IMPROVING RAPID UNSUPERVISED SPEAKER ADAPTATION BASED ON HMM SUFFICIENT STATISTICS

Randy Gomez, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano

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

ABSTRACT

In real-time speech recognition applications, there is a need to implement a fast and reliable adaptation algorithm. We propose a method to reduce adaptation time of the unsupervised speaker adaptation based on HMM-Sufficient Statistics. We use only a single arbitrary utterance without transcriptions in selecting the N-best speakers' Sufficient Statistics created offline to provide data for adaptation to a target speaker. Further reduction of N-best implies a reduction in adaptation time. However, it degrades recognition performance due to insufficiency of data needed to robustly adapt the model. Linear interpolation of the global HMM-Sufficient Statistics offsets this negative effect and achieves a 50% reduction in adaptation time without compromising the recognition performance. We have reduced the adaptation time from 10 sec to 5 sec without degradation of the word accuracy. Furthermore, we compared our method with Vocal Tract Length Normalization (VTLN), Maximum A Posteriori (MAP) and Maximum Likelihood Linear Regression (MLLR). Moreover, we tested in office, car, crowd and booth noise environments in 10 dB, 15 dB, 20 dB and 25 dB SNRs.

1. INTRODUCTION

Automatic Speech Recognition system has a very important role in human-machine interface. For the system to be practical, it should be usable to wide variety of speakers. Mismatch due to different age-groups and genders results in speaker variability problem which degrades the performance of the recognizer [1]. There are several methods in addressing this problem. For instance, training multiple classes of acoustic models with smaller variance [2]. Normalization of the vocal tract such as VTLN [3] has also been proposed. Model adaptation such as MLLR [4] and MAP [5] for example is proven to be very effective. Another method, is the transformation and combination of HMMs [6]. To achieve a good recognition performance, sufficient amounts of adaptation data in several utterances with phoneme transcriptions are needed in the case of MLLR and MAP [7], which raises the issues like execution time and size of adaptation data.

We have previously proposed a rapid unsupervised speaker adaptation based on HMM-Sufficient Statistics requiring only one adaptation utterance with a 10 sec adaptation time [7] [8]. Relevant and promising work in rapid adaptation includes the linear combination of rank-one matrices, which can handle very short adaptation data [9]. Also, a very fast compact context-dependent eigenvoice model adaptation is said to work even with minimal amount of data [10].

In this paper we extended the conventional unsupervised HMM Sufficient Statistics speaker adaptation using linear interpolation to further reduce the adaptation time. The proposed method can adapt in 5 sec time, which is 50% faster than the conventional method.

This paper is organized as follows. In section 2, HMM-Sufficient Statistics adaptation is introduced. Section 3 discusses the proposed method, then experimental results are presented in section 4 comparing different adaptation techniques. Finally, we conclude this paper in section 5.

2. HMM-SUFFICIENT STATISTICS ADAPTATION

Sufficient Statistics summarizes all the information in a sample about a target parameter which allows for an observation (training data) which is huge in size to be compactly represented in low-dimensional parameters. The concept of the unsupervised HMM-Sufficient Statistics speaker adaptation is summarized in two steps. First, we estimate the individual Sufficient Statistics of each speaker in the training database (offline). Next step is to make use of these Sufficient Statistics to provide data for adaptation to a target speaker through N-best speaker selection. Since estimation of Sufficient Statistics can be done offline, adaptation will not require any model estimation. Only updating of the model parameters using the Sufficient Statistics is needed. This renders the proposed method to execute very fast.

Figure 1 is a block diagram of the conventional HMM-Sufficient Statistics adaptation. First, the Speaker-Independent (SI) model is trained regardless of classes using all of the training data from the JNAS adult database consisting of 60K-utterance from 301 male and female speakers and the JNAS Senior database with 53K-utterance from 260 male and female speakers [1], where each speaker contributes 200 utterances. From this SI model, multi-template HMM models are created namely: Adult male, Adult female, Senior male and

