PIECEWISE-LINEAR TRANSFORMATION-BASED HMM ADAPTATION FOR NOISY SPEECH

Zhipeng Zhang and Sadaoki Furui

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

ABSTRACT

This paper proposes a new method using piecewise-linear transformation for adapting phone HMMs to noisy speech. Various noises are clustered according to their acoustical property and signal-to-noise ratios (SNRs), and noisy speech HMM corresponding to each clustered noise is made. Based on the likelihood maximization criterion, the HMM which best matches an input speech is selected and further adapted using linear transformation. The proposed method was evaluated by recognizing noisy broadcast-news speech. It was confirmed that the proposed method was effective in recognizing numerically noise-added speech and actual noisy speech 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. HMMs with Gaussian mixtures are usually used to model speech represented by cepstral coefficients, meaning that speech is modeled in the logarithmic spectral domain. However, noise is usually added to speech in the waveform or in the linear spectral domain, so the incorporation of additive noise into HMMs is not straightforward. Researches in this field have been actively carried out. Parallel model combination (PMC, also called HMM composition) [1][2] is one of the most practically 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). We previously proposed a method using a neural networks-based mapping [3] to deal with the problem and this method was confirmed to be effective in recognition under new speakers, new noise and various SNR conditions. However, these methods have a disadvantage that they require a large amount of computation including non-linear conversion. This problem is especially important when noise is time varying and the noise effect needs to be compensated for each utterance.

In order to solve such a problem, this paper proposes a piecewise-linear transformation (PLT) as an approximation of the non-linear effect of additive noise. The piecewise-linear transformation is performed in two steps: the noise-additive HMM parameter space selection and linear transformation for the selected HMM, both processes being performed based on the likelihood maximization criterion.

In this paper, we first explain the principal method, and then report on two experiments. The first experiment is carried out for numerically noise-added speech. In the next experiment, actual utterances by a wide range of speakers under various noise conditions are used. The paper concludes with a general discussion and issues related to future research.

2. PRINCIPLES OF NOISE-ADAPTATION USING PIECEWISE-LINEAR TRANSFORMATION

Noise-added speech spectra vary as a function of both noise spectra and signal-to-noise ratio (SNR). Therefore we first cluster noises, and then construct noise-added speech HMMs (noise-cluster HMMs) using a set of noisy utterances created by adding each cluster noise to clean speech at several SNRs. In the recognition phase, a noise-cluster HMM which best fits the input speech is selected, and further converted by maximum likelihood linear transformation. Figure 1 shows a system flow diagram of the method.

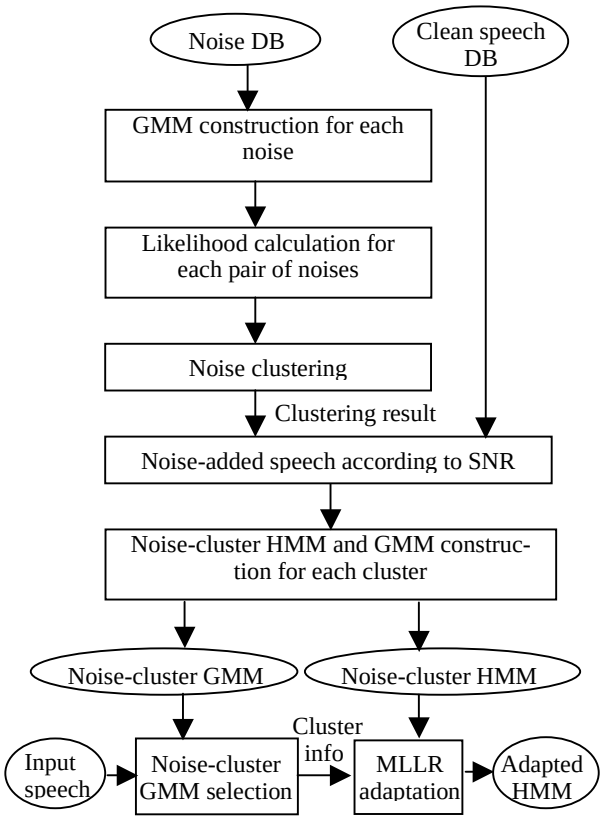


Fig. 1: Piecewise-linear transformation for HMM noise adaptation.

2.1 Noise clustering and HMM selection

Since it is difficult to directly cluster noise data, we first build GMM for each noise and cluster noise GMMs rather than directly clustering noise data. Likelihood for each pair of noise-GMM is calculated for the clustering. Based on the likelihood matrix, a clustering procedure originally proposed for the "SPLIT" speech recognition system [4] is carried out. This procedure has the advantage that any number of clusters can be made. As the number of clusters increases, the sum of likelihood values increases. The clustering is terminated automatically when the sum of likelihood or the number of clusters exceeds a pre-set threshold.

