論文 / 著書情報
Article / Book Information

Title	On-line Incremental Speaker Adaptation with Automatic Speaker Change Detection
Authors	Zhi-Peng Zhang, Sadaoki Furui, Katsutoshi Ohtsuki
Citation	IEEE ICASSP 2000, Vol. 2, No. , pp. 961-964
Pub. date	2000, 6
Copyright	(c) 2000 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
URL	http://www.ieee.org/index.html
DOI	http://dx.doi.org/10.1109/ICASSP.2000.859121
Note	This file is author (final) version.

ON-LINE INCREMENTAL SPEAKER ADAPTATION WITH AUTOMATIC SPEAKER CHANGE DETECTION

Zhi-Peng Zhang[1], Sadaoki Furui[1] and Katsutoshi Ohtsuki[2]

[1] Tokyo Institute of Technology, Department of Computer Science
2-12-1 Ookayama, Meguro-ku, Tokyo, 152-8552 Japan
[2] NTT Cyber Space Laboratories, Media Processing Project
1-1 Hikari-no-oka, Yokosuka-shi, Kanagawa, 239-0847 Japan
{zzp, furui}@cs.titech.ac.jp, ohtsuki@nttspch.hil.ntt.co.jp

ABSTRACT

In order to improve the performance of speech recognition systems when speakers change frequently and each of them utters a series of several sentences, a new unsupervised, on-line and incremental speaker adaptation technique combined with automatic detection of speaker changes is proposed. The speaker change is detected by comparing likelihoods using speaker-independent and speaker-adaptive GMMs. Both the phone HMM and GMM are adapted by MLLR transformation. In a broadcast news transcription task, this method reduces the word error rate by 10.0%. In comparison with the conventional method that uses HMMs for the speaker change detection, the GMM-based method requires a significantly less number of computations at the cost of only a slightly lower word recognition rate.

1. INTRODUCTION

Researchers in the speech recognition field have for many years acknowledged the existence of striking speech-recognition performance inhomogeneities with respect to speakers within a population. Coping with this problem requires speaker adaptation of acoustic models. In many applications of speech recognition, speakers change frequently, new speakers appear, and each of them utters a series of several sentences. In such a situation, an unsupervised and on-line adaptation method, which uses the unknown utterance itself for adaptation, is expected to be effective. The adaptation should also work incrementally within a segment in which one speaker utters several sentences. To create such a system, we must ensure that speaker change is detected automatically and correctly.

We have been developing a Japanese broadcast-news speech transcription system [1]-[2] as a part of a joint research project with NHK broadcasting. The goal is the closed-captioning of TV programs. In this task, speakers frequently change, and each speaker usually utters several sentences in succession. In addition, the speakers include not only professional announcers but also reporters.

The speaker adaptation method that this paper describes combines on-line, unsupervised and incremental speaker adaptation with automatic speaker-change detection. We test two methods of speaker-change detection in a broadcast news transcription task.

2. JAPANESE BROADCAST NEWS TRANSCRIPTION SYSTEM

2.1 Language Models

The language models were constructed using broadcast-news manuscripts taken from NHK TV news broadcasts made between July 1992 and May 1996. The broadcasts comprised roughly 500k sentences and 22M words (morphemes). To calculate word n-gram language models, we segmented the broadcast-news manuscripts into words using a morphological analyzer because 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 constructed a language model in which a word with multiple readings is split into different language model entries according to those readings [1]. We also introduced filled-pause modeling into the language model. A word-frequency list was derived for the news manuscripts, and the 20k most frequently used words were selected as vocabulary words. This 20k vocabulary covered approximately 98% of the words in the manuscripts. We calculated bigrams and trigrams and estimated unseen n-grams using Katz's back-off smoothing method.

2.2 Acoustic Models

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

The acoustic models were gender-dependent shared-state triphone HMMs and were designed using tree-based clustering. They were trained using phonetically-balanced sentences and dialogues read by 53 male speakers and 56 female speakers. The contents were completely different from the broadcast-news task. The total number of training utterances was 13,270 for the males and 13,367 for the females, and the total length of the training data was approximately 20 hours for each gender. The total number of HMM states was approximately 2,000 for each gender, and the number of Gaussian mixture components per state was four.