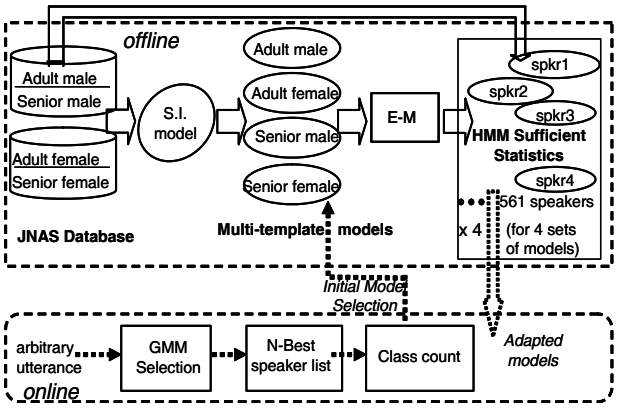

Fig. 1. Conventional HMM-Sufficient Statistics adaptation.

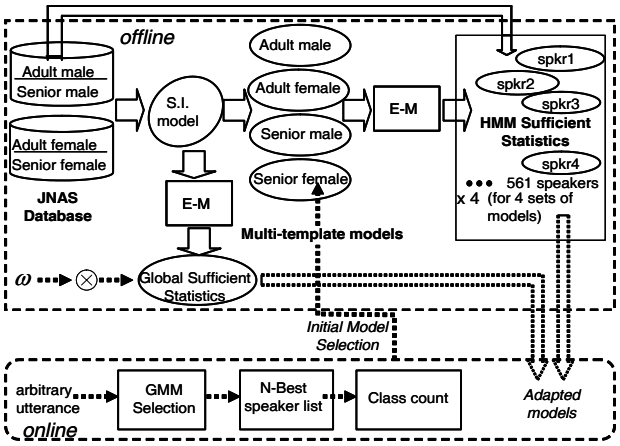

Fig. 2. Proposed HMM-Sufficient Statistics adaptation with linear interpolation.

Senior female. Consequently, four sets of HMM-Sufficient Statistics for each speaker are created which are equivalent to one-iteration of the Expectation Maximization (E-M) training with four multi-template HMMs.

2.1. Limitations of the Conventional HMM-Sufficient Statistics Adaptation

The recognition performance and adaptation speed of this approach are dependent on the number of N-best speakers, S. Experiments showed that the optimal N-best is $S_{optimal} = 40$ which corresponds to a 10-second adaptation time [7] [11] [8]. If S is further reduced such that $S < S_{optimal}$, adaptation time is reduced with a trade-off of the recognition performance. This is attributed to the fact that further decreasing S would result to insufficient data necessary to robustly estimate the target speaker's HMMs.

3. PROPOSED HMM-SUFFICIENT STATISTICS ADAPTATION WITH LINEAR INTERPOLATION

To address the problem discussed in section 2.1, we introduced linear interpolation using the global Sufficient Statistics. Figure 2 shows the proposed weighting of the global Sufficient Statistics. The proposed method makes it possible to robustly estimate the target speaker's HMMs even with N-best reduced ($S < S_{optimal}$) since the weighted global Sufficient Statistics offsets the negative effect of the removed statistical information. The adapted HMM parameters are as follows :

$$C_{im}^{adp_{new}} = \frac{\sum_{s=1}^{S} L_{im}^{s} + \omega L_{im}^{global}}{\sum_{m=1}^{M}(\sum_{s=1}^{S} L_{im}^{s} + \omega L_{im}^{global})}, \quad (1)$$

$$\boldsymbol{\mu}_{im}^{adp_{new}} = \frac{\sum_{s=1}^{S} \boldsymbol{m}_{im}^{s} + \omega \boldsymbol{m}_{im}^{global}}{\sum_{s=1}^{S} L_{im}^{s} + \omega L_{im}^{global}}, \quad (2)$$

$$\boldsymbol{\Sigma}_{im}^{adp_{new}} = \frac{\sum_{s=1}^{S} \boldsymbol{v}_{im}^{s} + \omega \boldsymbol{v}_{im}^{global}}{\sum_{s=1}^{S} L_{im}^{s} + \omega L_{im}^{global}} - \boldsymbol{\mu}_{im}^{adp} \boldsymbol{\mu}_{im}^{adp^T}, \quad (3)$$

$$a_{ij}^{adp_{new}} = \frac{\sum_{s=1}^{S} L_{i\to j}^{s} + \omega L_{i\to j}^{global}}{\sum_{j=1}^{J}(\sum_{s=1}^{S} L_{i\to j}^{s} + \omega L_{i\to j}^{global})}, \quad (4)$$