Noise-cluster HMMs are constructed using the noise-added speech made by adding noises that belong to a cluster to the clean speech. Clean HMM is used as an initial model.

2.2 Linear transformation

Gaussian mean parameters of the noise-cluster HMM are adapted according to the following equation:

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

where A is an $n \times n$ transformation matrix and b is an n-dimensional vector. These parameters are estimated using the MLLR method [5] such that the likelihood of the input speech is maximized. Since there is no closed form solution to the matrix estimation problem, it is solved by using the Expectation-Maximization (EM) algorithm. The transform sharing over Gaussian distributions can allow all the distributions in a system to be updated with only a relatively small amount of adaptation data.

3. EXPERIMENTS ON A JAPANESE BROADCAST NEWS TRANSCRIPTION SYSTEM

3.1 Language models

The broadcast-news manuscripts used for constructing the language models were taken from NHK news broadcasts over a period starting from July 1992 to May 1996 and comprised roughly 500k sentences and 22M words (morphemes). To calculate word n-gram language models, we segmented the broadcast-news manuscripts into words (morphemes) using a morphological analyzer since Japanese sentences are written without spaces between words. Since many Japanese words have multiple readings and correct readings can only be decided according to context, we have constructed a language model in which a word with multiple readings is split into different language model entries according to its reading [6]. We also introduced filled-pause modeling into the language model. A word-frequency list was derived from the news manuscripts and the 20k most frequently used words were selected as the vocabulary. These 20k words covered approximately 98% of the words in the broadcast-news manuscripts. We calculated bigrams and trigrams and estimated unseen n-grams using the Katz's back-off smoothing method.

3.2 Acoustic models

The feature vector extracted from speech consisted of 16 cepstral coefficients, the normalized logarithmic power, and their delta features (derivatives). The total number of parameters in each vector was 34. The cepstral coefficients were normalized by the cepstral mean subtraction (CMS) method.

The baseline acoustic model (Clean HMM) was gender-dependent shared-state triphone HMM and was designed using the tree-based clustering. It was trained by using phonetically-balanced sentences and dialogues read by 53 male speakers. The contents were completely different from the broadcast-news task. The total number of training utterances was 13,270, and the total length of the training data was approximately 20 hours. The total number of HMM states was approximately 2,000 with four Gaussian mixture components per state.

3.3 Noise data for training

28 kinds of noises collected by JEIDA (Japan Electronic Industry Development Association) were used for noise clustering. A noise GMM with 64 mixtures was trained for each noise by the Baum-Welch algorithm.

3.4 Evaluation data

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

- **Test-1**: 10 broadcast-news sentence utterances by a male speaker were extracted and two noises, exhibition hall and crowd noises, different from the 28 kinds of noises used for noise clustering, were numerically added to the utterances with three SNRs: 0, 10 and 15dB. Experiments were therefore performed under 6 different conditions (2 noises x 3 SNRs).
- **Test-2**: 50 sentence utterances from a wide range of speakers, superimposed with noise and/or music, were extracted from the real broadcast-news speech and used for the experiments. Average SNR was 17dB. This task was relatively difficult, since the noise was fairly unstationary.

3.5 Effectiveness of noise clustering and HMM selection

Recognition experiments were performed for evaluating the method of selecting a noise-cluster HMM for each input utterance.

Figures 2 and 3 show the word accuracy for Test-1 as a function of the number of noise clusters for the three SNR conditions. In these experiments, the SNRs of noisy input utterances were given. Therefore a noise-cluster HMM which maximized the likelihood for input speech was selected from those with the same SNR. The “Clean HMM” indicates the case using the clean HMM for recognition. These results indicate that 16-cluster condition gives the best performance.

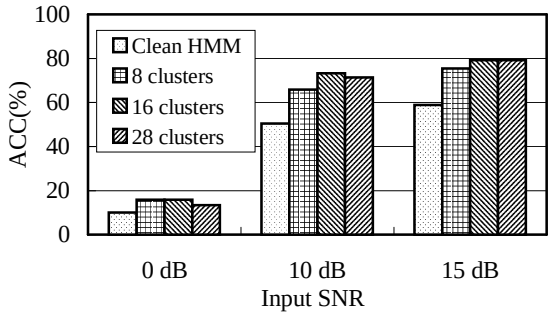
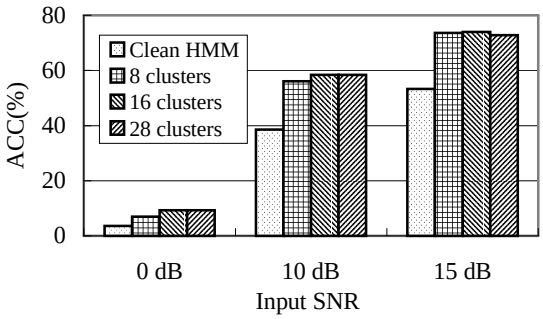


Fig. 2: Recognition results by noise-cluster HMM selection for Test-1 (crowd noise-added speech).



