IEICE TRANS. INF. & SYST., VOL.E90–D, NO.2 FEBRUARY 2007

PAPER

Reducing Computation Time of the Rapid Unsupervised Speaker Adaptation Based on HMM-Sufficient Statistics

Randy GOMEZ[†a)], *Nonmember*, **Tomoki TODA**[†], **Hiroshi SARUWATARI**[†], *and* **Kiyohiro SHIKANO**[†], *Members*

SUMMARY 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 rapid 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. 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.

key words: *HMM-sufficient statistics, unsupervised, rapid adaptation, speech recognition*

1. Introduction

Research in speech recognition has advanced very rapidly. With the availability of free softwares, speech database and tools, this field has become more accessible with a wider spectrum of applications. It is expected for this application to have a wide variety of users. Mismatch due to different age-groups and genders causes a problem of speaker variability which degrades the performance of the recognizer [1]. In line of the abovementioned application in speech recognition, it is imperative to design a flexible system that can adapt from various users and most importantly, can carry out adaptation rapidly using a minimum amount of arbitrary adaptation data.

Using multiple training database of different genders and age-group is inevitable to train an accurate Speaker-Independent (SI) acoustic model. However, this SI model will have an increase in variance due to the wide varieties of speakers in the multiple training database. There are several methods in addressing this problem [2]. A trivial approach to deal with speaker variability problem is to train multiple classes of acoustic models with smaller variances. An improvement in the recognition performance using cluster-based modeling approach is possible [3] especially when using an appropriate model selection method. The utilization of normalization techniques such as VTLN [4], [5] 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 [6].

Model adaptation effectively adjusts the SI model to reflect the inherent characteristics of the adaptation data to the adapted model. MLLR [7] and MAP [8] for example are proven to be very effective. Another method is the transformation and combination of HMMs [9] and the smooth N-best based speaker adaptation approach [10]. Works relevant to fast speaker adaptation include the linear combination of rank-one matrices, which can handle very short adaptation data [11]. Also, a very fast compact context-dependent eigenvoice model adaptation works even with minimal amount of data [12].

Unsupervised speaker adaptation based on HMM-Sufficient Statistics is a promising approach for a fast adaptation using only one adaptation utterance [13]. We proposed multi-template HMM-Sufficient Statistics adaptation and further improved recognition performance while keeping adaptation time within 10 sec [14], [15]. In this paper we further improved [15] by looking into the possibility of further reducing the current adaptation time (10 sec) [15]. Techniques such as weighting of the N-best HMM-Sufficient Statistics, interpolation of the global HMM-Sufficient Statistics combining with the clustered speakers are explored. The proposed method has achieved 50% adaptation time reduction compared to [15] without degrading the recognition performance

This paper is organized as follows. In Sect. 2, we introduce the HMM-Sufficient Statistics adaptation together with the problems of reducing adaptation time when using the conventional approach. Sect. 3 discusses the proposed interpolation of the global HMM-Sufficient Statistics. In Sect. 4 we discuss the technique of weighting the HMM-Sufficient Statistics. Experimental results are presented in Sect. 5 with comparisons of different adaptation techniques such as VTLN, MAP and MLLR, and combining MAP and MLLR with VTLN. Finally, we conclude this paper in Sect. 6.

Manuscript received April 4, 2006.
Manuscript revised June 14, 2006.
[†]The authors are with Nara Institute of Science and Technology, Ikoma-shi, 630–0192 Japan.
a) E-mail: randy-g@is.naist.jp
DOI: 10.1093/ietisy/e90–d.2.554

2. HMM-Sufficient Statistics Adaptation

2.1 Background

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. In our application, these parameters are as follows:

$$L_{im}^{spkr} = \sum_{r=1}^{R_{spkr}} \sum_{t=1}^{T_r} L_{im}^{r}(t) , \tag{1}$$

$$\boldsymbol{m}_{im}^{spkr} = \sum_{r=1}^{R_{spkr}} \sum_{t=1}^{T_r} L_{im}^{r}(t)\boldsymbol{O}_t^r , \tag{2}$$

$$\boldsymbol{v}_{im}^{spkr} = \sum_{r=1}^{R_{spkr}} \sum_{t=1}^{T_r} L_{im}^{r}(t)\boldsymbol{O}_t^r\boldsymbol{O}_t^{rT} , \tag{3}$$

L_{im}^{spkr} is the accumulated probability of the state occupancy while $\boldsymbol{m}_{im}^{spkr}$ and $\boldsymbol{v}_{im}^{spkr}$ are the mean and variance of a particular state i and mixture component m respectively as represented by the subscript im, R_{spkr} is the total number of speakers in the training data and $\boldsymbol{O}$ is the observation vector.

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) given by the Eqs. (1)–(3). 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.

