Rapid Unsupervised Speaker Adaptation Based on Multi-template HMM Sufficient Statistics in Noisy Environments

Randy Gomez, Akinobu Lee, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate School of Information Science
Nara Institute of Science and Technology , JAPAN
E-mail: {randy-g,ri,sawatari,shikano}@is.naist.jp

ABSTRACT

This paper describes a multi-template unsupervised speaker adaptation based on HMM-Sufficient Statistics. Multiple class-dependent models based on gender and age are used to push up the adaptation performance while keeping adaptation time within few seconds with just one arbitrary utterance. Adaptation begins with the estimation of speaker's class from the N-best neighbor speakers using Gaussian Mixture Models (GMM) on the way of speaker selection. The corresponding template model is adopted as a base model. The adapted model is rapidly constructed using the selected HMM-Sufficient Statistics. Experiments in noisy environment conditions with 20dB SNR office, crowd, booth, and car noise are performed. The proposed multi-template method achieved 89.5% word correct rate compared with 88.0% of the conventional single-template method, while the baseline recognition rate without adaptation is 85.7%. Moreover, experiments using Vocal Tract Length Normalization (VTLN) and supervised Maximum Likelihood Linear Regression (MLLR) are also compared.

1. Introduction

Using sufficient amounts of training data results to an accurate Speaker-Independent (SI) acoustic model. When using multiple database having wide varieties of gender and age among speakers, speaker variability becomes an issue. The SI model will have an increase in variance, which in turn degrades the recognition performance. There are several methods in addressing this problem[1].

A trivial approach to deal with speaker variability problem is to train multiple classes of acoustic models with smaller variances. Cluster-based modeling is proposed [2][3] which results to a better recognition performance when using an appropriate model selection method. The utilization of normalization techniques such as VTLN [4] effectively compensates the different sizes of speakers' vocal tracts through frequency warping. Experiments in adult and children data yielded an improvement in recognition accuracy when using VTLN [5].

Model adaptation is one of the prominent approaches that effectively adjusts the SI model to reflect the inherent characteristics of the adaptation data to the adapted model. MLLR [6] for example is a very powerful adaptation technique. Combinations of the above-mentioned methods together with model adaptation is also common. Incorporating VTLN together with MLLR [7] is reported. Also, MLLR adaptation using class-dependent models through gender classification [1] also proved to be effective. However, to achieve a good recognition performance, sufficient amounts of adaptation data in several utterances with phoneme transcriptions are needed [8].

We have previously proposed an unsupervised speaker adaptation based on HMM-Sufficient Statistics [8]. Relevant and promising works similar to this approach include transformation and combination of HMM models for speaker selection and training [9], smoothed N-Best based and eigenvoices speaker adaptation [10] [11] but these methods are more of an offline adaptation scheme.

In this paper we extend HMM sufficient statistics adaptation using multiple template models in preparing the HMM-Sufficient Statistics. The proposed method performs better than the conventional approach [8] while maintaining the execution time in few seconds with just one utterance of adaptation data without its corresponding phoneme transcriptions. We performed recognition experiments in 20dB SNR office, car, crowd and booth noise environments. Furthermore the proposed method which requires only one arbitrary utterance without phoneme transcriptions performs better than MLLR using 10 utterance adaptation data with phoneme transcriptions, and also compared with VTLN.

2. HMM-Sufficient Statistics Adaptation

The basic concept of this adaptation technique is the calculation of the sufficient statistics per speaker offline, and use these in creating an adapted model. Sufficient statistics are the statistical parameters that include the means, variances, Expectation Maximization EM counts of the Hidden Markov Models. Model adaptation by means of HMM-Sufficient Statistics refers to the reconstruction of the pre-stored sufficient statistics of the individ-