. **Fig. 3**: Recognition results by noise-cluster HMM selection for Test-1 (exhibition hall noise-added speech).

Figure 4 shows results for Test-2. Since the SNR of each input utterance is unknown, a noise-cluster HMM which gave the maximum likelihood for each input speech was selected among all noise-cluster HMMs with 0, 10, 15 or 20dB SNR. These results show that 8-16 cluster conditions give the best performance in general.

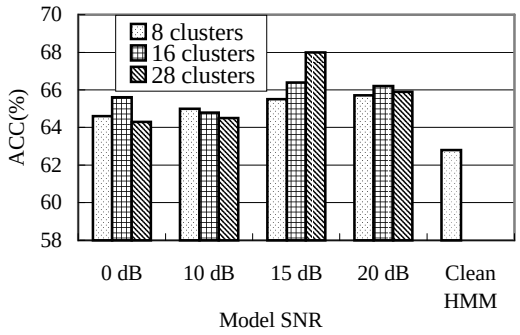


Fig. 4: Recognition result by noise-cluster HMM selection for Test-2.

3.6 Effectiveness of piecewise-linear transformation

Evaluation experiments for the piecewise-linear transformation, that is the combination of noise-cluster HMM selection and MLLR-based linear transformation, were carried out. Specifically, a noise-cluster HMM which gave the maximum likelihood for each input speech was selected among all noise-cluster HMMs with 0, 10, 15 and 20dB SNR, and then the MLLR transformation was performed. Since it needs huge amount of computation if we calculate the likelihood values using all noise-cluster HMMs in parallel for input speech, noise-cluster GMMs were made using the same noise-added speech used to construct the noise-cluster HMMs and used for the best cluster selection, instead of using the noise-cluster HMMs. A noise-cluster HMM corresponding to the selected noise-cluster GMM that yields the largest likelihood for input speech was used as the best model. The MLLR was performed for the selected noise-cluster HMM.

Figures 5 and 6 show the results for the Test-1. The "x-cluster" indicates the conditions without linear transfomation. The "cluster+MLLR" indicates the condition using the piecewise-linear transformation method. These results show that the "cluster+MLLR" condition gives the best performance, although the improvement by combining the MLLR is marginal.

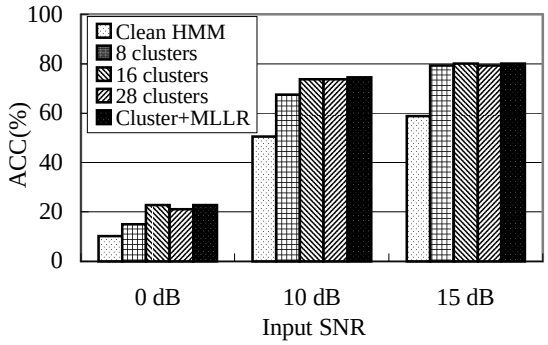


Fig. 5: Recognition results using piecewise-linear transformation for Test-1 (crowd-noise-added speech).

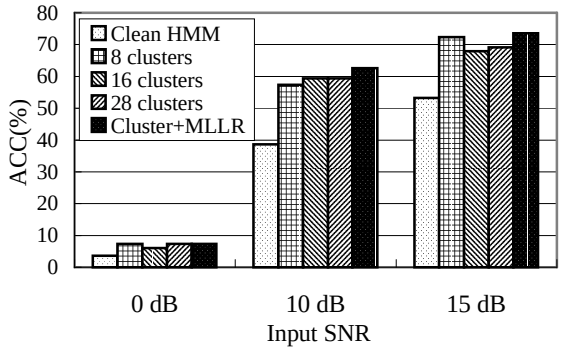


Fig. 6: Recognition results using piecewise-linear transformation for Test-1 (exhibition hall noise-added speech).

Another experiment for Test-2 was performed using the piecewise-linear transformation method. Figure 7 shows the word accuracy as a function of the number of clusters. It is clearly shown that the proposed method of combining cluster selection and MLLR gives significantly better results than cluster-selection in the case of real noisy speech.