2.2 Conventional HMM-Sufficient Statistics Adaptation

Model adaptation by means of Sufficient Statistics refers to the updating of the target speaker's model parameters using the pre-estimated HMM-Sufficient Statistics through N-best speaker selection. The updated model parameters are as follows:

$$C_{im}^{adp} = \frac{\sum_{s=1}^{S} L_{im}^{s}}{\sum_{s=1}^{S}\sum_{m=1}^{M} L_{im}^{s}} , \tag{4}$$

$$\boldsymbol{\mu}_{im}^{adp} = \frac{\sum_{s=1}^{S} \boldsymbol{m}_{im}^{s}}{\sum_{s=1}^{S} L_{im}^{s}} , \tag{5}$$

$$\boldsymbol{\Sigma}_{im}^{adp} = \frac{\sum_{s=1}^{S} \boldsymbol{v}_{im}^{s}}{\sum_{s=1}^{S} L_{im}^{s}} - \boldsymbol{\mu}_{im}^{adp}\boldsymbol{\mu}_{im}^{adpT} , \tag{6}$$

$$a_{ij}^{adp} = \frac{\sum_{s=1}^{S} L_{i\to j}^{s}}{\sum_{s=1}^{S}\sum_{j=1}^{J} L_{i\to j}^{s}} , \tag{7}$$

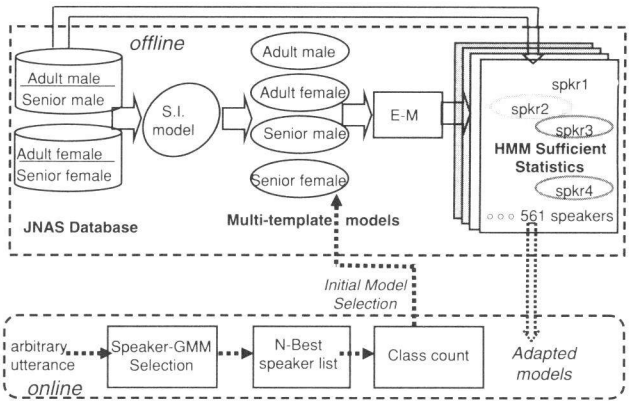


Fig. 1 Block diagram of the conventional HMM-Sufficient Statistics adaptation.

where C_{im}^{adp}, $\boldsymbol{\mu}_{im}^{adp}$ $\boldsymbol{\Sigma}_{im}^{adp}$, and a_{ij}^{adp} are the updated mixture, mean weight, covariance matrix and updated transition probability respectively. L_{im}^{S}, $L_{i\to j}^{S}$, $\boldsymbol{m}_{im}^{S}$, and $\boldsymbol{v}_{im}^{S}$ are the probability of mixture component occupancy, the accumulated probability of the state occupancy, means and variance respectively.

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 60 K-utterance from 301 male and female speakers and the S-JNAS Senior database with 53 K-utterance from 260 male and female speakers [1]. From this SI model, multi-template HMM models are created namely: Adult male, Adult female, Senior male and 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.

2.3 N-Best 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 is performed for the 4 different templates. Class counting is carried out using the speaker labels that are present in the speaker IDs, and template model is selected based on the class count. 5) Template model and N-best HMM-Sufficient Statistics are prepared for adaptation.

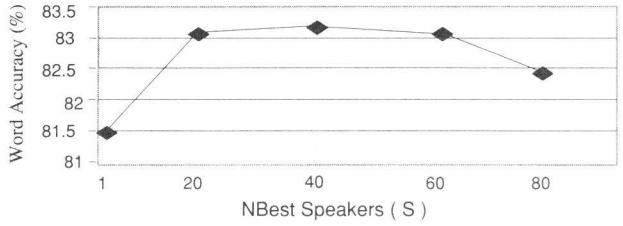


Fig. 2 Relationship between Nbest and recognition performance.

2.4 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 [13], [14], [17]. If S is further reduced such that $S < S_{optimal}$, adaptation time is reduced with a trade-off of the recognition performance as illustrated in Fig. 2. 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. HMM-Sufficient Statistics Adaptation with Linear Interpolation

3.1 Effects of Linear Interpolation

To address the problem discussed in Sect. 2.4, we introduced linear interpolation using the global Sufficient Statistics. Figure 3 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:

$$C_{im}^{adp_{new}} = \frac{\sum_{s=1}^{S} L_{im}^{s} + \omega L_{im}^{global}}{\sum_{m=1}^{M}\left(\sum_{s=1}^{S} L_{im}^{s} + \omega L_{im}^{global}\right)}, \tag{8}$$

$$\boldsymbol{\mu}_{im}^{adp_{new}} = \frac{\sum_{s=1}^{S} \boldsymbol{m}_{im}^{s} + \omega \boldsymbol{m}_{im}^{global}}{\sum_{s=1}^{S} L_{im}^{s} + \omega L_{im}^{global}}, \tag{9}$$