where $C_{im}^{adp_{new}}$,$\boldsymbol{\mu}_{im}^{adp_{new}}$, $\boldsymbol{\Sigma}_{im}^{adp_{new}}$, $a_{ij}^{adp_{new}}$ are the newly updated mixture weight, means, covariance matrix and updated transition probability using linear interpolation. L_{im}^{S}, $L_{i\to j}^{S}$, $\boldsymbol{m}_{im}^{S}$, $\boldsymbol{v}_{im}^{S}$ are the probability of mixture component occupancy, the accumulated probability of the state occupancy, means and variance respectively of the selected N-best speakers S. L_{im}^{global}, $L_{i\to j}^{global}$, $\boldsymbol{m}_{im}^{global}$, $\boldsymbol{v}_{im}^{global}$ are the probability of the mixture occupancy, the accumulated probability of the state occupancy , means and variance respectively which are estimated using all of the training data which constitute the global Sufficient Statistics. ω is the weighting factor of the global HMM-Sufficient Statistics. In this paper, we used the following weighting factors :

$$\omega = \tau_1, \quad (5)$$

$$\omega = \frac{\tau_2}{\tau_2 + L_{i\to j}^{global}}, \quad (6)$$

where in eqn (5) we used a multiplying constant τ_1 and in eqn (6), the weighting factor ω is normalized by the accumulated probability of the state occupancy, $L_{i\to j}^{global}$.

3.1. Speaker Selection

Speaker selection process starts with 1) the denoising of the noisy test utterance using Spectral Subtraction (SS), then the parameterization to MFCC. To reduce the effects of the residual noise that is present in the silence or unvoiced region of the speech utterance, the low power parts are removed prior to speaker selection. 2) We find the log-likelihood scores given the arbitrary test utterance and the individual-speaker GMMs. 3) From the log-likelihood scores, only N-best speakers are selected for adaptation. 4) From the N-best list, a class count

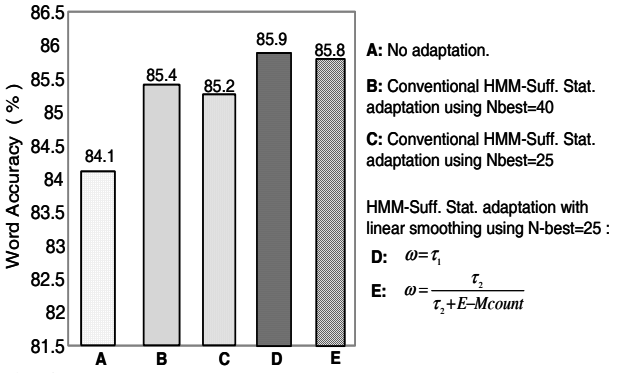

Fig. 3. Recognition performance in 25 dB office noise environment

Table 1. Word accuracy in four different noisy environment conditions (proposed/conventional).

Noise	10 dB	15 dB	20 dB	25dB
office	67.0/66.1	77.2/76.3	83.5/82.7	85.9/85.2
car	81.4/79.7	85.1/84.9	86.3/85.6	87.0/86.3
crowd	65.8/65.1	79.3/78.6	83.7/83.1	84.5/83.9
booth	44.6/44.0	69.1/68.4	82.8/82.1	83.4/82.8

is performed for the 4 different templates. Class counting is carried out using the speaker labels in the form of speaker IDs. Template model is selected based on this count. 5) Template model, N-best HMM-Sufficient Statistics and the global HMM-Sufficients Statistics are used for adaptation.

4. EXPERIMENTAL RESULTS

Phonetically tied mixture models (PTM) are trained by superimposing 25 dB office noise to the database [11] in creating the multi-template models. In the acoustic modeling part, office noise is superimposed to the clean speech from the database that results to 25 dB SNR [11] 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 3.1. Lastly, for the actual recognition test, the SS-denoised test utterances are superimposed with 30 dB office noise prior to recognition to neutralize the residual noise [11].