2.3 Evaluation Data

Clean speech data consisting of 50 male and 50 female utterances with no background noise were extracted from TV news broadcasted in July 1996 and used as evaluation utterances. The male and female sets included utterances by five or six speakers, respectively. All utterances were manually segmented into sentences.

3. BASIC SPEAKER ADAPTATION PROCEDURE

3.1 Overall Procedure

Figure 1 is a flow chart of the speaker adaptation process [2]. The maximum likelihood linear regression (MLLR) [3], MAP, and vector-field smoothing (VFS) [4] methods are instantaneously and incrementally carried out for each utterance. Each input utterance is re-recognized using the adapted model.

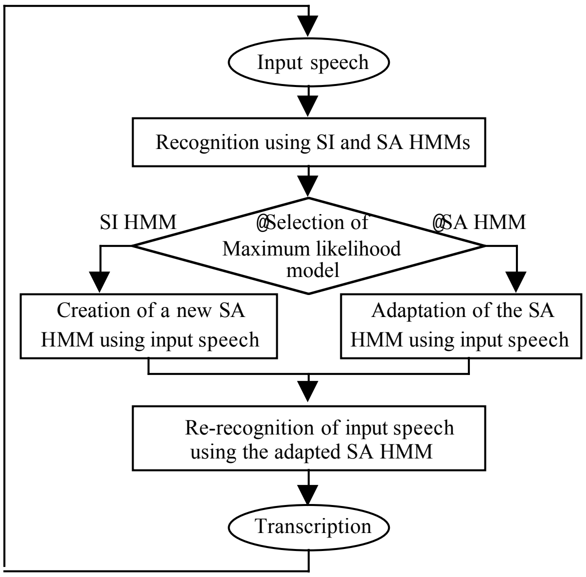
Input speech
Recognition using SI and SA HMMs
SI HMM
@Selection of Maximum likelihood model
@SA HMM
Creation of a new SA HMM using input speech
Adaptation of the SA HMM using input speech
Re-recognition of input speech using the adapted SA HMM
Transcription

Figure 1. Basic speaker adaptation process (SI: speaker independent, SA: speaker adapted).

The adaptation process is as follows.

(1) For the first input utterance, the speaker-independent phone HMM is used for both recognition and adaptation. The resulting model is then stored.

(2) For the second input utterance, the likelihood value of the utterance for the speaker-independent model and that for the speaker-adapted model are calculated and compared. If the former value is larger, the utterance is considered to be from a new speaker, and a new speaker-adapted model is created. Otherwise, the existing speaker-adapted model is incrementally adapted.

(3) For the succeeding input utterances, speaker changes are again detected by comparing the acoustic likelihood values obtained from the speaker-independent model and the speaker-adapted models. If the speaker-independent model yields a larger likelihood than any of the speaker-adapted models, a speaker change is detected and a new speaker-adapted model is constructed. In the experiment, to take advantage of our broadcast-news structure and to reduce the computational time, the two most recently constructed speaker-adapted models were kept and older models were removed.

3.2 Experimental Results

We tried to determine the optimum number of transformation matrices, or clusters, for MLLR given the experimental conditions. We evaluated MLLR with a single cluster, two clusters (silence, and all phonemes), three clusters (silence, vowels, and consonants), seven clusters (silence, consonants, and five Japanese vowels), and nine clusters (silence, five Japanese vowels, and three Japanese consonant categories) using male speech and the bigram language model. Table 1 shows recognition performances. “Baseline” is the case of no speaker adaptation. MLLR with seven clusters achieved the best performance in this experiment.

Table 1. Word error rates [%] for various numbers of clusters for MLLR (male utterances, bigram language model).

Baseline		18.0
MLLR type	1 cluster	15.6
	2 clusters	16.0
	3 clusters	16.4
	7 clusters	14.7
	9 clusters	15.8

Table 2 shows the results of the speaker adaptation experiments for MLLR with a single cluster and with seven clusters. The performance of the MLLR model with a single cluster was better than that of the speaker-independent models, and further improvement was achieved by using the seven phoneme clusters. The latter method reduced the word error rate by 11.8%, averaged over male and female, relative to the results for the speaker-independent models in the case of the trigram language model.

