SNR and noise spectral characteristics were variable. However, this method is impractical in that it has huge computation costs and the noise HMM must be trained in advance.

We have recently proposed using piecewise linear-transformation (PLT) as an approximation of the non-linear effect of additive noise [2]. PLT is performed in two steps: noisy speech HMM selection from clustered noisy HMMs and linear transformation of the selected HMM. Both processes use the likelihood maximization criterion.

To cope with the noise and SNR variations, we have also proposed using a set of tree-structured noisy speech HMMs to handle multiple SNR conditions [3]. In this method, the effects of noise variation are modeled by tree-structured HMMs separately for each SNR condition. Therefore, the adaptation process conducts a two-step search to find the best model. In the first step, the model having the largest likelihood is selected for each SNR condition by tracking the tree downward from the top (root). Next, the best model among all SNR conditions is selected. Experimental results confirmed that this method greatly improved the recognition rate in noisy speech recognition. Although this method is an easy way to treat variations of both noise and SNR, it has a disadvantage in that it incurs large computation cost to find the best model.

This paper proposes a new clustering method that integrates noise and SNR variations simultaneously and creates a single tree. In the recognition phase, a noisy speech HMM is selected by a one-step search. We first explain the new method, and then report several experiments. The paper concludes with a general discussion and issues related to future research.

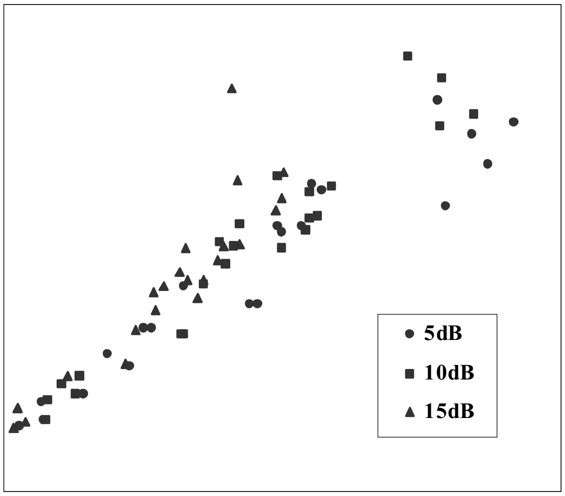

Fig 1: Analysis by Hayashi's quantification theory for noise added speech at three SNR conditions.

0-7803-8484-9/04/$20.00 ©2004 IEEE

2. PLT-based noise adaptation using tree-structured noise-adapted HMM

2.1 Tree-structured HMM construction

Figure 1 shows a projection of noise added speech data with three SNR values, 5, 10 and 15dB, and 30 kinds of noise on a two-dimensional space made by Hayashi's quantification theory [4]. The noise added speech data at 5, 10, and 15dB are indicated by circles, squares, and triangles, respectively. It is clearly shown that noise added speech data are not necessarily separated by the SNR value. In other words, noise added speech data at different SNR values are closer than different noise-added speech at the same SNR. This means that we should combine noise added speech data with different SNRs to create a single tree. Therefore, we first construct noise-added speech data by adding various noises to clean speech at multiple SNR levels. We then cluster all the noise-added speech data with all SNR conditions to build a tree, and a noisy speech HMM is made for each node in the tree.

As it is difficult to cluster noise-added speech data directly, noise-added speech GMMs for all the combinations of noises and SNRs are made and used for clustering. The noise-added speech data set corresponding to each cluster (node in the tree) is used to construct a noisy speech HMM for recognition. While the model located in the root is trained by all-noise added speech at all SNR conditions, models located in the leaves are trained by single-noise added speech at a single SNR condition. Since HMM models at the intermediate levels in the tree represent mixtures of several noises as well as several SNRs, the tree structured HMMs are expected to be robust against noise variations, even when the noise and SNR changing within an utterance.

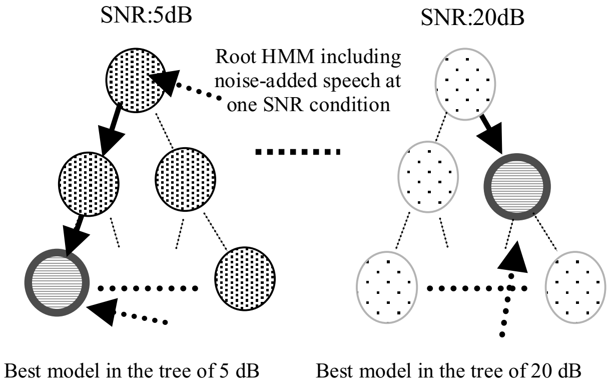
SNR:5dB
SNR:20dB
Root HMM including noise-added speech at one SNR condition
Best model in the tree of 5 dB
Best model in the tree of 20 dB

Fig 2: Tree-structured noise added speech HMM for each SNR condition.

