

Tree-Structured Noise-Adapted HMM Modeling for Piecewise Linear-Transformation-Based Adaptation

Zhipeng Zhang¹, Kiyotaka Otsuji¹ and Sadaoki Furui²

¹Multimedia Laboratories, NTT DoCoMo
3-5 Hikari-no-oka, Yokosuka, Kanagawa, 239-8536 Japan
{zpz,otsuji}@mml.yrp.nttdocomo.co.jp

²Department 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 tree-structured clustering to various noise samples or noisy speech in the framework of piecewise-linear transformation (PLT)-based noise adaptation. According to the clustering results, a noisy speech HMM is made for each node of the tree structure. Based on the likelihood maximization criterion, the HMM that best matches the input speech is selected by tracing the tree from top to bottom, and the selected HMM is further adapted by linear transformation. The proposed method is evaluated by applying it to a Japanese dialogue recognition system. The results confirm that the proposed method is effective in recognizing noise-added speech 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. Research in this field has been very active. Parallel model combination (PMC, also called HMM composition) [1][2] is one of the most practical and useful methods for handling additive noise. PMC can derive a noisy speech HMM by combining a clean speech HMM, a noise HMM, and a signal-to-noise ratio (SNR). One of the problems with PMC is that it is difficult to integrate it with the cepstral mean subtraction method. To overcome this problem, we proposed a method that uses neural networks-based mapping [3]; this method was confirmed to be effective in providing recognition under various noise and SNR conditions. However, both methods have two disadvantages: one is that their computation costs, which include non-linear conversion, are high. The other is that they have difficulty in dealing with time varying noisy speech.

We have recently proposed piecewise linear-transformation (PLT) as an approximation of the non-linear effect of additive noise [4]. PLT is performed in two steps: noisy speech HMM selection and linear transformation of the selected HMM. Both processes use the likelihood maximization criterion. In a previous study [4], the number of noise clusters was decided experimentally by setting various numbers and choosing the best condition among all noise clusters. One problem that arises with

the PLT method is that it is difficult to decide the best number of noise clusters resulting in both a failure of model selection and huge computation costs.

This paper proposes a new tree-structured noise clustering method for PLT-based HMM adaptation and compares two methods for constructing the tree structure: one method is based on noise clustering and the other is based on noise-added speech clustering. We first explain the basic method, and then report some experiments. The paper concludes with a general discussion and issues related to future research.

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

2.1 Tree-structured HMM by noise clustering

Noise-added speech spectra vary as a function of both noise spectra and the signal-to-noise ratio (SNR). In the first method, which uses noise clustering, we first build a hierarchical structure of noise, and then the same structure is used to create noise-added speech HMMs (noise-cluster HMMs) for different SNR conditions. While models located in the upper layers of the tree structure represent the spectral features of global noise, models located in the lower layers represent specific features of noise. Tree-structured hierarchical noise-cluster HMMs enable easy selection of the optimum model according to the properties of the input noisy speech.

Noise clustering is performed by a procedure originally proposed for the "SPLIT" speech recognition system [5]. Since it is difficult to directly cluster noise data, we first build a GMM for each noise and cluster noise GMMs. We construct noise-added speech HMMs using a set of noisy utterances created by adding noise signals for each cluster to clean speech at each SNR.

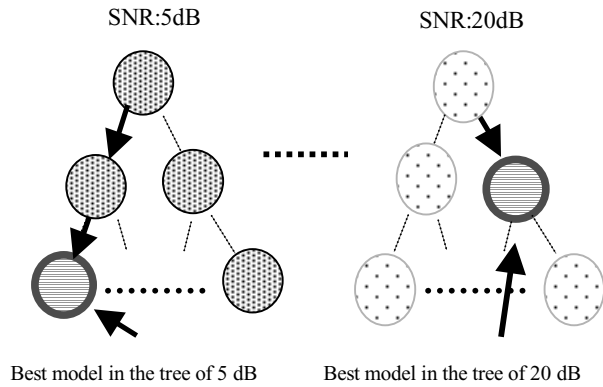
2.2 Tree-structured HMM by noise-added speech clustering

In the first method, noise-added speech HMMs are constructed using the tree made by noise clustering. Therefore, there is no guarantee that these HMMs are optimally clustered. In the second method, we directly cluster noise-added speech to construct the tree structured noisy speech HMM. Various noise-added speech data are made by adding noise signals to

clean speech at each SNR condition. In the same way as noise clustering, noise-added speech GMMs are made and clustered for each SNR. The noise-added speech data set corresponding to each cluster is used to construct the tree-structured noisy speech HMMs for recognition. While the first method uses the same tree structure regardless of the SNR, the second method uses different structures according to the SNR.