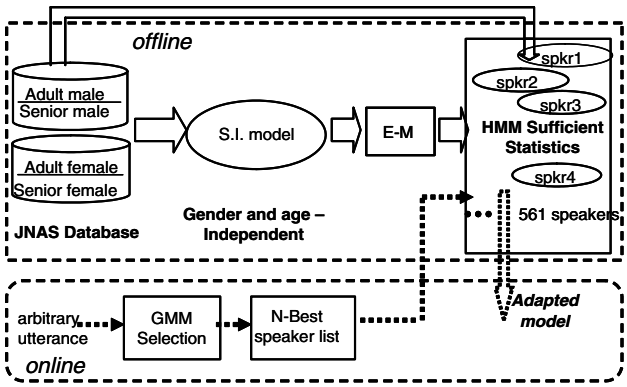

Figure 1: Block Diagram of the Conventional HMM-Sufficient Statistics Adaptation

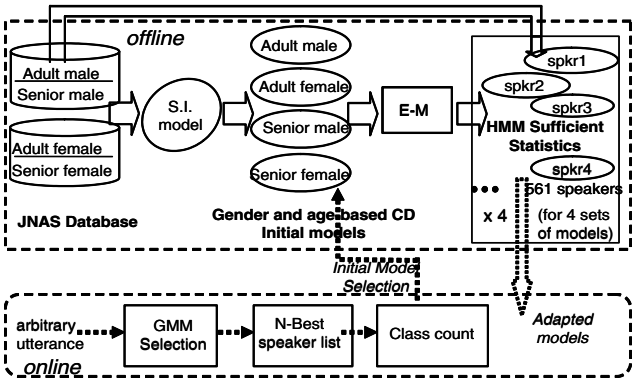

Figure 2: Block Diagram of HMM-Sufficient Statistics Multiple Models Adaptation

ual speaker. Model reconstruction is facilitated through a model selection process which will be explained in latter sections. Figure 1 is the block diagram of the conventional HMM-Sufficient Statistics adaptation. In this approach, only one set of HMM-Sufficient Statistics per speaker corresponding to the speaker independent model is created.

3. Proposed Method

The proposed method in Figure 2, we take advantage of using the multi-template models in creating the HMM-Sufficient Statistics. Gender and age information of the training data are emphasized and embedded in the HMM-Sufficient Statistics. During the reconstruction process, the adapted model has an improved discrimination performance among different classes of speaker's acoustical characteristics. The proposed method gives the system more degrees of freedom to choose the appropriate template model and its corresponding HMM-Sufficient Statistics that is close to the single arbitrary utterance.

3.1. Acoustic Modeling and HMM-Sufficient Statistics

In the offline portion of Figure 2, an SI model is trained irregardless of classes using all of the training data from JNAS adult database consisting of 60K-utterance from 301 male and female speakers and JNAS Senior database

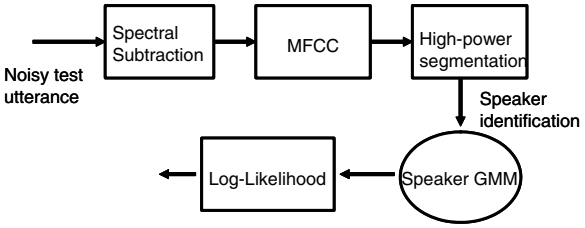

Figure 3: Proposed method : GMM Selection using the Noisy Utterance

Table 1: Execution Time of the Proposed Method

Process	Execution time
GMM Selection	2 sec
Adaptation	10 sec
Total	12 sec

with 53K-utterance from 260 male and female speakers. From this SI model, multi-template HMM models are created namely: Adult male, Adult female, Senior male and Senior female. Note that these models are both age and gender dependent. Consequently, four sets of HMM-Sufficient Statistics for each speaker are created which are equivalent to one-iteration of the Expectation Maximization (EM) training with four multi-template HMMs. These HMM-Sufficient Statistics are then stored for online adaptation in later part.

3.2. Speaker Selection

Speaker selection shown in Figure 2 of the proposed adaptation method are explained below:

1) N-best speakers close to the test utterance are selected using the GMM speaker dependent models. This process returns a list of log-likelihood among all speakers in the GMM model. In Figure 3, the noisy test utterance is being denoised using Spectral Subtraction (SS) and then parameterized (MFCC). To minimize the effects of the residual noise that is present in the silence or unvoiced region of the speech utterance, the low power part is removed and only the MFCCs that have high energy is retained for speaker selection.

2) From the log-likelihood list, only N-best speakers are selected for adaptation, narrowing down the log-likelihood list to N-speakers that are close to the test utterance basing the log-likelihood scores.

3) From the N-speakers list, a class count is performed for the 4 different templates from the speaker labels using the speaker IDs.

4) Template model is selected based on the class count. The class that has the most counts will correspond to the selected template model.

3.3. HMM-Sufficient Statistics Adaptation

As soon as a model template has been selected, the system will rapidly carry through the online adaptation procedures. Table 1 is the summary of the actual adaptation

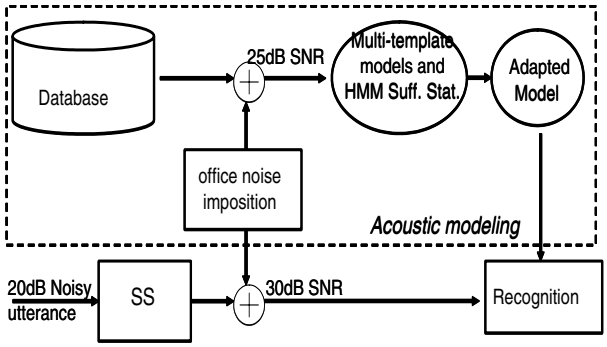

Figure 4: Block Diagram of the Overall System

time using one arbitrary utterance without phoneme transcriptions

4. Experimental Results

The performance using the conventional HMM-Sufficient Statistics adaptation is compared with the proposed multi-template acoustic models in creating HMM-Sufficient Statistics (gender and age-based) in 20dB noisy environments. In addition, VTLN is also examined. Lastly, speaker adaptation using MLLR trained with 10 and 50 utterances is also compared.