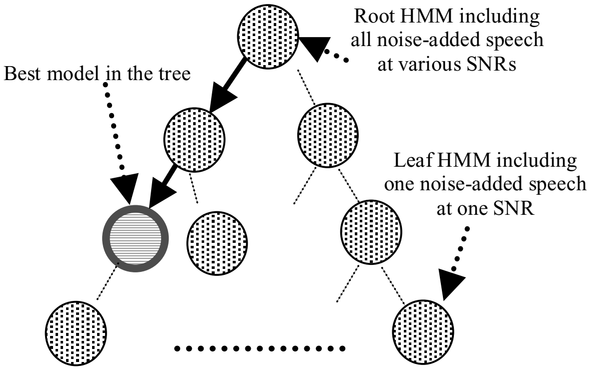
Root HMM including all noise-added speech at various SNRs
Best model in the tree
Leaf HMM including one noise-added speech at one SNR

Fig 3: Tree-structured noise added speech HMM integrating noise and SNR.

2.2 HMM selection

In our previous method, two-step search was needed to find the best model as shown in Figure 2. In our new method, a noise-cluster HMM that best fits the input speech is selected using one-step search by tracing the tree downward from the top (root) as shown in Figure 3. It is obvious that the new method greatly reduces the computation cost for model search.

2.3 Linear transformation

Gaussian mean parameters of the selected noise-cluster HMM are further adapted to the input speech as indicated by the following equation:

$$\hat{\mu} = A\mu + b \qquad (2)$$

where A is an $n \times n$ transformation matrix, μ is the Gaussian mean value, and b is an n-dimensional vector. These parameters are estimated using the MLLR method [5] so that the likelihood of the input speech is maximized. Transform sharing over Gaussian distributions can allow all distributions in a system to be updated with just a relatively small amount of adaptation data.

3. Experiments on a dialogue system

3.1 Task

The task of the system is retrieving information about restaurants and food stores. To narrow down the retrieval candidates, a user utters 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 of 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 [6].

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 "Spontaneous Speech Corpus"[7] produced by male speakers (approximately 59 hours) were used for training.

3.4 Noise data for training

28 kinds of noises collected by JEIDA (Japan Electronic Industry Development Association) were used for noise clustering [8]. Noise-added speech were made at three SNR values, 5dB,10dB, and 15dB, and noise-added speech GMM (64 mixture) was trained for each noise using the Baum-Welch algorithm. Noise-added speech GMMs were then clustered based on the likelihood matrix (84rows x 84columns) in which each term was calculated from a pair of noise GMMs.

3.5 Evaluation data

The following test data were used to evaluate the proposed method.

- **Test1**: 50 sentences uttered by male speakers were used to evaluate the proposed method. Two noises, "Station" and "Hall" recorded at a station concourse and a department store elevator hall, respectively, which differed from the 28 noise samples used for noise clustering, were numerically added to the utterances at three SNR levels: 5, 10 and 15dB. Experiments were therefore performed under 6 different conditions (2 noises x 3 SNRs).
- **Test2**: 540 sentence utterances from 12 speakers (45 per speaker) collected over three days (2003/01/20-22), were recorded in real noisy environments ("Station" and "Office") and used in the experiments. The average SNRs were 10dB ("Station" noisy speech) and 12dB ("Office" noisy speech). This task was relatively difficult, since the noise was non-stationary.

3.6 Comparison of model selection by one-step tree search and two-step tree search methods

Recognition experiments were performed on Test1 to compare two methods; two-step tree search (previous method) and one-step tree search (new method). The best matching noise-adapted HMM was selected from the tree and used to recognize the input speech. In this experiment, MLLR adaptation was not applied.

Figures 4 and 5 show the word accuracy on the two kinds of noise added speech at the three SNR values, 5, 10 and 15dB. The "Baseline" indicates the case wherein the clean HMM was used for recognition. These results indicate that the one-step tree search method gives better performance than the two-step tree search method. It was also observed that the processing time for selecting the best model using the one-step search method for Test1 data was only 1/3rd that using the two-step method.