2.3 HMM selection

In the recognition phase, a test utterance is first decoded using the clean HMM to produce a phone label sequence. The likelihood values averaged over the test utterance length using various HMMs are calculated according to the phone labels. The noise-cluster HMM that best fits the input speech is selected using a two-step search method. In the first step, the model having the largest likelihood is selected for each SNR by tracking the tree from the root (top) to the leaves (bottom), as shown in Figure 1. Next, the best model among all SNR conditions is selected. This method yields the best HMM that matches the noise property as well as the SNR of input speech.



Best model in the tree of 5 dB Best model in the tree of 20 dB

Fig. 1: Tree-structured noise-clustered HMM for each SNR.

2.4 Linear transformation

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

$$\hat{\mu} = A\mu + b \quad (1)$$

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 [6] so 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. A user utters a kind of food, a station name, and conditions for narrowing down the retrieval candidates. The database of restaurants and food stores known to the Internet was

used. The database consists of 80 business categories and 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 are used. The models are trained using text corpora that are prepared separately for each dialogue content (topic) category. Some training texts are transcribed from real dialogue utterances, and other texts are manually typed in by several human subjects on the assumption that they are 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 ‘‘Spontaneous Speech Corpus’’[8] 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 [9]. Both noise GMM and noise-added speech GMM (64 mixtures) were trained for each noise using the Baum-Welch algorithm.

3.5 Evaluation data

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 \times 3 SNRs).

3.6 Effectiveness of model selection using the tree-structured noise-adapted HMM

Recognition experiments were performed to evaluate the first method; the tree structure was made by noise clustering instead of noisy-speech clustering. The best matching noise-adapted HMM was selected from the tree and used to recognize the input speech. In this experiment, MLLR was not applied.

Figure 2 shows the word accuracy when SNR of noisy input utterances was set at 15dB. The noise-cluster HMM that

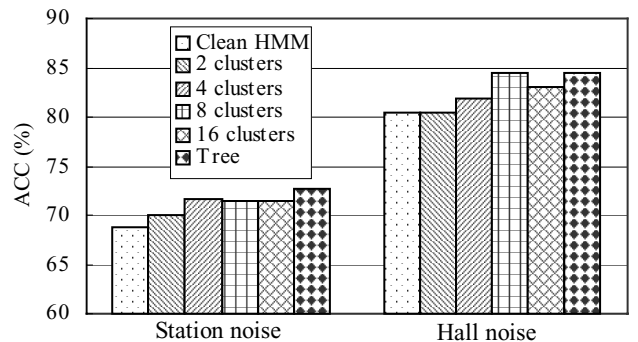


Fig. 2: Results by model selection using the tree-structured HMMs made by the noise clustering method. (SNR:15dB)

maximized the likelihood of the input speech was selected from those with the same SNR. The “Baseline” indicates the case wherein the clean HMM was used for recognition. The figure also shows the results when noises were directly clustered into different numbers of clusters without using the tree structure. These results indicate that the tree-structured method gives the best performance.

3.7 Comparison of noise clustering and noise-added speech clustering

An evaluation experiment was performed to compare the results of the noise clustering and the noise-added speech clustering methods. Figure 3 shows the word accuracy for the two proposed methods and the “Baseline” at the 15dB SNR condition. Noise clustering “Method 1” and noise added speech clustering “Method 2” yielded the same result for “Station” noise-added speech. Method 2 had better performance than Method 1 for “Hall” noise-added speech. These results indicate that the noise-added speech clustering method is slightly better than the noise clustering method.

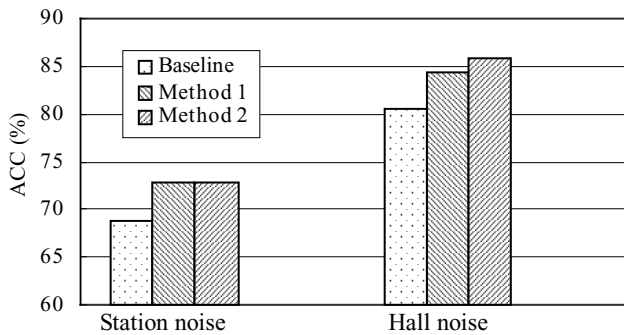


Fig. 3: Comparison of noise clustering and noise-added speech clustering methods. (SNR:15dB)

