論文 / 著書情報
Article / Book Information

Title	Tree-Structured Clustering Methods for Piecewise Linear-Transformation-Based Noise Adaptation
Authors	Zhipeng Zhang, Toshiaki Sugimura, Sadaoki Furui
出典 / Citation	IEICE Trans.Inf.&Syst., Vol. E88-D, No. 9, pp. 2168-2176
発行日 / Pub. date	2005, 9
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 2005 Institute of Electronics, Information and Communication Engineers.

IEICE TRANS. INF. & SYST., VOL.E88-D, NO.9 SEPTEMBER 2005

PAPER

Tree-Structured Clustering Methods for Piecewise Linear-Transformation-Based Noise Adaptation

Zhipeng ZHANG[†a)], *Nonmember*, Toshiaki SUGIMURA[†], *Member, and* Sadaoki FURUI[††], *Fellow*

SUMMARY This paper proposes the application of tree-structured clustering to the processing of noisy speech collected under various SNR conditions in the framework of piecewise-linear transformation (PLT)-based HMM adaptation for noisy speech. Three kinds of clustering methods are described: a one-step clustering method that integrates noise and SNR conditions and two two-step clustering methods that construct trees for each SNR condition. 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 methods are evaluated by applying them to a Japanese dialogue recognition system. The results confirm that the proposed methods are effective in recognizing digitally noise-added speech and actual noisy speech issued by a wide range of speakers under various noise conditions. The results also indicate that the one-step clustering method gives better performance than the two-step clustering methods.

key words: *robust speech recognition, noise adaptation, piecewise-linear transformation, tree-structured noise clustering, GMM*

1. Introduction

The performance of current speech recognition systems degrades significantly when applied to real-world systems. With the increase of real-world applications such as dialogue systems and transcription systems, the need for robust speech recognition systems is becoming critical. Noise-added speech $\hat{s}$ is modeled by:

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

where s, n and SNR represent the 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 handle variable noise through the use of the maximum likelihood (ML) estimation criterion. Experiments confirmed that this method greatly improves the recognition rate even when SNR and noise spectral characteristics are 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 an ML-based piecewise linear-transformation (PLT) HMM adaptation method for additive noise. This method offers an approximation of the non-linear effect of additive noise [2]. In developing our proposed method, a wide variety of noise data were collected and classified into noise clusters. Noise-added speech HMMs (noise-cluster HMMs) were constructed using noisy utterances created by adding noise signals classified into each cluster to clean speech at several SNR conditions. In the recognition phase, the noise-cluster HMM that best fitted the input speech was selected and further converted to reduce mismatches with the input speech by the MLLR method. In both processes, noise-cluster HMM selection and linear transformation use the likelihood maximization criterion. Figure 1 shows the flow diagram of the proposed method.

Our original method has a fundamental problem: since the optimum cluster number varies with the input noisy speech, it is necessary to prepare multiple sets of clusters with different numbers of clusters, and choose the optimum model from among them. Therefore, it suffers from huge computational costs. For the previous method [2], we carried out experiments using multiple sets of clusters made by doubling the cluster number such as $1, 2, 4, \ldots 2^i, \ldots N$, where N is the total number of noise-cluster HMMs. Since the amount of computation needed to choose the best matching model is proportional to each cluster number, we needed to evaluate $1 + 2 + 4 + \ldots + N = 2N - 1$ models in the above case.

To avoid this problem, we propose a new method that constructs a tree-structured noisy HMM for a PLT-based adaptation method. The tree-structured clustering method has been successfully applied for speaker adaptation [3]. In this paper, we apply the tree-structured clustering method for PLT-based noise adaptation. We construct a tree-structured HMM that represents both noise spectra and the SNR. The root node includes all noises and SNR conditions and each leaf node consists of only one noise at one SNR condition. A noise-added speech HMM is constructed for each node. An HMM in the leaf layer is selected if the input noise speech is close to one of the kinds of noise and the SNR condition used for training while a model in the upper layer should be selected if the input noise speech dif-

Manuscript received December 1, 2004.
Manuscript revised April 20, 2005.

[†]The authors are with Multimedia Laboratories, NTT DoCoMo, Inc., Yokosuka-shi, 239-8536 Japan.
[††]The author is with the Department of Computer Science, Tokyo Institute of Technology, Tokyo, 152-8552 Japan.
a) E-mail: zzp@mml.yrp.nttdocomo.co.jp