$$\boldsymbol{\Sigma}_{im}^{adp_{new}} = \frac{\sum_{s=1}^{S} \boldsymbol{v}_{im}^{s} + \omega \boldsymbol{v}_{im}^{global}}{\sum_{s=1}^{S} L_{im}^{s} + \omega L_{im}^{global}} - \boldsymbol{\mu}_{im}^{adp_{new}} \boldsymbol{\mu}_{im}^{adp_{new}\,T}, \tag{10}$$

$$a_{ij}^{adp_{new}} = \frac{\sum_{s=1}^{S} L_{i\to j}^{s} + \omega L_{i\to j}^{global}}{\sum_{j=1}^{J}\left(\sum_{s=1}^{S} L_{i\to j}^{s} + \omega L_{i\to j}^{global}\right)}, \tag{11}$$

where $C_{im}^{adp_{new}}$, $\boldsymbol{\mu}_{im}^{adp_{new}}$, $\boldsymbol{\Sigma}_{im}^{adp_{new}}$, $a_{ij}^{adp_{new}}$ are the newly updated

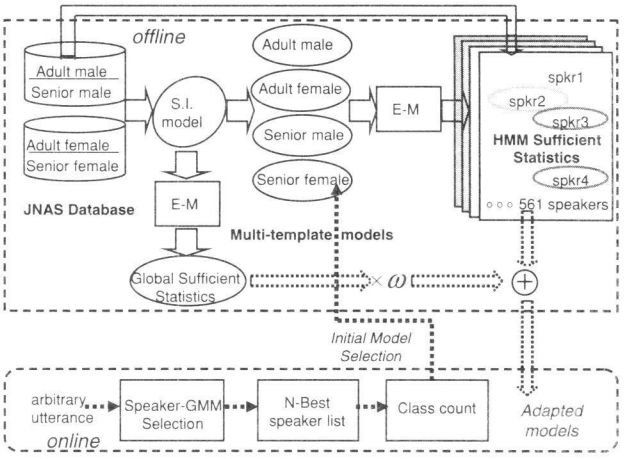


Fig. 3 Block diagram of HMM-Sufficient Statistics adaptation with linear interpolation using individual speakers' Sufficient Statistics.

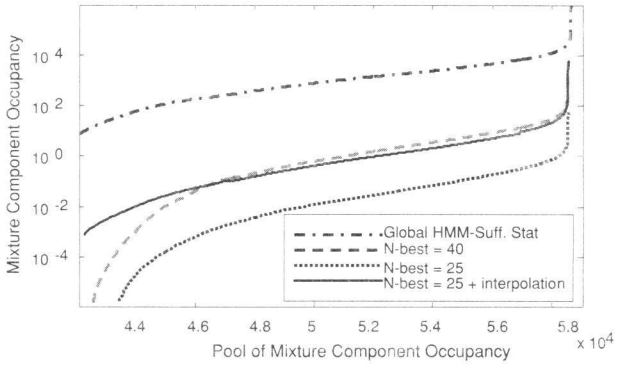


Fig. 4 Effects of linear interpolating the global HMM-Sufficient Statistics adaptation.

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 Fig. 4 we show the graph of the HMM-Sufficient Statistics particularly the mixture component occupancy (in logscale) versus the pool of all Gaussian mixtures. In this figure, we show the effect of merely reducing N-best from 40 to 25 (without interpolation). This is manifested by a decrease in the mixture component occupancy as depicted by the shifting of the envelope (N-best=25) downwards relative to N-best=40. This can be translated to a reduction in the recognition performance because reducing the number of selected N-best means reducing the adaptation data. On the other hand, the effect of the proposed linear interpolation pushes back the envelope of the N-best=25 close to N-

best=40. The supposed decrease in the mixture component occupancy is compensated by the interpolation of the global HMM-Sufficient Statistics. This would mask the detrimental effect in the recognition performance brought by decreasing N-best.

3.2 Clustered Speakers' HMM-Sufficient Statistics

We extended the proposed adaptation method by clustering the speakers in the database as opposed to using only individual speakers. In this scheme, the individual-speaker GMMs are changed to cluster-based GMMs. Likewise, the individual HMM-Sufficient Statistics are changed to clustered speakers' HMM-Sufficient Statistics. The N-best generates the list of clusters that are close to the target speaker. The motivation of this approach 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, 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 [18]. We also implemented K-means clustering but the former has a better recognition performance.

4. Sufficient Statistics Weighting

HMM-Sufficient Statistics adaptation makes use of the N-best speakers to select the HMM-Sufficient Statistics as adaptation data. The selected N-best speakers has a corresponding likelihood scores. We utilize this likelihood scores to introduce weighting of the individual sufficient statistics prior to adaptation as shown in Fig. 5. Weighting of the N-best HMM-Sufficient Statistics emphasizes the ones that are close to the test utterance while it attenuates those that are not so close. In effect, it would be possible to reduce the N-best needed for adaptation. The weight of the HMM-Sufficient Statistics is defined as,

$$W_i = \frac{P(\mathrm{O}|\lambda_i)}{\sum_{s=1}^{S} P(\mathrm{O}|\lambda_s)} \quad , \tag{12}$$

where W_i is the weight, $P(O|\lambda_i)$ is the likelihood of the observation O given the GMM model λ and S is the the number of selected speakers.