We analyzed the selected models in the tree structure. It was found that in most cases, the selected model came from the mid level of the tree as shown in Figure 4. Figure 5 shows the result of an analysis using Hayashi’s quantification theory (Type 4) for noisy speech. “Station” and “Hall” correspond to points “3” and “8” respectively. The model selected for “Station” noise-added speech was trained by the noises of “Station passage”, “Air-conditioner” and “Sorting factory”. The three noises are represented by “4”, “6” and “7”. It is clearly shown

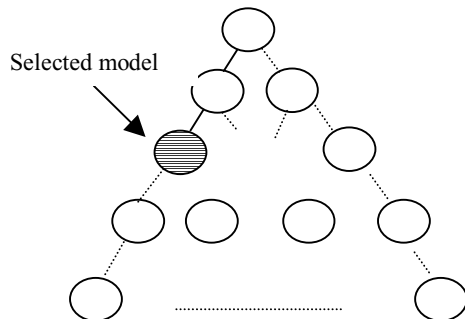


Fig. 4: An example of model selection in the tree-structured speech HMMs (“Station” noise-added speech).

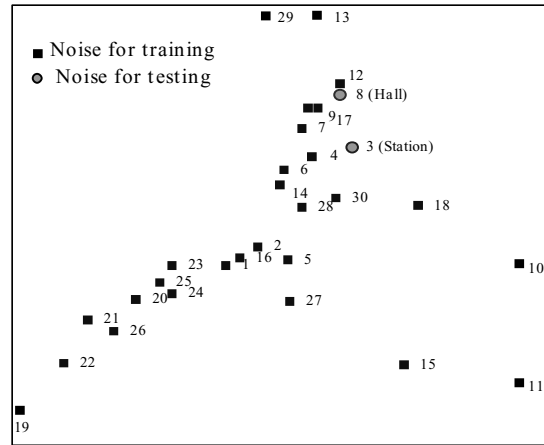


Fig.5: Noisy speech analysis by Hayashi’s quantification theory

that these noises are close to “Station”, so it is reasonable to use this combination to model the “Station” noise-added speech. A similar result was obtained for the “Hall” noise- added speech.

3.8 Effectiveness of piecewise-linear transformation

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 experiments. Specifically, the noise-cluster HMM that gave the maximum likelihood for each input speech was selected from all the noise-cluster HMMs at SNR values of 5, 10 and 15dB, and then the MLLR transformation was performed.

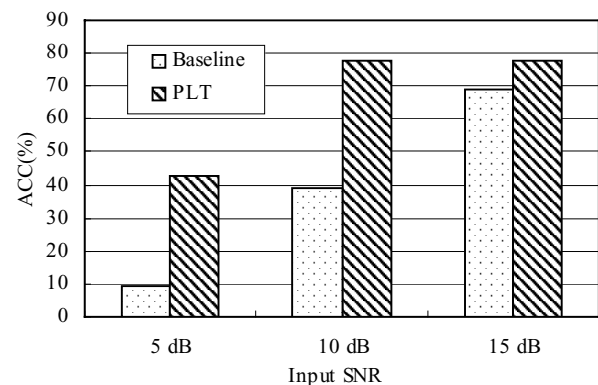


Fig. 6: Results by PLT (“Station” noise-added speech).

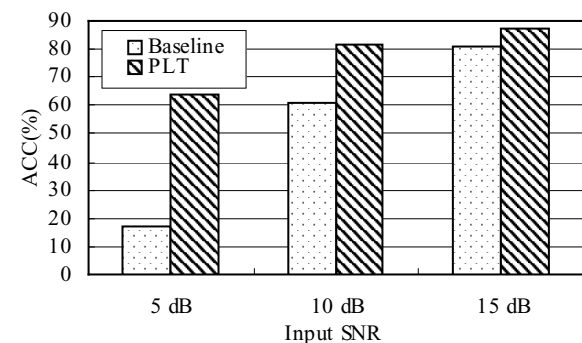


Fig. 7: Results by PLT (“Hall” noise-added speech).

Figures 6 and 7 show the recognition results. "PLT" indicates the use of the piecewise-linear transformation method. These results show that the "PLT" method improves the performance significantly. It reduces the word error rate by 36.1% on average relative to the "Baseline" results.