The test set is composed of four classes, namely: adult male, adult female, senior male and senior female. Each class is of 100 utterances from 23 speakers which are taken outside of the training speakers. This sums up to 400 total test utterances from 92 test speakers across different genders and age-groups. Recognition experiments are carried out using JULIUS with 20K-word on Japanese newspaper dictation task from JNAS.The language model is provided by the IPA dictation toolkit.

Weighting factors given in equations (5) and (6) achieved best results when $0 < \tau_1 < 0.2$ and $1 \leq \tau_2 \leq 2$. In particular we used $\tau_1 = 0.015$ and $\tau_2 = 2$.

4.1. General Result

In Figure 3, the word accuracy when using no adaptation is 84.1% (A), while the conventional HMM-Sufficient Statistics adaptation is 85.4% using N-best $S = 40$ (B). It is apparent that when N-best is reduced to $S = 25$ (C), the word accuracy drops to 85.2%. This points to the fact that merely reducing the selected N-best in the conventional approach results to an insufficient statistical data needed to robustly estimate the target speaker's HMMs as mentioned in section 2.1. The proposed HMM-Sufficient Statistics adaptation with linear interpolation using the two different weighting factors given in equations (5) and (6) has a recognition performance of 85.9% (D) and 85.8% (E) respectively which is approximately 0.7% higher than (C) when using the same amount of N-best $S = 25$. It also outperforms the conventional approach even when using the optimal N-best $S_{optimal} = 40$. It clearly shows that the negative effect in the estimation of the HMMs caused by reducing N-best from $S_{optimal} = 40$ to $S = 25$ is compensated by the linear interpolation of the global Sufficient Statistics. As a result, execution time becomes faster owing to fewer N-best. In Table 1, the summary of recognition performance in office, crowd, car and booth noise environments with different SNRs are given. In this table the proposed method has an adaptation time of 6 sec.

4.2. Experiments with Clustering

We extended the proposed adaptation method by clustering the speakers in the database as opposed to using only individual speakers in Figure 2. In this scheme, the individual-speaker GMMs and HMM-Sufficient Statistics are changed to cluster-based. The N-best generates the list of clusters that are close to the target speaker. The motivation of this appproach is to further reduce adaptation time by reducing N-best. Although, a further reduction of N-best poses a problem due to the insufficient statistical data as dicussed in section 2.1, this problem is minimized by combining 2 speakers statistical information in each cluster and at the same time incorporate linear interpolation.

In order to keep the statistical information uniform in the N-best list, we impose that each cluster be composed of a uniform number of speakers (i.e 2 speakers per cluster) by using Minimax [12]. We also implemented K-Means clustering but the former has a better recognition performance. Figure 4 is the plot of the word accuracy comparing 1) individual speakers (unclustered) with interpolation, 2) clustered speakers with and without linear interpolation as a function of N-best. The N-best list for the unclustered speakers are the individual speakers itself while the latters' N-best list is composed of clustered speakers. It is very clear that the proposed linear interpolation improves the performance of the clustered speakers as opposed to the clustered speakers without linear interpolation. More interestingly, the clustered speakers with linear interpolation using N-best =20 achieved a better recog-

Table 2. Execution time of the proposed method using Intel XEON 2.4 GHz processor with 1GB of memory

HMM-Suff. Stat. adaptation	Execution time
Conventional	10 sec
Linear interp. w/ individual speakers	6 sec
Linear interp. w/ clustered speakers	5 sec

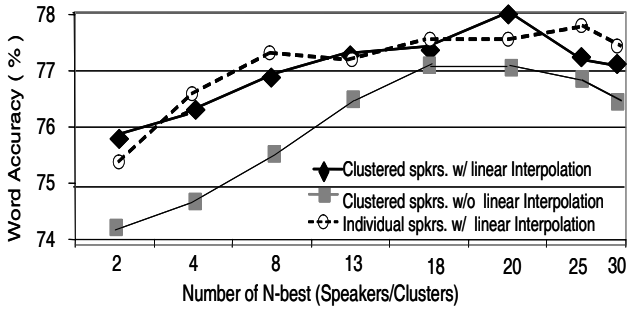

Fig. 4. Clustered speakers' HMM-Sufficient Statistics adaptation with linear interpolation (Averaged in all noisy environment conditions and SNRs).

nition performance with that of using the individual speakers (unclustered) with N-best = 25, thus a reduction in adaptation time is further achieved. Table 2 is the summary of the adaptation time.