DOI: 10.1093/ietisy/e88-d.9.2168

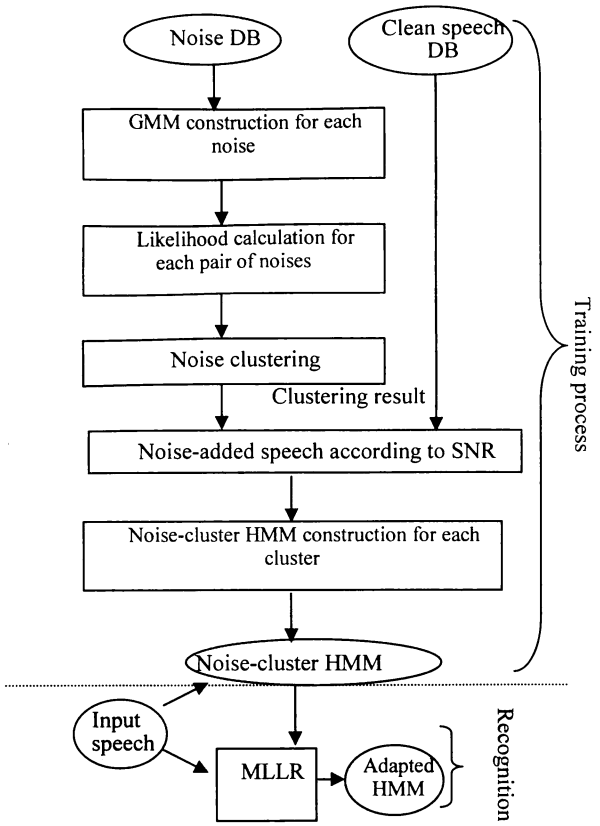


Fig. 1 System flow of piecewise-linear transformation for HMM noise adaptation.

fers from the noises and SNR condition used for training. The tree-structured hierarchical clustering method makes it easy to select the optimum model, and reduces the computational cost of doing so. By applying the tree-structured model space, the computational cost of finding the closest model is reduced to $2^* \log_2 N$ or less.

The remainder of the paper is organized as follows. Section 2 describes tree-structured clustering methods. Section 3 describes the speech recognition experiments conducted to evaluate the proposed methods. Sections 4 and 5 provide and analyze the results of the experiments. The first experiment uses noisy speech created by adding noise to clean speech. In the second experiment, actual samples from a wide range of speakers collected under various noise conditions are used. The paper concludes with a general discussion and issues related to future research.

2. Tree-Structured Clustering Methods for Piecewise Linear-Transformation-Based Noise Adaptation

Noise-added speech spectra vary as a function of both the noise spectra and SNR. It follows that any tree-structured HMM should represent both the noise spectra and SNR. We propose three tree-structured clustering methods for PLT-based noise adaptation: two two-step clustering methods that construct trees for each SNR condition and a one-step clustering method that integrates noise and SNR conditions

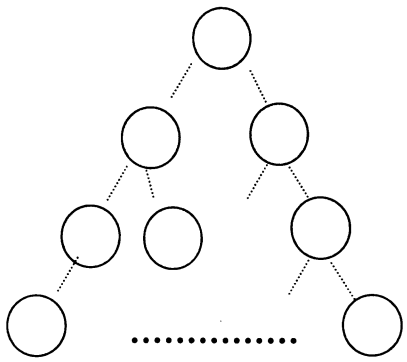


Fig. 2 Noise clustering method (Method 1); a common tree is used for all SNR conditions.

yielding a single tree covering all SNR conditions. The two-step clustering methods are the noise clustering method, which constructs the same tree for all SNR conditions, and the noise-added speech clustering method, which constructs a different tree for each SNR condition.

2.1 Two-Step Clustering Methods

2.1.1 Noise Clustering Method

The first method ("Method 1"), which uses noise clustering, starts by building a hierarchical structure of noise, and then uses it 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-added speech, models located in the lower layers represent specific features of noise-added speech. Figure 2 shows the concept of Method 1.

Since it is difficult to directly cluster noise data, we first build a GMM for each noise and then cluster the noise GMMs. The number of GMM components is 64. Noise GMM clustering is performed using a procedure originally proposed for the "SPLIT" speech recognition system [4]. This method clusters these noise GMMs in a top-down manner: select the cluster having the smallest intra-likelihood for clustering; choose two noise GMMs, the combination of which maximizes the sum of likelihood from among all noise GMMs; and then split all noise GMMs. This process is continued until all noise GMMs are individually clustered as leaves. This procedure has an advantage in that as the number of clusters increases, the sum of likelihood values increases. As the SPLIT method makes a binary tree that can be easily handled for finding the best node, we use it in noise clustering. The likelihood between each pair of noises is defined as the value of feature vectors of one noise yielded by GMM model of the other noise.

According to the noise clustering result, we construct noise-added speech HMMs from a set of noisy utterances created by adding noise signals for each cluster to clean speech at each SNR. As the noise clustering result is directly applied to all SNR conditions, these trees at different SNR conditions have the same structure.

In the recognition phase, a test utterance is first decoded using the clean HMM to produce a phoneme label sequence. In this paper, we assume correct sentence boundaries are given by some automatic method or the push-to-talk strategy. The likelihood values averaged over each sentence length using various HMMs are calculated according to the phoneme labels. The noise-cluster HMM that best fits the input sentence 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). Next, the best model among all SNR conditions is selected. This method yields the HMM that best matches the noise property as well as the SNR of the input speech. The best-matching HMM selection process is repeated for each sentence.

Although it is possible to reproduce phoneme labels having a higher accuracy by using the selected HMM, and reselect the best-matching HMM, this method is not used in order to reduce the amount of computation needed. Since the phoneme labels are only used for model selection, we consider the reproduction process using the selected HMM has no significant impact on the final recognition results.