Figure 6 shows the mixture component occupancy of the selected N-best speakers (top) and its corresponding likelihood scores (bottom). On top, the light shaded bars are the unweighted mixture component occupancy (HMM-Sufficient Statistics) which is in general, flat over N-best while the dark bars represent its weighted version, based on likelihood. Furthermore, the weigted HMM-Sufficient Statistics has a decreasing trend over N-best speakers depending on the likelihood scores of the individual speaker as shown in the bottom.

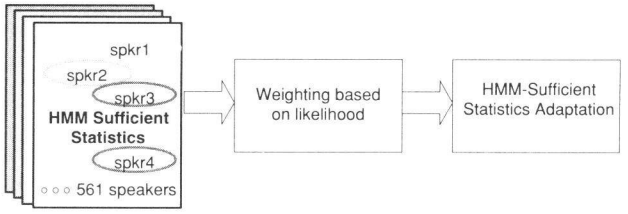


Fig. 5 Block diagram of weighting HMM-Sufficient Statistics prior to adaptation.

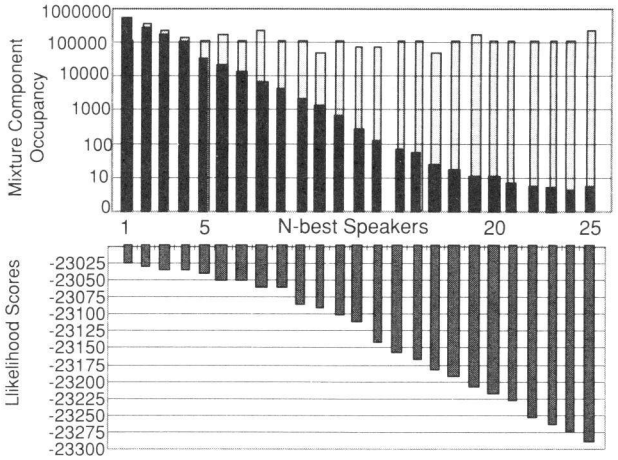


Fig. 6 Effects of Weighting the individual HMM-Sufficient Statistics.

5. Experimental Evaluation

Phonetically tied mixture models (PTM) [16] are trained by superimposing 25 dB office noise to the database [17] 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 [17] 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 Sect. 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. Recognition experiments are carried out using JULIUS with 20 K-word on Japanese newspaper dictation task from JNAS. The language model is provided by the IPA dictation toolkit. Summary of the basic experimental condition parameters used in this set-up is provided in Table 1. 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. The speakers used in testing are differents speakers from that of the training database.

5.1 HMM-Sufficient Statistics Set-up

The JNAS database is composed of Adult and Senior. The

Table 1 System specifications.

Sampling frequency	16 kHz
Frame length	25 ms
Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCC, 12-order ΔMFCCs 1-order ΔE
HMM	PTM, 2000 states
Training data	Adult and Senior by JNAS
Test data	Adult and Senior by JNAS

Table 2 Set-up of the HMM-Sufficient Statistics using JNAS Adult and Senior Database.

Gender	No. of Speakers	Utterances per speaker
Adult Male	150	150
Adult Female	150	150
Senior Male	130	200
Senior Female	130	200

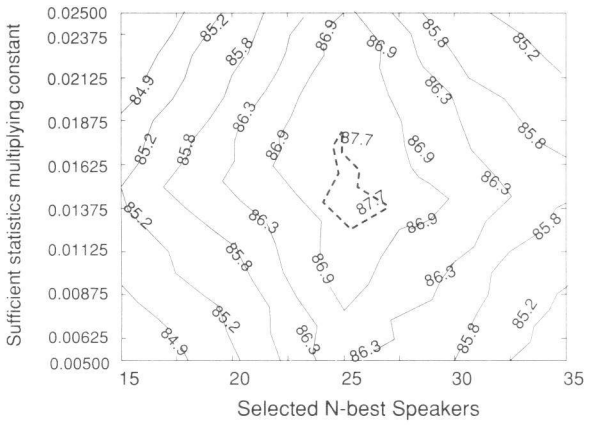


Fig. 7 Setting the value of the multiplying factor for interpolation.

speakers are classified as Adult Male, Adult Female, Senior Male and Senior Female. Every speaker in the database has four HMM-Sufficient Statistics corresponding to the four classes. Each of these is of approximately 5.5 MB in size stored in the disk. Details of the number of speakers and utterances used in creating HMM-Sufficient Statistics offline is given in Table 2.