4.1. Experimental Conditions

The language model is provided by the IPA dictation toolkit. Phonetically tied mixture models are trained by superimposing 25dB office noise to the database in creating the multi-template models [12]. Figure 4 shows the overall block diagram of the system. In the acoustic modeling part, office noise is superimposed to the clean speech from the database that results to 25dB SNR which is used in training. In the adaptation part, the single arbitrary noisy utterance is denoised with SS which is used for speaker selection as outlined in section 3.2. Lastly, for the actual recognition test, the SS-denoised test utterances are superimposed with 30dB office noise prior to recognition [12].

The testsets are grouped into four classes namely: Adult male, Adult female, Senior male, Senior female respectively. Each testset is of 100 utterances from 23 speakers which are taken outside from the training database. Office, crowd, booth and car noise are superimposed which results to 20dB SNR noise utterances respectively. Denoising of the test utterance as shown in Figure 4 is done using SS, the denoised test utterance is then superimposed with 30dB office noise in order to neutralize the residual noise effects after SS, and this kind of approach has been successfully implemented under different types of noise [12]. The 30dB office noise-superimposed to the denoised utterance is tested for recognition performance using JULIUS with 20K-word on Japanese newspaper dictation task from JNAS.

The significance of using a single 25dB office noise acoustic model of this type instead of noise-matched

Table 2: Word Correct Summary of Figure 5 (Average of the Four Testsets)

Model	Average
SI(no adapt)	85.7%
SI(HMM-Suff.Stat adapt)	88.0%
Multi-template(no adapt)	87.9%
Proposed Method	89.5%
Multi-template(VTLN adapt)	88.2%

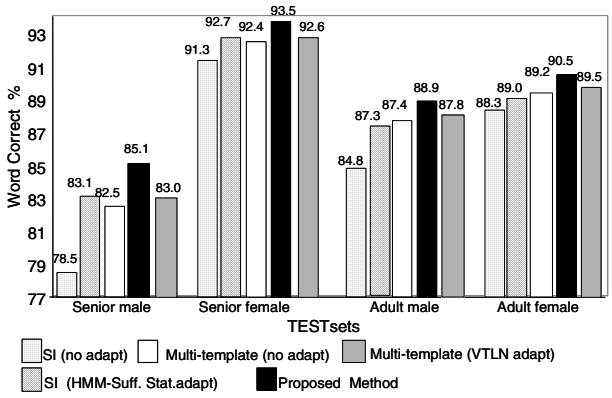

Figure 5: Average Recognition Performance Under Four Noisy Environment Conditions

models is the fact that in real application there exist so many types of noise and it would be impractical to create matched models for each of these noise types. The 25dB office noise acoustic modeling together with SS and 30dB office noise superimposition to the test uterrances is robust enough against various noisy conditions [12]. We can check for the robustness of the proposed method under several types of noise using only a single noise-adapted model rather than several noise-matched models.

4.2. Recognition Results

Figure 5 is the average recognition performance using the four types of noise for all testsets. This is summarized in Table 2 given the model conditions and the average word correct rate over all the testsets. The baseline result using SI model without adaptation SI(no-adapt) is 85.7% while the conventional approach using SI model with HMM-Sufficient Statistics adaptation SI(HMM-Suff. Stat. adapt.) is 88.0%. The multi-template models without adaptation Multi-template(no adapt) is higher than the SI(no adapt) with 87.9% word correct. Lastly, the proposed method which is the multi-template with HMM-Sufficient Statistics adaptation has 89.5% of word correct. The proposed method has a 1.5% absolute improvement compared to the conventional HMM-Sufficient Statistics adaptation with the same amount of execution time.