2.1.2 Noise-Added Speech Clustering Method

In Method 1, noise-added speech HMMs are constructed using the tree made by noise clustering. Therefore, optimal clustering is not guaranteed for noise-added speech. In the second method ("Method 2"), 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. In this method, the noise-added speech GMMs for clustering at each SNR condition are different. Therefore, the clustering results at each SNR condition are different so that the tree structures at each SNR condition are not the same. This means that nodes of the tree contain different kinds of noise from that of "Method 1". The noise-added speech data set corresponding to each cluster is used to construct the tree-structured noisy speech HMMs.

The model selection process is performed in the same way as Method 1.

2.2 One-Step Clustering Method

For Methods 1 and 2, we cluster the noise samples or noise-added speech data to build a hierarchical structure of noise at each SNR condition. The tree structure is then used to create noise-added speech HMMs for different SNR conditions. In such a tree-structured space, two-step search is needed 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. Although this method is an easy way to handle the variations of both noise and SNR, it has a disadvantage in that it incurs large computation cost to find the best model.

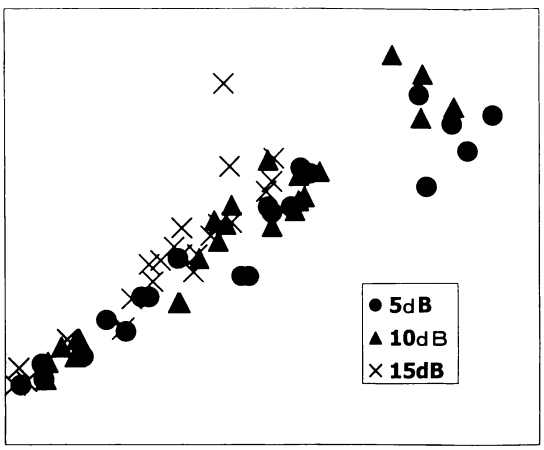


Fig. 3 A scatter diagram for noise-added speech at three SNR conditions.

Figure 3 shows a projection of noise-added speech data with three SNR values, 5, 10 and 15 dB, and 30 kinds of noise on a two-dimensional space made by Hayashi's quantification theory [5]. The x and y axes indicate the two eigenvectors having the largest eigenvalues, which best represent the differences between noise samples. The noise-added speech data at 5, 10, and 15 dB are indicated by circles, triangles, and crosses, 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, in this subsection, we propose a clustering method ("Method 3") that integrates noise as well as SNR variations. 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 at all SNR conditions to build a tree, and a noise-added speech HMM is made for each node in the tree. The HMM models in this tree are robust to the noisy speech in which SNR can vary within one sentence as they were constructed from data made at several SNR conditions. Figure 4 shows the concept of Method 3.

In the recognition phase, the best matching HMM is selected by tracking the tree from the root to the lower nodes for each test sentence utterance.

2.3 Linear Transformation

In all clustering and model selection methods, the selected HMM is converted by the MLLR adaptation method [6] to further adapt the selected model to the input sentence utterance. Transform sharing over Gaussian distributions can allow all distributions in a system to be updated with just a relatively small amount of adaptation data. A global trans-

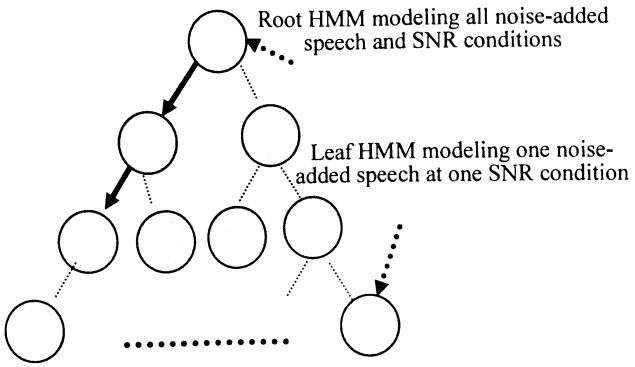


Fig. 4 One-step clustering method (Method 3).

formation is applied to every Gaussian component in the model.

3. Experiments

3.1 Task

The task of the system described below is retrieving information about restaurants and food stores. A user utters one kind of food, a station name, and conditions for narrowing down the retrieval candidates. 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-based word bigrams and reverse class-based word trigrams 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 typed in 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-based language models. Words belonging to each class were given an equal word occurrence probability.

3.3 Acoustic Models

The acoustic features were 25-dimensional vectors consisting of 12 MFCCs, 12Δ MFCCs, and Δ log energy. 42 phonemes were used in the acoustical model. 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)" [7] produced by male speakers (approximately 59 hours) were used for training the clean speech HMM. The same data were used to make noise-added speech.

3.4 Noise Data for Training

28 kinds of noises collected by JEIDA (Japan Electronic Industry Development Association) were used as training noises [8]. Noise-added speech were made using three SNR levels (SNR = 5 dB, 10 dB, 15 dB).

3.5 Evaluation Data

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

Test-1: 50 sentences uttered by 5 male speakers were used to evaluate the proposed method. The average duration of the test utterances was 1.54 seconds. 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 digitally added to the utterances at three SNR levels: 5, 10 and 15 dB. Experiments were therefore performed under 6 different conditions (2 noises $\times$ 3 SNRs).