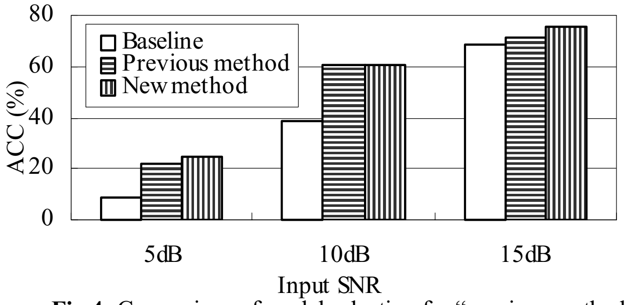

Fig 4: Comparison of model selection for "previous method" and "new method" on Test1 ("Station" noise-added speech).

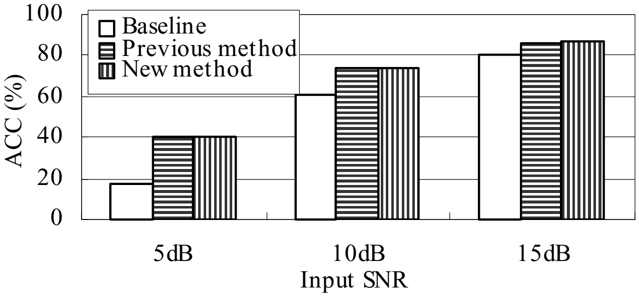

Fig 5: Comparison of model selection for "previous method" and "new method" on Test1 ("Hall" noise-added speech).

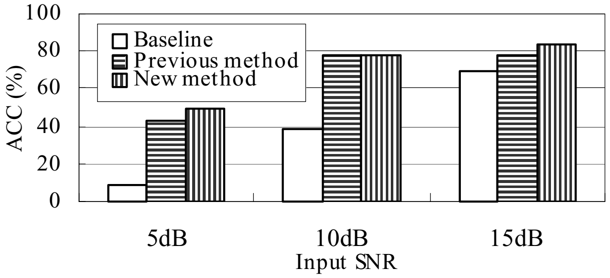

Fig 6: Comparison of PLT for previous method and new methods on Test1 ("Station" noise-added speech)

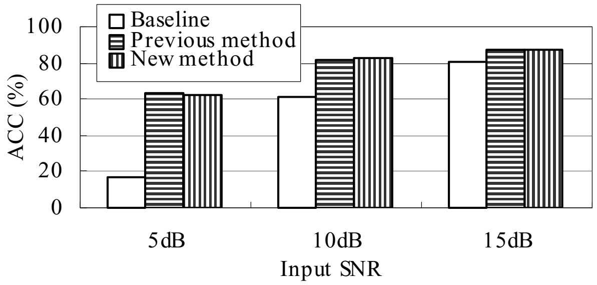

Fig 7: Comparison of PLT for previous method and new methods on Test1 ("Hall" noise-added speech).

3.7 Comparison of PLT to one-step tree search and two-step tree search methods

The PLT-based method, that is the combination of the tree-structured noise-adapted HMM selection and the MLLR-based linear transformation, was evaluated by recognition experiments.

Figures 6 and 7 show the results on Test1. These results show that the one-step tree search method gives better performance than the two-step tree search method for most cases. It reduced the word error rate by 49.8% on average relative to the "Baseline" results.

3.8 GMM-based model selection

In the experiments described above, the best model for each input noisy speech was selected from among the HMMs for the nodes in the trees. Since it needs a huge amount of computation to calculate the likelihood values using HMMs, GMMs were made using the same noise-added speech used 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 input speech was used as the best model. The MLLR method was performed using the selected noise-adapted HMM.

Figures 8 and 9 show the results for three conditions: no adaptation "Baseline", the basic adaptation method "HMM-based method", and the improved method "GMM-based method". These results show that the "GMM-based method" reduced the word error rate by 47.5%. Since the HMM-based method resulted in a 49.8% reduction, the GMM-based approach is slightly worse than the basic method, but the reduction in the computation costs made possible by using the GMM-based method is so significant that it more than makes up for the slight loss in performance.

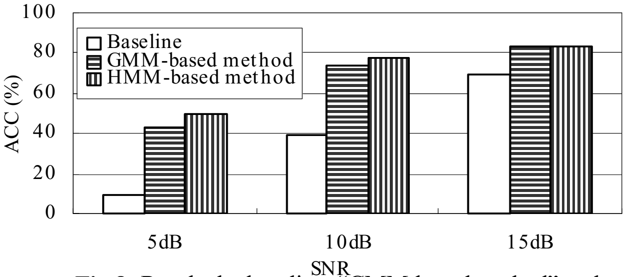

Fig 8: Results by baseline, "GMM-based method" and "HMM-based method"("Station" noise-added speech).

