論文 / 著書情報
Article / Book Information

Title	Evaluation of tree-structured piecewise-linear-transformation-based noise adaptation on AURORA2 database
Authors	Zhipeng Zhang, Tomoyuki Ohya, Sadaoki Furui
Citation	Interspeech 2004 - ICSLP, Vol. , No. 1, pp. 113-116,
Pub. date	2004, 10
Copyright	(c) 2004 International Speech Communication Association, ISCA
DOI	http://dx.doi.org/

EVALUATION OF TREE-STRUCTURED PIECEWISE-LINEAR-TRANSFORMATION-BASED NOISE ADAPTATION ON AURORA2 DATABASE

Zhipeng Zhang[1],Tomoyuki Ohya [1] and Sadaoki Furui[2]

[1] Multimedia Laboratories, NTT DoCoMo
3-5 Hikari-no-oka, Yokosuka, Kanagawa, 239-8536 Japan
{zzp,ohya}@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 uses the AURORA2 task to investigate the performance of our proposed tree-structured piecewise-linear transformation (PLT) noise adaptation. In our proposed method, an HMM that best matches the input speech is selected based on the likelihood maximization criterion by tracing a tree structured HMM space that is prepared in the training step, and the selected HMM is further adapted by linear transformation. Experimental results show that our method achieves a significant improvement for the AURORA2 database.

1. Introduction

The performance level 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 demand for robust speech recognition systems is becoming critical. To compare the performance of different algorithms, the AURORA working group that belongs to the technical body STQ (Speech processing, Transmission and Quality aspects) as an ETSI standardization activity has prepared a database for evaluation. The main activity is to develop a front-end feature that will be more robust against noise for DSR (Distributed Speech Recognition). Many researchers have used the AURORA database to evaluate their feature-processing techniques, e.g. [1,2].

One effective approach to robust speech recognition is model adaptation; the acoustic models are adapted to the noise condition [3, 4, 5]. Model adaptation methods have an advantage in that these methods can adapt not only expected values but also distribution functions. Model adaptation methods, such as MLLR and MAP, adapt the observation probabilities of Gaussian mixture components according to the phone list yielded by input speech recognition. Although model adaptation methods generally need two passes, the first pass for phoneme segmentation and the second pass for re-recognition using adapted models, they have an advantage in that they can achieve phoneme dependent adaptation, which is impossible with feature level compensation/normalization methods.

Sasou et al. have proposed a method which produces, on-line, an extended acoustic model by combining a mismatch model with a clean acoustic model trained using only clean speech data [6]. The mismatch model is modeled by time-varying population parameters using a Gaussian Mixture Model (GMM) and a Gain-Adapted Hidden Markov Model (GA-HMM) decomposition method. This method was confirmed to offer significantly improved performance on the AURORA2 database. However, to estimate the GMM, this method needs non-speech frames that are usually detected by a speech/non-speech detector module. The performance largely depends on the result of speech/non-speech detection (SND) and it is inapplicable when the number of detected non-speech frames is insufficient. Another problem is the huge computational cost of performing 3-dimensional Viterbi decoding in the decomposition process.

We have recently proposed the use of tree-structured piecewise linear-transformation (PLT)-based adaptation [7]. PLT is performed in two steps: noisy speech HMM

selection from tree-structured noisy HMM and linear transformation of the selected HMM. Both processes use the likelihood maximization criterion. This method has several advantages; it is unnecessary to use SND for estimating noise, since it does not need any knowledge about the input noise, and also the computational cost is low. The proposed method has been tested in a large vocabulary speech recognition system and shown to achieve better performance than either the MLLR or PMC method [8].

Most researchers have reported combining front-end feature processing and model adaptation methods and AURORA2 database testing [9,10,11]. Our proposed method offers a simple and effective method that can be combined with front-end feature processing methods. This paper investigates the performance of our proposed method on the AURORA2 task. We first briefly explain the method, and then report 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

Noise-added speech spectra vary as a function of both noise spectra and SNR. In our proposed method, a wide variety of noise data are collected and a large number of noisy speech (noise-added speech) data are created by adding each noise signal to a large set of clean speech utterances to recreate several SNR conditions. The noisy speech data with all the combinations of noise and SNR conditions are classified into a tree structure. As it is difficult to cluster noisy speech data directly, noisy speech GMMs for all the conditions are made and used for clustering. The noisy 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.

In the recognition phase, the noise-cluster HMM that best fits the input speech is selected by tracing this tree from the top (root). First, the likelihood value averaged over the test utterance length using the root HMM is calculated. The likelihood values are then calculated using the HMMs of its children nodes. The model that yields the largest likelihood value among these models is selected. If one HMM of the children nodes yields the largest likelihood, it is selected and the search is continued in the same manner. If the parent model yields the largest likelihood, the model is selected as the best model and the search is stopped. The selected HMM model is further converted to reduce mismatches with the input speech by the MLLR-based unsupervised adaptation method. Figure 1 shows a flow diagram of the proposed method.