Test-2: 540 sentence utterances from 12 speakers (45 per speaker) collected over three days (2003/01/20–22), were recorded in actual noisy environments ("Station" and "Office") and used in the experiments. The average duration of the test utterances were 4.86 seconds ("Station" noise-added speech) and 4.89 seconds ("Office" noise-added speech). The average SNRs were 10 dB ("Station" noise-added speech) and 12 dB ("Office" noise-added speech). The noise power was estimated using the noise periods immediately before and after each sentence utterance. The power of noise-added speech was estimated as the mean value averaged over the utterance period. The SNR was estimated based on these values. This task was relatively difficult, since the noise was non-stationary.

4. Experimental Results for Test-1

4.1 Effectiveness of Model Selection Using the Tree-Structured Noise-Adapted HMM Made by Method 1

Recognition experiments were performed to evaluate Method 1. The best matching noise-adapted HMM was selected from the tree and used to recognize the input speech, sentence by sentence. In these experiments, MLLR was not applied.

Figures 5, 6 and 7 show the word error rates (WER %) when the SNR of noisy input utterances was set at 5, 10 and 15 dB, respectively. The noise-cluster HMM that 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. To compare the effectiveness of the tree structure with that of the previous method, we conducted experiments using fixed numbers of noise clusters: 2, 4, 8, and 16, throughout the 50 sentences. This means that a model is selected only from the

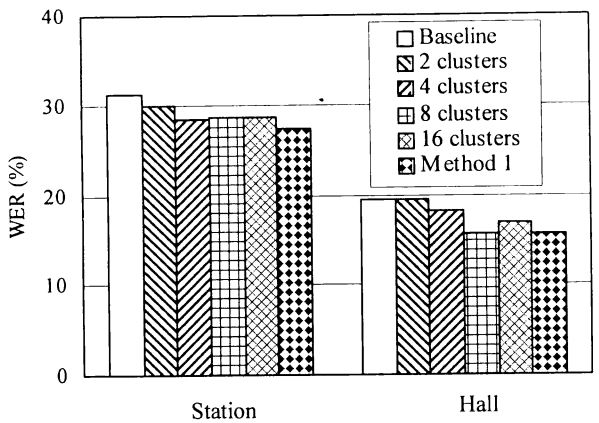


Fig. 5 Comparison of WER for baseline, model selection by "Method 1" using tree-structured clusters, and "Method 1" using fixed number of clusters on Test-1 data (SNR: 15 dB).

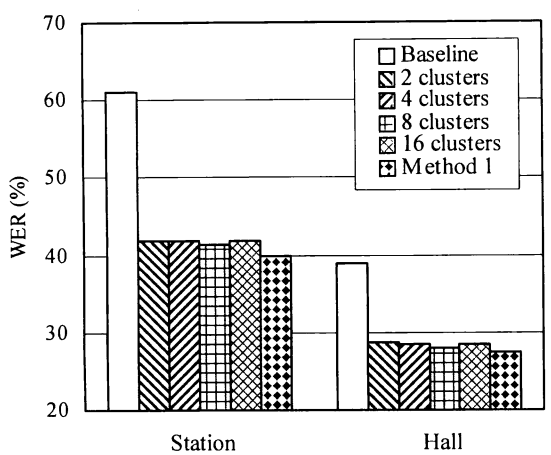


Fig. 6 Comparison of WER for baseline, model selection by "Method 1" using tree-structured clusters, and "Method 1" using fixed number of clusters on Test-1 data (SNR: 10 dB).

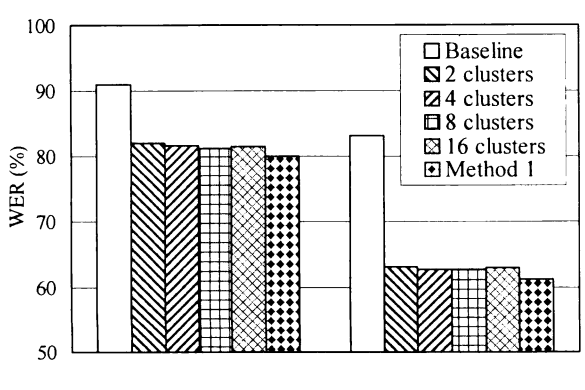


Fig. 7 Comparison of WER for baseline, model selection by "Method 1" using tree-structured clusters, and "Method 1" using fixed number of clusters (SNR: 5 dB).

given number (2, 4, 8 or 16) of models for each input sentence. These results indicate that the tree-structured method gives the best performance at all SNR conditions.