5.2 Optimization of ω

In our experiment, the value of the multiplying factor used for interpolation is set heuristically to maximize the recognition performance of the testing database. Figure 7 is the contour plot of the WA at different values of the multiplying constant ω and corresponding N-best speakers selected. With the aid of this figure we set $\omega = 0.015$ with N-best=25.

5.3 General Results

In Fig. 8, the WA 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

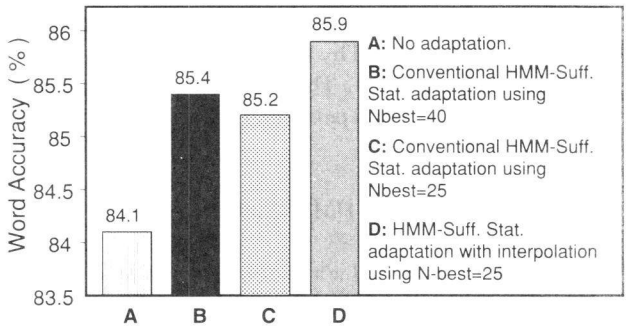


Fig. 8 Recognition performance for 25 dB Office noise environment.

when N-best is reduced to $S = 25$ (C), the WA 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 Sect. 2.4. The proposed HMM-Sufficient Statistics adaptation with linear interpolation has a recognition performance of 85.9% 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 2, the summary of recognition performance in office, crowd, car and booth noise environments with different SNRs are given. Here, it is shown that reducing N-best of the conventional approach to $S = 25$ degrades the recognition performance of the conventional approach using $S_{optimal} = 40$. However, when using the proposed linear interpolation, we can use $S = 25$ without degrading the recognition performance. More interestingly, the result is consistent in all noisy environments and SNRs.

5.4 Results of Clustered Speakers' HMM-Sufficient Statistics

Figure 9 is the plot of the WA 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. 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 can achieve the same recognition performance with that of using the individual speakers (unclustered) with N-best = 25, thus a reduction in adaptation time is further achieved.

The results in different noisy environment conditions and different SNRs using clustered speakers' HMM-Sufficient Statistics are given in Table 3. In this table, it is

Table 3 Word accuracy using individual speakers' HMM-Sufficient Statistics (conventional: $S_{optimal}$ = 40 / conventional: S = 25 / proposed with linear interpolation S = 25).

Noise	10 dB	15 dB	20 dB	25 dB
office	66.5 / 66.1 / 67.0	76.7 / 76.3 / 77.2	83.1 / 82.7 / 83.5	85.4 / 85.2 / 85.9
car	80.0 / 79.7 / 81.4	85.0 / 84.9 / 85.1	85.8 / 85.6 / 86.3	86.6 / 86.3 / 87.0
crowd	65.5 / 65.1 / 65.8	79.0 / 78.6 / 79.3	83.5 / 83.1 / 83.7	84.2 / 83.9 / 84.5
booth	44.3 / 44.0 / 44.6	68.7 / 68.4 / 69.1	82.5 / 82.1 / 82.8	83.2 / 82.8 / 83.4

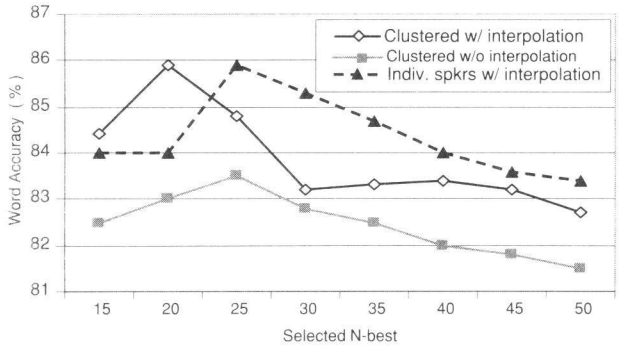


Fig. 9 Performance of the clustered speakers' HMM-Sufficient Statistics adaptation with linear interpolation.

Table 4 WA using clustered speakers' HMM-Sufficient Statistics (conventional: S = 20 / proposed: with linear interpolation S = 20).

Noise	10 dB	15 dB	20 dB	25 dB
office	62.3 / 67.1	74.3 / 77.2	78.9 / 83.6	82.4 / 85.9
car	76.8 / 81.5	82.0 / 85.1	83.3 / 86.4	84.2 / 87.0
crowd	61.5 / 65.9	73.9 / 79.3	79.8 / 83.8	80.6 / 84.5
booth	40.3 / 44.6	64.7 / 69.2	78.6 / 82.9	80.1 / 83.4

apparent that employing linear interpolation improves WA performance. Also, the recognition performance did not show any degradation with S = 20 relative to S = 25 of Table 4. Thus, a further reduction of N-best is possible when using linear interpolation with clustered speakers' HMM-Sufficient Statistics.