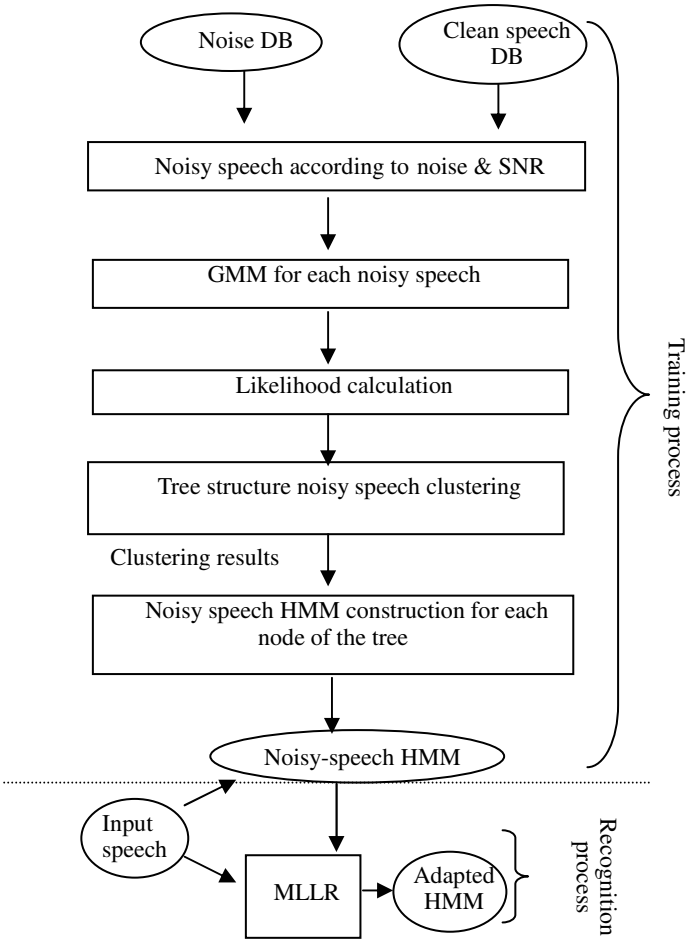

Fig. 1 Tree-structured Piecewise-linear transformation for HMM noise adaptation.

3. Experiments

3.1 Task

The performance of the proposed method is evaluated on the AURORA2 [12] speaker independent connected digit recognition task, i.e. the TIdigit database is used. This contains the recordings of male and female US-American adults speaking isolated digits and sequences of up to 7 digits.

In the AURORA2 evaluation, two training modes

are considered: training on clean data and multi-condition training on noisy data. "Clean" data corresponds to the TIdigit training data filtered with a G712 characteristic. "Noisy" data corresponds to the TIdigit training data filtered with a G712 characteristic and contaminated by artificially added noise at several SNR conditions. In our experiment, we evaluate the performance using the "Clean" training condition.

In the AURORA2 evaluation, three different sets of speech data are subjected to recognition.

- Set "A" consists of TIdigits test data filtered with a G712 characteristic and contaminated by artificially added noise (subway, babble, car, and exhibition noises) at several SNRs.
- Set "B" consists of TIdigits test data filtered with a G712 characteristic and contaminated by artificially added noise (restaurant, street, airport, and train station noises) at several SNRs.
- Set "C" consists of TIdigits test data filtered with a MIRS characteristic and contaminated by artificially added noise (subway and street noises) at several SNRs.

Since the intention of test set C is the investigation of a different frequency characteristic (MIRS instead of G712), we evaluate the performance on A and B sets.

3.2 Acoustic Models

The digits are modeled by whole word HMMs with the following parameters:

- 16 states per word
- simple left-to-right models without skips over states
- mixture of 3 Gaussians per state

A vector size of 39 is defined by using 12 cepstral coefficients (without the zeroth coefficient) and the logarithmic frame energy plus the corresponding delta and acceleration coefficients. Two pause models are defined. The first one, called "sil", consists of 3 states and models the pauses before and after the utterances. A mixture of 6 Gaussians models each state. The second pause model, called "sp", is used to model pauses between words. It consists of a single state, which is tied to the middle state of the first pause model. The training is performed in several steps by applying the Baum-Welch re-estimation scheme.

3.3 Noise Data for Noise Clustering

90 kinds of noises collected by NTT-AT (NTT Advance Technology) were used for noise clustering [13]. Noisy speech samples were made at several SNR values (-5dB, 0dB, 5dB, 10dB, 15dB, 20dB) and noisy speech GMM (64 Gaussian mixture) was trained for each noisy speech sample using the Baum-Welch algorithm. Noisy speech GMMs were then clustered based on the likelihood matrix in which each term was calculated from a pair of noise GMMs.

3.4 Experimental Results

Recognition experiments were performed to evaluate our method. The best matching noise-adapted HMM was selected from the tree and then transformed by MLLR. In the experiments, instantaneous MLLR adaptation was carried out. One sentence for testing was used for both HMM selection and MLLR adaptation.