Since the noises used in the experiments are non-stationary, the optimum number of clusters varies for each sentence utterance. Therefore, the optimum models for the 50 sentence utterances cannot be selected from the limited number of clusters. On the other hand, the tree-structured method can select the best-matching model from the tree for each sentence utterance. If we select a model from all the (flat structure) clusters without using the tree, the same performance as the tree-structured method can be obtained, but the amount of computation becomes excessive. In this experiment, since 28 noise conditions were used for clustering, the computational cost of searching all the clusters by using the tree structure is approximately 1/6 ($[2 \times \log_2(28)]/[(2 \times 28 - 1)]$) of that using the flat structure. Furthermore, the process of tracing the tree from the root to the leaves for model selection stops when the optimum model is selected. Therefore, the actual computational cost with the tree structure is often much less than $2 \times \log_2(28)$, and the relative computational cost becomes less than 1/6.

4.2 Model Selection by Methods 1, 2 and 3 and Comparison with the Method Using Fixed Number of Clusters

Experiments were performed to evaluate the effectiveness of the tree-structured clustering by Methods 1, 2 and 3. We conducted the experiments to compare these tree-structured methods ("tree structure") with the previous method ("previous method").

Figures 8 and 9 show the word error rates for the three methods (Methods 1, 2 and 3) and the "Baseline" at three SNR conditions (SNR = 5, 10, 15 dB). In these figures, "previous method" indicates the best result among the 2, 4, 8, and 16 clusters. These results show that the tree-structured methods consistently give better performance than the previous method at all SNR conditions. It is also indicated that Method 2 has better performance than Method 1 in most cases. The reason is that Method 2 consistently uses noise-added speech for clustering and constructing the HMMs. These results also showed that using Method 3 yielded the lowest processing time for selecting the best model compared to Method 1 and Method 2.

Since these experiments used 28 noises at 3 SNR conditions for clustering, the total number of leaf nodes was 3×28 and the reduction in computational cost achieved by Method 3 was approximately 1/13 ($[2 \times \log_2(3 \times 28)]/[(2 \times 3 \times 28) - 1]$).

4.3 Comparison of MLLR, Model Selection and Piecewise-Linear Transformation

Experiments were performed to compare the performances of MLLR, model selection and piecewise-linear transformation ("PLT", the combination of model selection and MLLR). In "Model selection", the noise-cluster HMM that

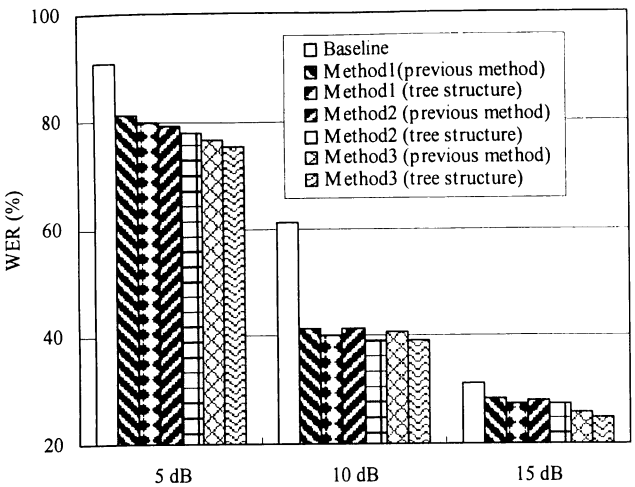


Fig. 8 Comparison of baseline, proposed method (tree-structured-clusters) and previous method (fixed number of clusters) for the three clustering methods on Test-1 data (model selection only, "Station" noise-added speech).

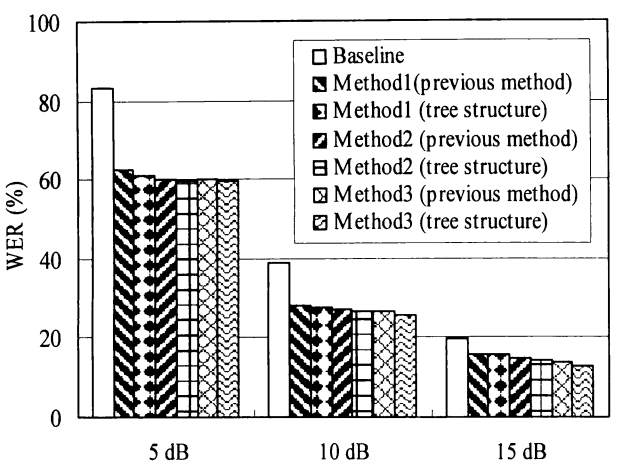


Fig. 9 Comparison of baseline, proposed method (tree-structured-clusters) and previous method (fixed number of clusters) for the three clustering methods on Test-1 data (model selection only, "Hall" noise-added speech).

gave the maximum likelihood for each input speech was selected from all the noise-cluster HMMs. In the PLT, for each input utterance, the best matching noise-adapted HMM was selected from the tree (created by Method 3) after which MLLR transformation was performed. The selected model was then used to recognize the input utterance after MLLR adaptation. In the "MLLR", MLLR adaptation was performed directly for the clean HMM.