5.5 Results of Individually Weighting the N-Best Speakers' HMM-Sufficient Statistics

In Fig. 10, it is shown that in both cases (1) and (2) where interpolation is not carried out, the word accuracy performance benefits from weighting the HMM-Sufficient Statistics. The graph shows that with N-best=34, (2) can achieve its best performance while it takes N-best=38 for (1). In cases of (3) and (4) where linear interpolation is implemented, the performance of both weighting and no-weighting is at par with N-best<30. With very few N-best, the system suffers from insufficiency of adaptation data (see Fig. 2). Thus interpolation of HMM-Sufficient Statistics far outweighs the effect of weighting when using only few N-best. On the othe hand, as N-best is further increased way passed N-best=30, weighting the HMM-Sufficient Statistics (4) tends to be robust in degradation of the recognition performance as opposed to (3). It is in this condition where the system is faced with too much adaptation data (see Fig. 2) and suffers a decrease in the recognition performance. The negative effect of too much adaptation data is reduced through HMM-Sufficient Statistics weighting.

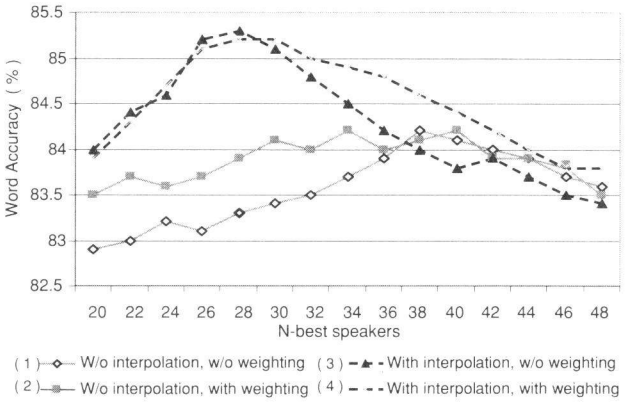


Fig. 10 Recognition performance using weighted HMM-Sufficient Statistics.

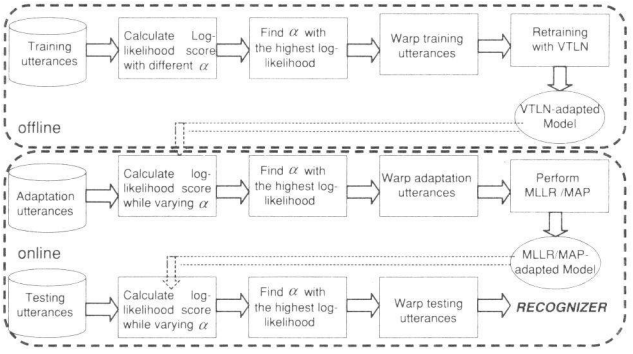


Fig. 11 Block diagram of the supervised VTLN adaptation in finding for the optimum α.

5.6 Comparisons with VTLN, MAP and MLLR

We carried out experiments with MAP and MLLR. We also combined VTLN with MLLR (VTLN+MLLR) and VTLN with MAP (VTLN+MAP). Figure 11 shows the case of combining VTLN and MAP/MLLR. In the *offline* part of this figure, we search for the warping parameter α that maximizes the log-likelihood score of the training database. This is used to warp all of the training utterances and used to re-estimate the VTLN-adapted model. Consequently, in the *online* part, we do the same process of finding α of the adaptation utterances using the VTLN-adapted model and warped these utterances prior to MLLR/MAP adaptation. The process of finding α is repeated again for the last time using the MLLR/MAP adapted model to the test utterances.

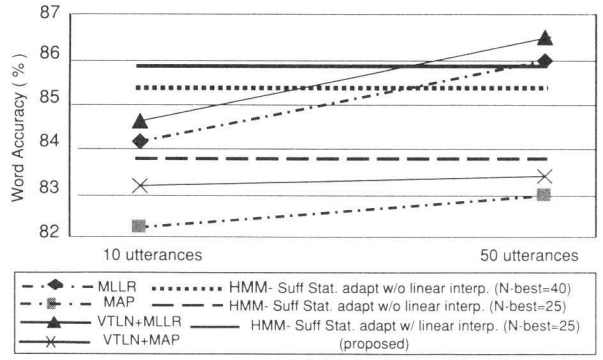


Fig. 12 Recognition performance with various adaptation techniques.

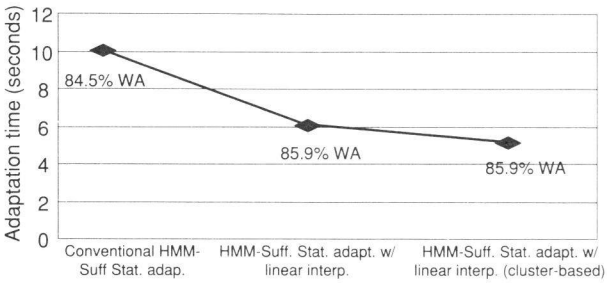


Fig. 13 Summary of adaptation time reduction using Intel XEON 2.4 GHz with 1 GB of memory.