4.3. Supervised MLLR, MAP, and VTLN Results

Figure 5 compares the proposed method against MLLR and MAP. We also combined VTLN with MLLR (VTLN+MLLR) and VTLN with MAP (VTLN+MAP) by warping both the training and the testing data before performing MLLR using 128 classes and MAP respectively. In the abscissa, the labels 10 and 50 utterances correspond to the adaptation data for the MLLR and MAP variants.

The proposed method works better than that of the supervised MLLR, MAP, VTLN+MLLR and VTLN+MAP when using 10-utterance adaptation data. When using 50 utterances, MLLR and VTLN+MLLR has a better performance than the proposed method while MAP together and VTLN+MAP are still outperformed by the proposed method. It should be noted that when using 50-utterances of adaptation data, MLLR and MAP are performed offline while the proposed method can execute the adaptation process in 5 sec using only a single arbitrary adaptation utterance without transcriptions.

5. CONCLUSION

We have succesfully reduced the adaptation time from 10 sec to 6 sec with linear interpolation of the global HMM-Sufficient Statistics. A further reduction to 5 sec is obtained by clustering the speakers' HMM-Sufficient Statistics together with linear interpolation. Most interestingly, the reduction in N-best which reduces adaptation time is achieved without degrading the recognition performance. In fact, it slightly improved the

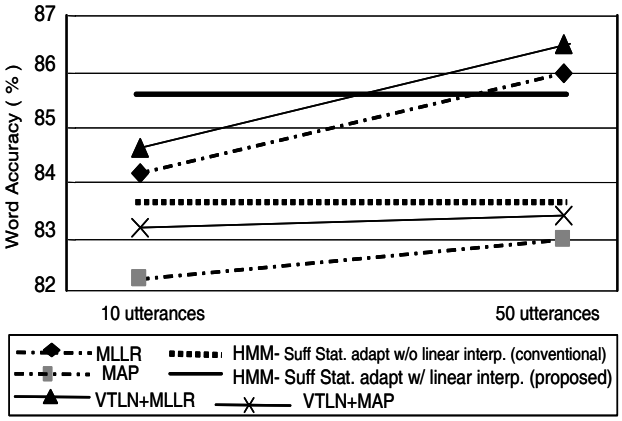

Fig. 5. Recognition performance with various adaptation techniques.

word accuracy. Furthermore, the system works well under office, crowd, booth and car noise and in different SNRs.

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

6. REFERENCES

[1] A. Baba,"Elderly Acoustic Model for Large Vocabulary Continuous Speech Recognition" *In Proceedings of EUROSPEECH*, pp. 1657-1660, 2001.

[2] C. Huang, et al., "Analysis of Speaker Variability", *In Proceedings of EUROSPEECH*,Vol. 2, pp 1377-1380 September 2001

[3] 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

[4] 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

[5] S. Young, et al., "The HTK Book"

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

[7] S. Yoshizawa, et al., "Unsupervised Speaker Adaptation Based on Sufficent HMM Statistics of Selected Speakers" *In Proceedings of ICASSP*,2001

[8] R. Gomez, et al., "Rapid Unsupervised Speaker Adaptation Based on Multi-template HMM Sufficient Statistics in Noisy Environments" *In Proceedings of EUROSPEECH*,pp 296-301, 2005.

[9] G. Vatbhava, V. Karthik, G. Ramesh, "Rapid Adaptation with Linear Combinations of Rank-one Matrices", *In Proceedings of ICASSP*, 2001

[10] R. Kuhn, F. Perronnin, P. Nguyen, J. Junqua, L. Rigazio, "Very Fast Adaptation with a Compact Context-Dependent Eigenvoice Model", *In Proceedings of ICASSP*, 2002

[11] S. Yamade, et al., "Spectral Subtraction In Noisy Environments Applied To Speaker Adaptation Based on HMM Sufficient Statistics" *In Proceedings of ICSLP* , pp. I-1045-1048 2000.

[12] R. Gomez, et al., "Speaker-Class Reduction for HMM-Sufficient Statistics Adaptation Using Multiple Acoustic Models" *In Proceedings of Acoustical Society of Japan*, 2005.