Figures 10 and 11 show the word error rate for two kinds of noise-added speech at three SNR conditions (SNR = 5, 10, 15 dB). "Baseline" indicates the case wherein the clean HMM was used for recognition. These results show that the model selection and MLLR methods have almost equal effectiveness, and that their combination further improves the performance. Relative to the "Baseline" results, the PLT (Method 3) reduced the word error rate by

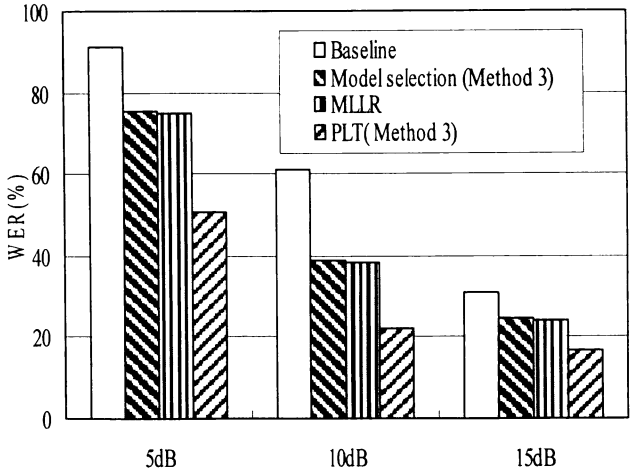


Fig. 10 Comparison of model selection, MLLR and PLT using Method 3 on Test-1 data ("Station" noise-added speech).

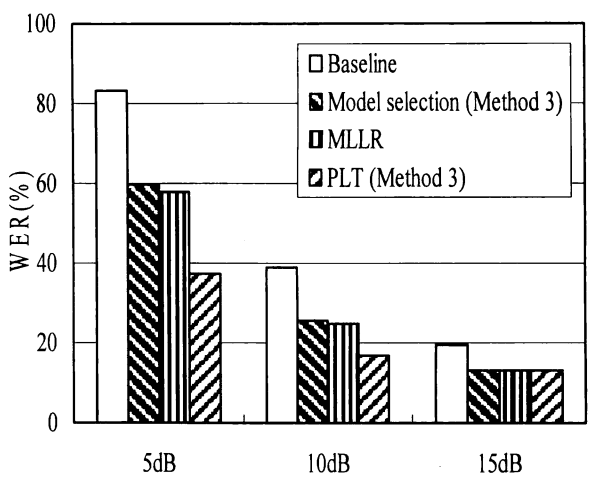


Fig. 11 Comparison of model selection, MLLR and PLT using Method 3 on Test-1 data ("Hall" noise-added speech).

50% at 5 dB, 59% at 10 dB, and 33% at 15 dB on average.

4.4 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 number of mixtures in GMM was set to 64. 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 selected model was then used to recognize the input utterance after MLLR adaptation ("PLT"). In these experiments, Method 3 was used for clustering.

Figures 12 and 13 compare the tree-structured method "tree structure" to the previous method "previous method"

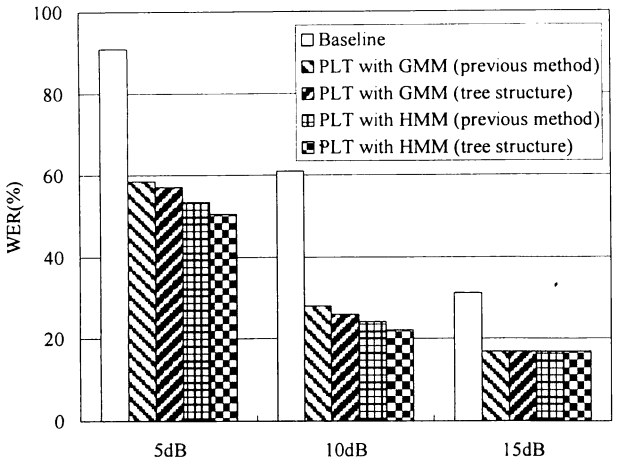


Fig. 12 Results by baseline, previous method and proposed method for GMM-based and HMM-based PLT methods on Test-1 data ("Station" noise-added speech).

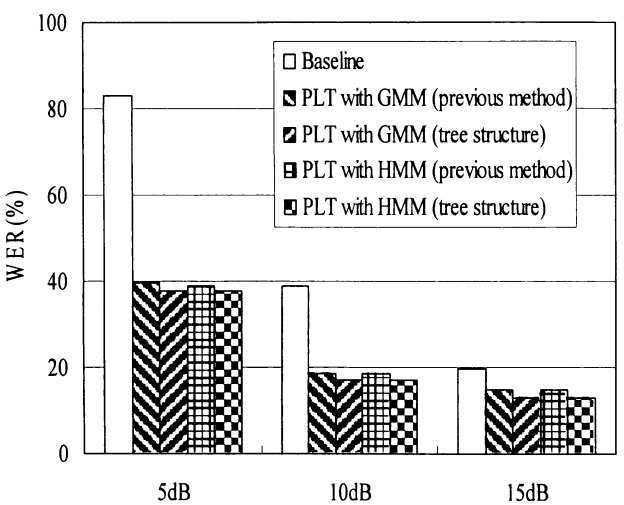


Fig. 13 Results by baseline, previous method and proposed method for GMM-based and HMM-based PLT methods on Test-1 data ("Hall" noise-added speech).

for the HMM-based PLT "PLT with HMM", and the GMM-based PLT "PLT with GMM". Figures 12 and 13 also show the result of "Baseline". These results show that the tree-structured method gives better performance than the previous method at all SNR conditions examined. These results also show that "PLT with GMM" reduced the word error rate by 47% at 5 dB, 57% at 10 dB, and 33% at 15 dB. The computational costs of the GMM-based method are approximately 1/1000 of that using the HMM-based method. The GMM-based method has slightly worse performance than the HMM-based 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.