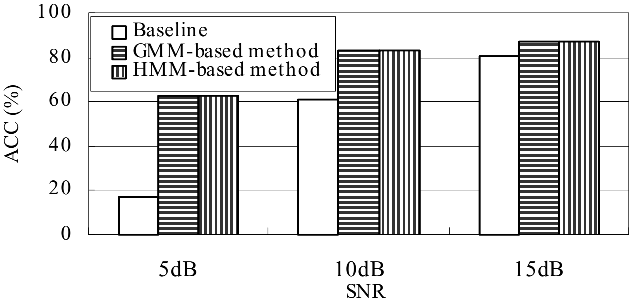

Fig 9: Results by baseline, "GMM-based method" and "HMM-based method"("Hall" noise-added speech).

3.9 Recognition result on Test2

Another experiment was performed on Test2 to evaluate the proposed method. Figures 10 and 11 show the results for three conditions: no adaptation "Baseline", the basic adaptation method "HMM-based method", and the improved method "GMM-based method". These results show that the "HMM-based method" reduced the word error rate by 35.0% while the "GMM-based method" reduced the word error rate by 33.3%.

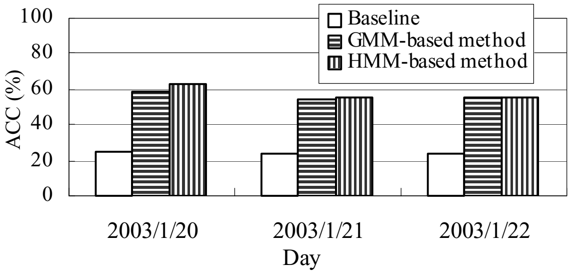

Fig 10: Results by baseline, "GMM-based method" and "HMM-based method" on Test2 ("Station" noisy speech).

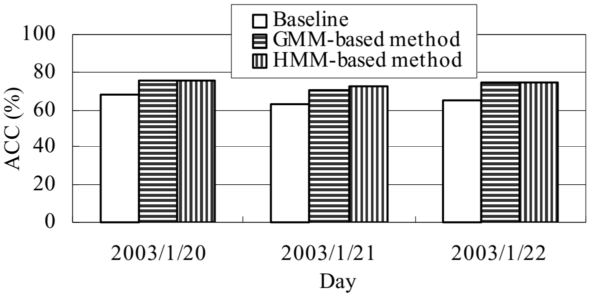

Fig 11: Results by baseline, "GMM-based method" and "HMM-based method" on Test2 ("Office" noisy speech).

4. Conclusion

This paper has reported a new tree-structured construction method that integrates noise and SNR effects simultaneously for piecewise linear-transformation (PLT)-based noise adaptation. This method consists of two parts: best matching HMM selection and linear transformation of the selected HMM. Both processes are based on the likelihood maximization criterion. The proposed method has two advantages over the previous method: improved recognition performance and reduced computation cost. The proposed method was evaluated using a dialogue system with two kinds of test data. Experimental results show that the proposed method reduced the error rate from 49.8% and 35.0% relative to that obtained with a clean speech HMM. Compared to the previous method, which uses two-step tree search, the proposed method reduces the computation cost for identifying the best HMM model by 2/3^{rds} and slightly improves the recognition performance.

Future research includes increasing the variation of noises for both training, testing, and automatic noise/speech segmentation.

References

[1] Y. Minami et al.: "A maximum likelihood procedure for a universal adaptation method based on HMM composition", Proc. ICASSP, pp. 129-132 (1995)

[2] Z.P. Zhang et al.: "Tree-structured noise adapted modeling for piecewise linear transformation-based HMM adaptation ", Proc. Eurospeech, pp. 669-672 (2003)

[3] Z.P. Zhang et al.: "Piecewise-linear transformation-based HMM adaptation for noisy speech", Speech Communication (to be published)

[4] T. Kan, "Multivariate analysis", Gendai-Sugakusha (1993)

[5] 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)

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

[7] S. Furui et al.: "Toward the realization of spontaneous speech recognition -Introduction of a Japanese priority program and preliminary results-"", Proc. ICSLP, pp. 518-521 (2000).

[8] http://www.milab.is.tsukuba.ac.jp/corpus/noise_db.html