Finally, we warp the testing utterances for recognition experiment.

In Fig. 12, we show the recognition performance using MAP, MLLR, VTLN+MAP and VTLN+MLLR. In the abscissa, the labels 10 and 50 utterances correspond to the adaptation data for the MLLR and MAP variants. The proposed method works best among the supervised MLLR, MAP, VTLN+MAP and VTLN+MLLR when using 10-utterance adaptation data. When using 50 utterances, MLLR and VTLN+MLLR has a better performance compared to the proposed method, while MAP and VTLN+MAP are still outperformed by our 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 adapt in 5 sec time using only a single arbitrary adaptation utterance without transcriptions. We have succesfully reduced the adaptation time from 10 sec [15] to 6 sec when using linear interpolation of the global HMM-Sufficient Statistics as shown in Fig. 13. A further reduction to 5 sec is obtained by clustering the speakers' HMM-Sufficient Statistics together with the proposed linear interpolation. In the case of the supervised MLLR and MAP using 50 utterances, execution time can be as high as 60 sec excluding the time to collect and transcribed these utterances. On the other hand, VTLN+MLLR and VTLN+MAP require much more time to carry out adaptation.

6. Conclusion

In this paper, we proposed linear interpolation of the global HMM-Sufficient Statistics to reduce adaptation data by reducing N-best speakers HMM-Sufficient Statistics to reduce adptation time. The reduction of adaptation time is achieved without degrading the recognition performance. Furthermore, the system works well under office, crowd, booth and car noise and in different SNRs. With the proposed linear interpolation of the HMM-Sufficient Statistics, it is possible to reduce N-best and adapt to a robust model. We will focus our future research to make use of existing powerful adaptation techniques to using HMM-Sufficient Statistics for a more rapid adaptation and an improved recognition performance.

Acknowledgment

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

References

[1] A. Baba, S. Yoshizawa, M. Yamada, A. Lee, and K. Shikano, "Elderly acoustic model for large vocabulary continuous speech recognition," Proc. EUROSPEECH, pp.1657–1660, 2001.
[2] C. Huang, T. Chen, S. Li, and J.L. Zhou, "Analysis of speaker variability," Proc. Eurospeech, vol.2, pp.1377–1380, Sept. 2001.
[3] B. Xiang, L. Nguyen, S. Matsoukas, and R. Schwartz, "Cluster-dependent acoustic modeling," Proc. ICASSP, vol.1, pp.677–680, 2005.
[4] D. Pye and P.C. Woodland, "Experiments in speaker normalisation and adaptation for large vocabulary adaptation," Proc. ICASSP, vol.2, no.1, pp.1047–1051, April 1997.
[5] P. Zhan, M. Westphal, M. Finke, and A. Waibel, "Speaker normalization and speaker adaptation—A combination for conversational speech recognition," Proc. Eurospeech, vol.10, pp.2087–2090, Sept. 1997.
[6] D. Giuliani and M. Gerosa, "Investigating recognition of children's speech," Proc. ICASSP, vol.2, pp.137–140, April 2003.
[7] C.J. Leggeter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Proc. Computer Speech and Language, vol.9, pp.171–185, 1995.
[8] J. Gauvain and C.H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observation of Markov chains," IEEE Trans. Speech Audio Process., vol.2, no.2, pp.291–298, 1994.
[9] C. Huang, T. Chen, and E. Chan, "Transformation and combination of hidden Markov models for speaker selection training," Proc. ICSLP, pp.1377–1380, 2004.
[10] T. Matsui, T. Matsuoka, and S. Furui, "Smoothed N-best based speaker adaptation for speech recognition," Proc. ICASSP, pp.1015–1018, 1997.
[11] G. Vaibhava, V. Karthik, and G. Ramesh, "Rapid adaptation with linear combinations of rank-one matrices," Proc. ICASSP, vol.1, pp.581–584, 2002.
[12] R. Kuhn, F. Perronnin, P. Nguyen, J. Junqua, and L. Rigazio, "Very fast adaptation with a compact context-dependent eigenvoice model," Proc. ICASSP, vol.1, pp.373–376, 2001.
[13] S. Yoshizawa, A. Baba, K. Matsunami, Y. Mera, M. Yamada, and K. Shikano, "Unsupervised speaker adaptation based on sufficent HMM statistics of selected speakers," Proc. ICASSP, pp.341–344, 2001.
[14] R. Gomez, A. Lee, H. Saruwatari, and K. Shikano, "Rapid unsupervised speaker adaptation based on multi-template HMM sufficient statistics in noisy environments," Proc. EUROSPEECH, pp.296–301, 2005.

[15] R. Gomez, A. Lee, T. Toda, H. Saruwatari, and K. Shikano, "Improving rapid unsupervised speaker adaptation based on HMM-sufficient statistics in noisy environments using multi-template models," IEICE Trans. Inf. & Syst., vol.E89-D, no.3, pp.998–1005, March 2006.