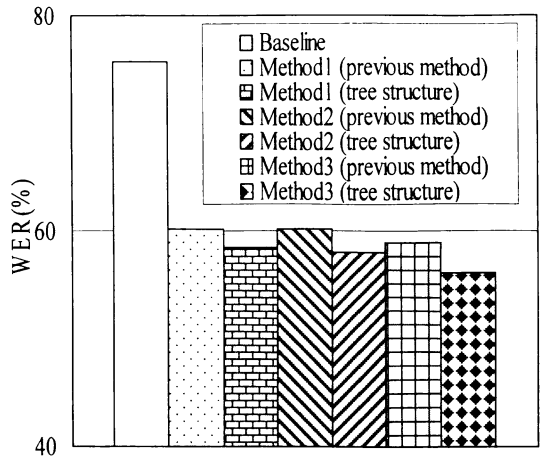


Fig. 14 Comparison result of previous method and tree-structured methods for three clustering methods on Test-2 data ("Station" noise-added speech).

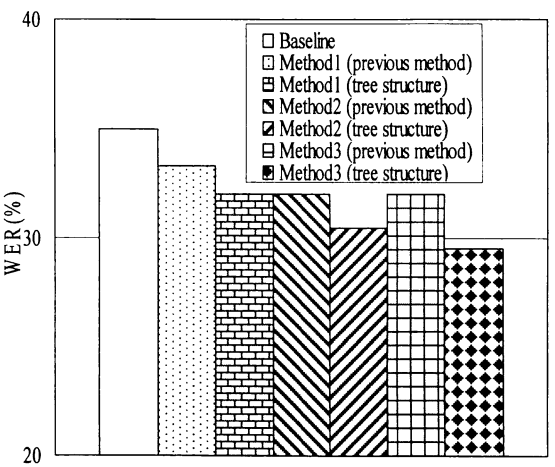


Fig. 15 Comparison result of previous method and tree-structured methods for three clustering methods on Test-2 data ("Office" noise-added speech).

5. Experimental Results for Test-2

5.1 Comparison of the Three Clustering Methods with the Previous Method

Experiments on Test-2 data were performed to compare the three clustering methods ("tree structure") and the previous method ("previous method") with a fixed number of noise clusters. Figures 14 and 15 show the results for "Method 1", "Method 2", and "Method 3". The "previous method" is the best result from among 2, 4, 8, 16 clusters. In these experiments, selection was performed by HMM, not GMM. These results indicate that the tree-structured methods give better performance than the previous method. This also confirms that Method 3 has the best performance among the three clustering methods. These results show that the proposed method is effective in actual noisy environments.

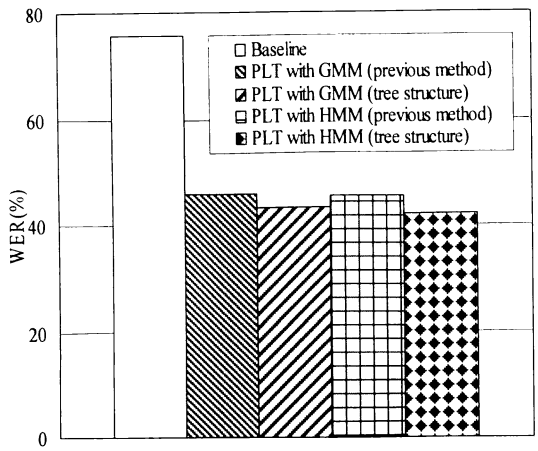


Fig. 16 Results of baseline, previous method and tree-structured method for GMM-based and HMM-based PLT methods on Test-2 data ("Station" noise-added speech).

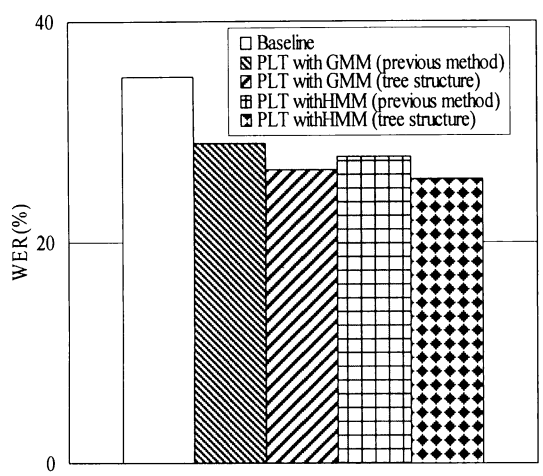


Fig. 17 Results of baseline, previous method and tree-structured method for GMM-based and HMM-based PLT methods on Test-2 data ("Office" noise-added speech).