3.9 GMM-based model selection

As 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 huge amount of computation to calculate the likelihood values using all HMMs in parallel, GMMs were made using the same noise-added speech used to construct the HMMs and used for cluster 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 33.0%. Since the HMM-based method resulted in a 36.1% reduction, the GMM-based method 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 loss in performance.

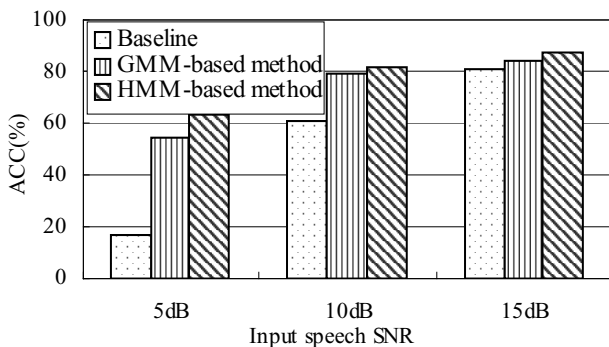


Fig. 8: Results by baseline, "GMM-based method" and "HMM-based method" ("Station" noise-added speech).

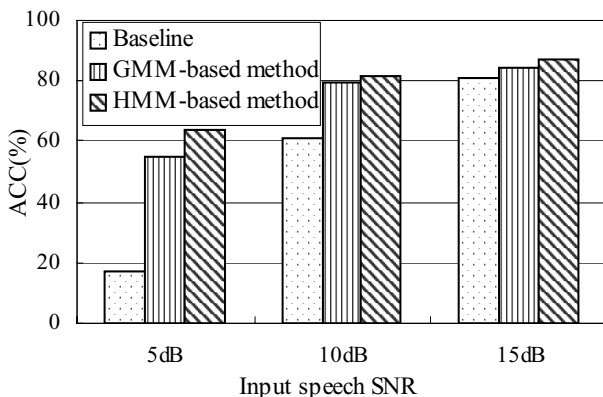


Fig. 9: Results of baseline, "GMM-based method" and "HMM-based method" ("Hall" noise-added speech).

4. Conclusion

This paper has reported a new method of HMM adaptation that uses piecewise-linear transformation (PLT); the goal is to improve large-vocabulary continuous-speech recognition accuracy for noise-added speech. The PLT method consists of two parts: best-matching HMM selection and linear transformation of the selected HMM based on the maximum likelihood criterion. In order to reduce the cost of model selection for input speech, this paper proposed two new methods. First, tree-structured noise-adapted HMMs are made by clustering noises or noisy speech, and model selection is performed by tracing the tree from the root to the leaves. Second, GMMs that correspond to the HMMs in the tree structure are made and used to select the best model instead of the HMMs. The HMM corresponding to the selected GMM is further adapted to match the input speech.

The proposed methods have been evaluated using a dialogue system, in which two kinds of real noise were added to speech at three different SNR levels (5, 10 and 15dB). The noises differed from those used for creating noise-adapted HMMs and GMMs. Experimental results show that the proposed method with HMM-based and GMM-based model selection achieved error rate reductions of 36.1% and 33.0%, respectively.

Future research includes increasing the variation of noises for both training and testing, improving the tree structure of noise-adapted HMM, and automatic noise/speech segmentation. Although this paper investigated only the influence of additive noise, actual speech usually involves the combination of various distortions including multiplicative distortions. Since the framework of the proposed method is flexible enough to cope with various distortions simultaneously, it is worth trying to apply our method to more complex conditions.

References

- [1] M. J. F. Gales et al.: "An improved approach to the hidden Markov model decomposition of speech and noise", Proc. ICASSP, pp. 233-236 (1992)
- [2] F. Martin et al.: "Recognition of noisy speech by composition of hidden Markov models", Proc. Eurospeech, pp. 1031-1034 (1993)
- [3] S. Furui et al.: "Noise adaptation of HMMs using neural networks", Proc. ITRW ASR, pp. 160-167 (2000)
- [4] Z.P. Zhang and S. Furui: "Piecewise linear transformation-based HMM adaptation for noisy speech", Proc. IEEE ASRU, (2001)
- [5] N. Sugimura et al.: "A method of word-multitemplate extraction based on minimization of distance", Proc. Speech Tech. Committee Meeting, ASJ, S82-64 (1982)
- [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. ICML, pp. 53-58 (2002).
- [8] 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).
- [9] http://www.milab.is.tsukuba.ac.jp/corpus/noise_db.html