Table 2. Word error rates [%] with several types of MLLR.

Language model	MLLR type	Evaluation sets		
		Male	Female	Average
Bigram	Baseline	18.0	16.1	17.0
	1 cluster	15.6	15.8	15.6
	7 clusters	14.7	14.4	14.5
Trigram	Baseline	14.2	12.9	13.5
	1 cluster	12.7	12.5	12.6
	7 clusters	12.1	11.8	11.9

3.3 Detected Speaker Changes vs. Given Speaker Changes

For comparison with the above experiments, we conducted a speaker adaptation experiment in which correct speaker changes were given. In this experiment, the speaker-independent model was used as the initial model for adaptation, irrespective of the likelihood value, every time the speaker changed. The experimental results, shown in Table 3, indicate that lower word error rates were obtained with automatically detected speaker changes than with given speaker changes. This shows that the acoustic model having the highest likelihood value with respect to the input speech is highly effective as an initial model for speaker adaptation.

Table 3. Comparison of word error rates [%] with detected and given speaker changes (seven clusters MLLR).

Lang. model	Speaker changes	Evaluation sets		
		Male	Female	Average
Bigram	Given	15.7	14.3	15.0
	Detected	14.7	14.4	14.5
Trigram	Given	12.3	11.9	12.1
	Detected	12.1	11.8	11.9

4. MODIFIED SPEAKER ADAPTATION PROCEDURE

4.1 Overall Procedure

The method described in the previous section requires a large number of computations, because each input utterance is recognized using a speaker-independent HMM and multiple speaker-adapted HMMs in parallel. To reduce the computational load, a single-state Gaussian mixture model (GMM) can be used instead of phone HMMs to detect speaker changes.

Figure 2 is a flow chart of the new speaker adaptation process. A single-state speaker-independent GMM with 64 Gaussian distributions is constructed in the training stage using the same utterances that were used to construct the speaker-independent phone HMM. The speaker-independent GMM is then adapted using a MLLR transformation matrix obtained by applying a single-cluster adaptation procedure to the phone HMM.

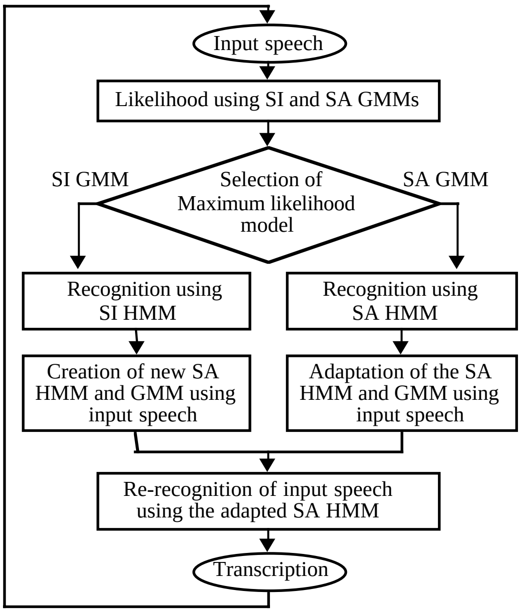
Input speech
Likelihood using SI and SA GMMs
SI GMM
Selection of Maximum likelihood model
SA GMM
Recognition using SI HMM
Recognition using SA HMM
Creation of new SA HMM and GMM using input speech
Adaptation of the SA HMM and GMM using input speech
Re-recognition of input speech using the adapted SA HMM
Transcription

Figure 2. Modified speaker adaptation process (SI: speaker independent, SA: speaker adapted).

For every input utterance, the likelihood value for the speaker-independent GMM and that for speaker-adapted GMMs are calculated and compared. If the former value is larger than any of the values for speaker-adapted GMMs, the utterance is considered to be from a new speaker, and a new speaker-adapted phone HMM and GMM are constructed. If one of the speaker-adapted GMMs yields the largest likelihood for the input utterance, it, along with the the corresponding speaker-adapted HMM, is adapted. The adapted HMM is used for re-recognizing the input utterance.