5.2 GMM/HMM-Based PLT

Another experiment was performed to evaluate GMM/HMM-based PLT methods including Method 3 using Test-2 data. Figures 16 and 17 show the results for five conditions: no adaptation "Baseline", tree-structured method ("tree structure") and previous method ("previous method") for the HMM-based PLT "PLT with HMM", and the GMM-based PLT "PLT with GMM". These results show that the tree- structured methods give better performance than the previous method. These results also show that the "PLT with HMM" reduced the word error rate by 35.0% while the "PLT with GMM" reduced the word error rate by 33.3%. This confirms the effectiveness of the GMM-based method. These results show that the proposed method is effective in actual noisy environments.

6. Conclusions

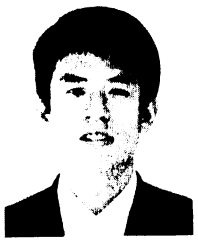
This paper proposed three tree-structured clustering methods for piecewise linear-transformation-based (PLT) adaptation: the noise clustering method (Method 1), noise-added speech clustering method (Method 2), and one-step clustering method (Method 3). Model selection is performed by tracing the tree from the root to the leaves and the selected model is further transformed by MLLR. Both model selection and linear transformation are based on the ML criterion.

The proposed methods were evaluated using a dialogue system and two kinds of test data. Experimental results show that the proposed tree-structured methods give better performance than the previous method for various test data with a large reduction in computational cost. In this experiment, which used 28 noises and 3 SNR conditions, the reduction in computational cost was approximately 1/13. Experimental results also show that Method 3, which deals with noise and SNR variations simultaneously, gives the best performance among the three clustering methods. In combination with MLLR, it achieved error rate reductions of 49.8% and 35.0% on Test-1 (digitally noise-added speech) and Test-2 (actual noisy speech) data, respectively.

Future research includes increasing the extent of noise variation in both training and testing.

References

[1] Y. Minami and S. Furui, "A maximum likelihood procedure for a universal adaptation method based on HMM composition," Proc. IEEE International Conference on Acoustics Speech and Signal Processing, pp.129–132, 1995.
[2] Z.P. Zhang and S. Furui, "Piecewise-linear transformation-based HMM adaptation for noisy speech," Speech Commun., vol.42, no.1, pp.43–58, 2004
[3] T. Kosaka, S. Matsunaga, and S. Sagayama, "Speaker-independent speech recognition based on tree-structured speaker clustering," Comput. Speech Lang., vol.10, pp.55–74, 1996.
[4] N. Sugamura, K. Aikawa, K. Shikano, and M. Kohda, "Speaker independent recognition of isolated words based on multiple reference templates in SPLIT system," Trans. Committee on Speech Research, Acoustical Society of Japan, pp.505–512, 1982.
[5] T. Kan, Multivariate Analysis, Gendai-Sugakusha, 1993.
[6] M.J.F. Gales and P.C. Woodland, "Mean and variance adaptation within the MLLR framework," Comput. Speech Lang., vol.10, no.3, pp.249–264, 1996.
[7] S. Furui, K. Maekawa, H. Isahara, T. Shinozaki, and T. Ohdaira, "Toward the realization of spontaneous speech recognition—Introduction of a Japanese priority program and preliminary results," Proc. International Conference on Spoken Language Processing, pp.518–521, 2000.
[8] http://www.milab.is.tsukuba.ac.jp/corpus/noise_db.html



Zhipeng Zhang received the Ph.D. degree from Tokyo Institute of Technology in Computer Science at 2002. He is currently a research engineer of Multimedia Laboratories, NTT DoCoMo, Inc. His research activities are oriented toward speaker model adaptation and noise model adaptation. He is a member of the Acoustical Society of Japan (ASJ), the International Speech Communication Association (ISCA).

Toshiaki Sugimura received the MS degree from Tokyo Institute of Technology in Computer Science at 1980. He joined NTT in 1980. He is engaged in the research of natural language processing, intelligent processors, mobile multimedia, ubiquitous information-communication environments, and other subjects. He is a member of Association for Computing Machinery (ACM) and the Information Processing Society of Japan (IPSJ).

Sadaoki Furui is currently a Professor at Tokyo Institute of Technology, Department of Computer Science. He is engaged in a wide range of research on speech analysis, speech recognition, speaker recognition, speech synthesis, and multimodal human-computer interaction and has authored or coauthored over 400 published articles. He is a Fellow of the IEEE and the Acoustical Society of America. He is President of the Acoustical Society of Japan (ASJ), the International Speech Communication Association (ISCA), and the Permanent Council for International Conferences on Spoken Language Processing (PC-ICSLP). He is a Board of Governor of the IEEE Signal Processing Society. He is Editor-in-Chief of the Transaction of the IEICE and has served as an Editor-in-Chief of Speech Communication. He has received the Yonezawa Prize, the Paper Award and the Achievement Award from the IEICE (1975, 1988, 1993, 2003, 2003), and the Sato Paper Award from the ASJ (1985, 1987). He has received the Senior Award from the IEEE ASSP Society (1989) and the Achievement Award from the Minister of Science and Technology, Japan (1989). He has received the Book Award from the IEICE (1990). In 1993 he served as an IEEE SPS·Distinguished Lecturer.