4.3. VTLN Results

VTLN was applied to the training data to create the multi-template models and to the test data as well. Average

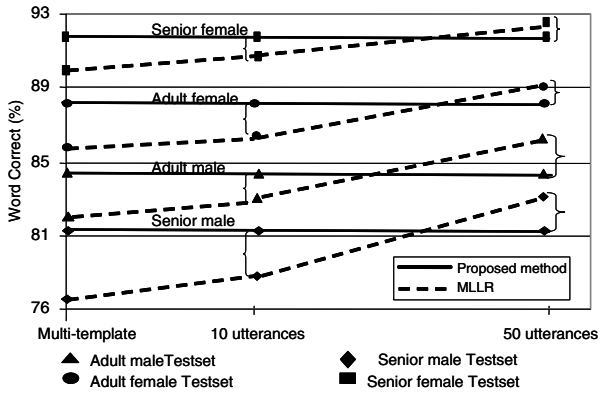

Figure 6: Comparison of Recognition Results Using 10 and 50 Adaptation Utterances for MLLR and 1 Adaptation Utterance for the Proposed Method

recognition rate over all testsets in Figure 5 is 88.2% which is 1.3% lesser than the proposed multi-template HMM-Sufficient Statistics adaptation which has a recognition performance of 89.5%.

4.4. Recognition Results Using MLLR

MLLR results for the gender and age-based multiple acoustic models are given in Figure 6. In the abscissa, the labels 10 and 50 utterances correspond to the adaptation data for MLLR, while the label multi-template corresponds to the recognition performance without adaptation data, or simply using the multi-template models only. This result is the average of all the different noisy conditions. The broken lines represent the supervised MLLR adaptation results for 10 and 50 utterances, while the unbroken lines show the results of the proposed multi-template HMM-Sufficient Statistics unsupervised adaptation which only needs one arbitrary utterance without phoneme transcriptions.

It is shown that the proposed method works better than that of the supervised MLLR when using 10-utterance adaptation data. Although, 50 adaptation utterances for MLLR performs better than the proposed method, adaptation time also goes much further beyond the 12-second execution time of the proposed method. Braces in the graph show the difference between the proposed multi-template HMM-Sufficient Statistics adaptation and supervised MLLR. Moreover, it is obvious that the recognition rate using the multi-template models alone without adaptation is having the least recognition performance as compared to the proposed method and MLLR.

5. Conclusion

We successfully extended the conventional unsupervised HMM-Sufficient Statistics speaker adaptation using a large database into using multi-template acoustic models HMM-Sufficient Statistics unsupervised speaker adaptation. While the proposed method performs better than VTLN and MLLR using 10 adaptation utterances, it executes adaptation in less than 12 seconds. In the future, we will be working on cluster-based HMM-Sufficient Statistics adaptation to investigate the effects of further reducing the adaptation time and the recognition performance.

6. Acknowledgment

This work is supported by the MEXT e-Society project of Japanese Ministry of Education.

7. References

[1] C. Huang et al. “Analysis of Speaker Variability”, *In Proceedings of Eurospeech*,Vol. 2, pp 1377-1380 September 2001

[2] Gales M. “Cluster-Adaptive Training For Speech Recognition”, *In Proceedings of ICSLP*, pp. 1783-1786 1998.

[3] B. Xiang et al. “Cluster-Dependent Acoustic Modeling”, *In Proceedings of ICASSP*,Vol.1, pp. 677-680 2005.

[4] P.C. Woodland et al. “Experiments in Speaker Normalisation and Adaptation for Large Vocabulary Adaptation”, *In Proceedings of ICASSP*,Vol.2,No.1,pp.1047-1051,Apr1997

[5] Giuliani and M. Gerosa et al. “Investigating Recognition of Children‘s Speech”, *In Proceedings of ICASSP*,Vol 2, pp. 137-140 April 2003

[6] C.J.Leggeter and Woodland “ Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models” *In Proceedings of Computer Speech and Language*, vol.9,pp.171-185, 1995

[7] P. Zhan et al. “Speaker Normalization and Speaker Adaptation- A combination for conversational Speech Recognition”, *In Proceedings of Eurospeech*,Vol. pp. 2087-2090 September 1997

[8] S. Yoshizawa K. Shikano, “Unsupervised Speaker Adaptation Based on Sufficent HMM Statistics of Selected Speakers”, *In Proceedings of ICASSP*, 2001

[9] C. Huang et al., “Transformation and Combination of Hidden Markov Models for Speaker Selection Training” *In Proceedings of ICSLP*, 2004.

[10] Matsui T. et al. “Smoothed N-Best Based Speaker Adaptation for Speech Recognition ”,*In Proceedings of ICASSP*, pp. 1015-1018, 1997.

[11] Kuhn R. et al. “EigenVoices For Speaker Adaptation”, *In Proceedings of ICSLP* , pp. 1771-1774 1998.

[12] S. Yamade, K. Matsunami, A. Baba, A. Lee, H. Saruwatari, K. Shikano , “Spectral Subtraction In Noisy Environments Applied To Speaker Adaptation Based on HMM Sufficient Statistics” *In Proceedings of ICSLP*, pp. I-1045-1048, 2000.