Table 1 shows the resulting word accuracy. These results show that the proposed method improved the performance significantly; 80.83% and 81.91% averaged over 4 noises and 5 SNR conditions (20, 15, 10, 5, 0dB) were achieved for sets A and B, respectively. They correspond to reductions in the relative error rate of 50.40% and 59.12%, respectively. These results are equivalent to or better than the best performances achieved by model adaptation methods reported so far in the framework of AURORA2 evaluations.

4. Conclusion

This paper has investigated our proposed tree-structured piecewise linear-transformation (PLT)-based noise adaptation using the AURORA2 database. This method consists of two parts: best matching noisy speech HMM selection and linear transformation of the selected HMM. Both processes are based on the likelihood maximization criterion. Experimental results show that the proposed method improved performance significantly.

The proposed method has several advantages compared to other recently investigated methods, such as HMM decomposition and spectral subtraction. That is, it is unnecessary to use Speech/Non-speech Detection (SND) for estimating noise, since it does not need any knowledge about the input noise. In addition, this method needs a relatively small amount of computation.

Future research topics include combining it with front-end processing techniques to further improve its performance.

References

[1] P. Jain et al., "Distributed speech recognition using noise-robust MFCC and traps-estimated manner features", *Proc. ICSLP*, pp. 473-476 (2002)

[2] Y. Wang et al., "Average instantaneous frequency (AIF) and average log-envelopes (ALE) for ASR with the AURORA 2 Database", *Proc. EUROSPEECH*, pp. 25-28 (2003)

[3] 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)

[4] J. L. Gauvain et al., "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains", *IEEE Transactions on Speech and Audio Processing*, Vol. 2, No. 2, pp.291-298 (1994)

[5] M.J.F.Gales et al., "Cepstral parameter compensation for HMM recognition in noise", *Speech Communication*, Vol.12, No.3,pp.231-239 (1993).

[6] A. Sasou et al., "Adaptation of acoustical model using the gain-adapted HMM decomposition method", *Proc. EUROSPEECH*, pp. 29-32 (2003)

[7] Z. Zhang et al., "A tree-structured clustering method integrating noise and SNR for piecewise linear-transformation-based noise adaptation", *Proc. ICASSP* (2004) to appear

[8] Z. Zhang et al., "Piecewise-linear transformation-based HMM adaptation for noisy speech", *Proc. ASRU* (2001)

[9] J. Chen et al., "Bell labs approach to AURORA evaluation on connected digit recognition", *Proc. ICSLP*, pp. 229-232 (2002)

[10] M. Lieb et al., "Progress with the philips continuous ASR system on the AURORA2 noisy digits database ", *Proc. ICSLP*, pp. 449-452 (2002)

[11] M. Fujimoto et al., "Evaluation of noisy speech recognition based on noise reduction and acoustical model adaptation on the AURORA2 tasks", *Proc. ICSLP*, pp. 465-468 (2002)

[12] http://www.ntt-at.co.jp/product/demwa02/index.html

[13] http://www.elda.fr/article52.html

Table 1: Result for clean training multi-condition testing

Cleaning training, multicondition testing - Results										
	A					**B**				
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average
Clean	98.96	99.00	99.20	99.24	99.10	98.98	99.00	99.02	99.34	99.09
20 dB	96.87	98.00	98.12	96.32	97.33	97.14	97.52	97.97	97.12	97.44
15 dB	95.18	95.70	97.52	94.34	95.69	93.92	96.22	96.65	94.90	95.42
10 dB	89.50	90.71	95.07	86.67	90.49	85.02	93.04	93.42	90.88	90.59
5 dB	77.67	74.02	86.43	69.57	76.92	67.54	83.12	81.56	79.84	78.02
0 dB	47.08	45.87	48.18	33.73	43.72	41.06	51.23	54.75	45.33	48.09
-5dB	35.50	35.74	35.61	28.79	33.91	32.47	44.10	47.23	40.23	41.01
Average	81.26	80.86	85.06	76.13	**80.83**	76.94	84.23	84.87	81.61	**81.91**
Relative performance										
	A					**B**				
	Subway	Babble	Car	Exhibition	Average	Restaurant	Street	Airport	Station	Average
Clean	2.80%	0.00%	23.08%	5.00%	7.72%	4.67%	0.00%	5.77%	17.50%	6.99%
20 dB	-6.10%	79.70%	27.41%	-1.94%	24.77%	71.43%	41.78%	78.31%	45.45%	59.24%
15 dB	25.96%	83.61%	75.10%	28.89%	53.39%	74.41%	67.27%	85.43%	68.81%	73.98%
10 dB	50.66%	81.63%	85.06%	45.23%	65.64%	66.88%	78.84%	85.74%	77.03%	77.12%
5 dB	53.32%	64.50%	79.41%	44.84%	60.52%	52.95%	72.58%	73.53%	72.03%	67.77%
0 dB	28.48%	40.33%	39.42%	19.13%	31.84%	33.81%	40.64%	47.13%	38.18%	39.94%
-5dB	27.38%	34.72%	28.94%	21.23%	28.07%	30.04%	37.57%	42.50%	34.71%	36.21%
Average	38.59%	61.81%	62.09%	31.01%	**50.40%**	51.35%	59.01%	67.64%	58.56%	**59.12%**