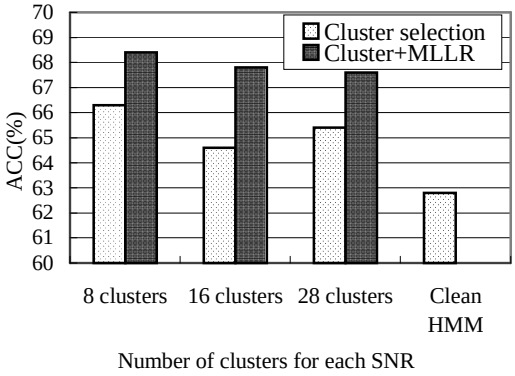


Fig. 7: Recognition results using piecewise-linear transformation for Test-2.

3.7 Comparison with PMC

A supplementary experiment using Test-1 was performed to compare the results of the proposed method (PLT) with that of the PMC method. The CMS (cepstral mean subtraction) was not applied to the input signals in the experiments using PMC, since the CMS is in principle difficult to combine with PMC.

Experimental results for the two different noises are shown in Figs. 8 and 9. These results show that the proposed method performs significantly better than PMC. This is partly because baseline performance prior to adaptation for the PMC-based method was significantly worse than the former method due to the fact that CMS was not combined with PMC. On top of that, it is clearly indicated that the improvement by noise adaptation using PLT is much larger than the one obtained by PMC.

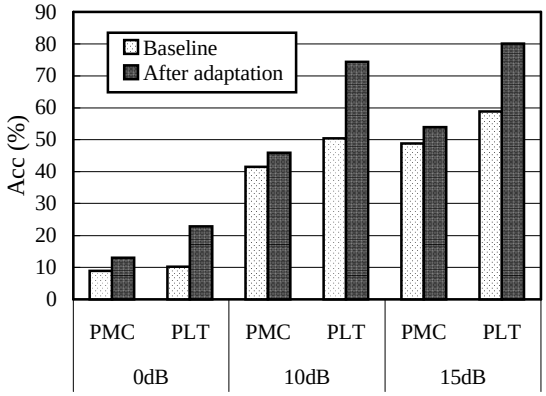


Fig. 8: Comparison of results by the proposed method (PLT) and PMC (crowd noise-added speech).

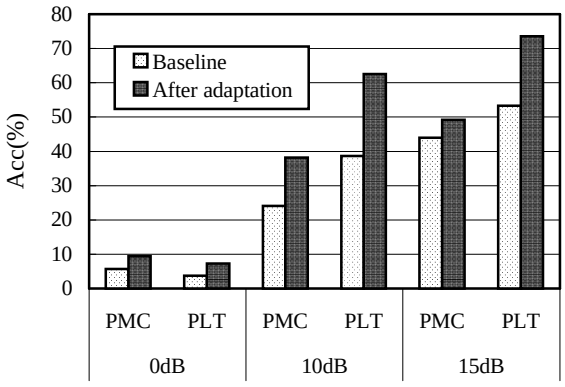


Fig. 9: Comparison of results by the proposed method (PLT) and PMC (exhibition hall noise-added speech).

4. CONCLUSION

This paper has reported investigations of HMM adaptation using a piecewise-linear transformation method, with the intent of improving large-vocabulary continuous-speech recognition accuracy for noise-added speech. The piecewise-linear transformation method consists of two parts: noise-added HMM parameter space clustering and linear transformation with the maximum likelihood criterion. In order to reduce the computational cost for selecting the noise-cluster HMM, a method that uses noise-cluster GMM to select the best model for input speech has been introduced.

Two experiments for numerically noise-added speech and real broadcast news speech distorted by various noises showed the effectiveness of the proposed method. The first experiment on numerically noise-added speech showed the effectiveness of the noise-cluster HMM selection method and marginal improvement by applying the MLLR for the selected HMM.

The second experiment using real noisy broadcast news speech, including reports from remote sites, showed that the combination of the MLLR significantly contributed to improve the recognition accuracy. This confirmed the effectiveness of the proposed method for real speech distorted by various types of noise.

Another experiment was carried out to compare the proposed method with PMC. Results show that the proposed method performed significantly better than PMC. Our future works include comparison of the proposed method with an improved PMC method [7] which maximizes likelihood by adapting HMMs taking additive noise as well as convolutional (multiplicative) distortion into consideration

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 will be worth trying to apply our method to more complex conditions.

ACKNOWLEDGMENTS

The authors would like to thank NHK (Japan Broadcasting Corporation) for providing the broadcast news database. This work is supported in part by the International Communications Foundations.

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 of the ISCA ITRW ASR2000 pp160-167 (2000)

[4] N. Sugimura et al.: “A method of word-multitemplate extraction based on minimization of distance”, Proc. Speech Tech. Committee Meeting, ASJ, S82-64 (1982)

[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] K. Ohtsuki et al.: “Improvements in Japanese Broadcast News Transcription”, Proc. DARPA Broadcast News Workshop, pp. 231-236 (1999).

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