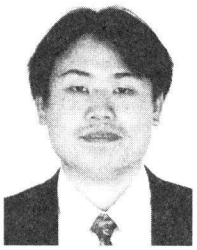
[16] A. Lee, T. Kawahara, K. Takeda, and K. Shikano, "A new phonetic tied-mixture model for efficient decoding," Proc. ICASSP, pp.1269–1272, 2000.

[17] S. Yamade, K. Matsunami, A. Baba, A. Lee, H. Saruwatari, and K. Shikano, "Spectral subtraction in noisy environments applied to speaker adaptation based on HMM sufficient statistics," Proc. ICSLP, pp.I-1045–1048, 2000.

[18] R. Gomez, A. Lee, H. Saruwatari, and K. Shikano, "Speaker-class reduction for HMM-sufficient statistics adaptation using multiple acoustic models," Proc. Acoustical Society of Japan, pp.133–134, March 2005.

Randy Gomez was born in Cebu City, Philippines on December 11, 1976. He received his B.S. degree in Electronics and Communication Engineering at the Mindanao State University-Iligan Institute of Technology in 1998 and served as an instructor immediately after graduation. Received the M. of Eng. Sci. degree in Electrical Engineering at the University of New South Wales (UNSW) in Sydney Australia in 2002. He obtained his Ph.D. in 2006 from the Graduate School of Information Science in Nara Institute of Science and Technology (NAIST). His research interests include robust acoustic modelling and rapid speaker adaptation for practical speech recognition applications. Currently he is connected with the Speech and Acoustics Laboratory in NAIST as a postdoctoral fellow. He is a member of Acoustical Society of Japan, IEEE, and ISCA.

Tomoki Toda was born in Aichi, Japan on January 18, 1977. He received the B.E. degree in electrical engineering from Nagoya University in 1999 and the M.E. and Ph.D. degrees in engineering from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST) in 2001 and 2003, respectively. During 2001–2003, he was an intern researcher and a visiting researcher at ATR Spoken Language Translation Research Laboratories. He was a Research Fellow of the Japan Society for the Promotion of Science (JSPS) in Graduate School of Engineering, Nagoya Institute of Technology during 2003–2005. He was a visiting researcher at Language Technologies Institute, Carnegie Mellon University from October 2003 to September 2004. He is currently an Assistant Professor of the Graduate School of Information Science, NAIST and a visiting researcher at ATR Spoken Language Communication Research Laboratories. His research interests include speech synthesis, speech analysis and speech recognition. He received the TELECOM System Technology Award for Student from the Telecommunications Advancement Foundation in 2003. He is a member of ASJ, IEEE, and ISCA.



Hiroshi Saruwatari was born in Nagoya, Japan, on July 27, 1967. He received the B.E., M.E. and Ph.D. degrees in electrical engineering from Nagoya University, Japan, in 1991, 1993 and 2000, respectively. He joined Intelligent Systems Laboratory, SECOM CO., LTD., Mitaka, Tokyo, Japan, in 1993, where he engaged in the research and development on ultrasonic array system for the acoustic imaging. He is currently an associate professor of Graduate School of Information Science, Nara Institute of Science and Technology. His research interests include array signal processing, blind source separation, and sound field reproduction. He received the Paper Awards from IEICE in 2000, and from TAF in 2004. He is a member of the IEEE, the VR Society of Japan, and the Acoustical Society of Japan.

Kiyohiro Shikano received the B.S., M.S. and Ph.D. degrees in electrical engineering from Nagoya University in 1970, 1972 and 1980, respectively. He is currently a professor of Nara Institute of Science and Technology (NAIST), where he is directing speech and acoustics laboratory. His major research areas are speech recognition, multi-modal dialogue systems, speech enhancement, adaptive microphone array, and acoustic field reproduction. From 1972, he had been working at NT Laboratories, where he had engaged in speech recognition research. During 1990–1993, he was the executive research scientist at NTT Human Interface Laboratories, where he supervised the research of speech recognition and speech coding. During 1986–1990, he was the Head of Speech Processing Department at ATR Interpreting Telephony Research Laboratories, where he was directing speech and speech synthesis research. During 1984–1986, he was a visiting scientist in Carnegie Mellon University, where he was working on distance measures, speaker adaptation, and statistical language modeling. He received the Yonezawa Prize from IEICE in 1975, the Signal Processing Society 1990 Senior Award from IEEE in 1991, the Technical Development Award from ASJ in 1994, IPSJ Yamashita SIG Research Award in 2000, and Paper Award from the Virtual Reality Society of Japan in 2001. He is a member of the Information Processing Society of Japan, the Acoustical Society of Japan, Japan VR Society, the Institute of Electrical and Electronics Engineers (IEEE), and International Speech Communication Society.