4.2 Experimental Results

Table 4 compares word error rates for three conditions: no adaptation, the basic adaptation method, and the modified method. In the case of the trigram language model, the modified method reduced the word error rate by 10.0% over that of a method using no adaptation, averaged over male and female. The modified method's recognition performance was slightly worse than the basic method's (11.8%), but the computational efficiency of the modified method was so great (computational time for model selection was reduced to roughly 1/1000 that of the basic adaptation method) that it made up for the lower recognition performance.

Table 4. Word error rates [%] by baseline and basic/modified speaker adaptation methods.

Language model	Adaptation method	Evaluation sets		
		Male	Female	Average
Bigram	Baseline	18.0	16.1	17.0
	Modified method	15.5	14.4	15.0
	Basic method	14.7	14.4	14.5
Trigram	Baseline	14.2	12.9	13.5
	Modified method	12.5	11.8	12.2
	Basic method	12.1	11.8	11.9

4.3 Discussion

Another way to reduce the number of computations required for model selection is to segment the input speech into a phoneme or word sequence by using a speaker-independent HMM and re-calculate the sentence likelihood for each speaker-adapted HMM based on the segmentation result [5]. A supplementary experiment was performed and the results showed that while this method achieved almost the same performance as our basic adaptation method, it still needed a much larger number of computations than did our modified method using GMM.

Table 5. Word error rates [%] for baseline and incremental/instantaneous adaptation.

Language model	Adaptation method	Evaluation sets		
		Male	Female	Average
Bigram	Baseline	18.0	16.1	17.0
	Incremental adaptation	15.5	14.4	15.0
	Instantaneous adaptation	16.1	14.9	15.5
Trigram	Baseline	14.2	12.9	13.5
	Incremental adaptation	12.5	11.8	12.2
	Instantaneous adaptation	12.6	12.3	12.4

Another supplementary experiment was performed with no incremental adaptation. That is, each input utterance was used to adapt the speaker-independent phone HMM to the speaker (instantaneous adaptation) and re-recognized by using the adapted HMM. In the framework of the adaptation method described in this paper, this corresponds to the condition that each input utterance is considered to be an utterance by a new speaker. Table 5 compares the word error rates with/without incremental adaptation. These results indicate that the incremental adaptation is better than the instantaneous adaptation for every case.

5. CONCLUSION

Combining an on-line, unsupervised and incremental speaker adaptation method with automatic detection of speaker changes is applicable to successive utterances spoken by multiple speakers. In the modified version of the method, speaker-independent and adapted GMMs are used for speaker-change detection. Using the maximum likelihood linear regression (MLLR) transformation with seven clusters combined with the vector-field smoothing (VFS) technique, the proposed method reduced word error rate by 10.0% in a Japanese broadcast news transcription task. In comparison with the conventional method using HMM for the speaker change detection, the GMM-based method needs a significantly less amount of computation at the cost of only slightly lower performance.

Future research includes using a better initial speaker-independent model for speaker-change detection and adaptation and combining the proposed speaker adaptation method with noise and channel adaptation methods to make the recognition system more robust against a wider range of mismatch conditions.

ACKNOWLEDGMENTS

The authors would like to thank NHK (Japan Broadcasting Corporation) for providing us with the broadcast news database. The authors would also like to thank Dr. T. Fuchi at NTT Cyber Space Laboratories for letting us use their morphological analysis program. This work is supported in part by the International Communications Foundation.

REFERENCES

[1] K. Ohtsuki et al., “Improvements in Japanese broadcast news transcription,” Proc. DARPA Broadcast News Workshop, pp. 231-236, 1999.

[2] K. Ohtsuki et al., “Recent advances in Japanese broadcast news transcription,” Proc. Eurospeech’99, pp. 671-674, 1999.

[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] K. Ohkura et al., “Speaker adaptation based on transfer vector field smoothing with continuous mixture density HMMs,” Proc. ICSLP’92, pp. 369-372, 1992.

[5] T. Hain et al., “Segment generation and clustering in the HTK broadcast news transcription system,” Proc. DARPA Broadcast News Transcription and Understanding Workshop, pp. 133-137, 1998.