社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2005－ＳＬＰ－59（３）
2005／12／21

HMM十分統計量と線形補間法に基づく高速教師なし話者適応の評価

Randy Gomez† 戸田 智基† 猿渡 洋† 鹿野 清宏†

† 奈良先端科学技術大学院大学情報科学研究科 〒630-0101 奈良県生駒市高山町 8916-5
E-mail: †randy-g@is.naist.jp

あらまし　話者性は音声認識性能に大きな影響を与える要因の一つであり、古くから話者適応技術が盛んに研究されている。実環境において音声認識技術を使用する際には、極少量の適応データによる高速な話者適応技術が求められる。これに対して、我々はこれまでにHMM十分統計量に基づく教師なし話者適応に関する研究を行ってきた。この手法では、ユーザーの任意の一発話のみを用いて、話者データベースから声質の近い話者を上位数十人選択し、選択された話者のデータを用いてユーザー用HMMを学習する。予め話者毎にHMM十分統計量を計算しておくことで、モデル学習時の計算量を大幅に削減する事ができる。選択話者数を減らすことで適応に要する時間をさらに減らせる一方で、学習データ量が不十分となるため認識性能は劣化する。本報告では、十分統計量の線形補間法を導入することで、高性能かつ高速な教師なし話者適応を実現する。提案法では、選択話者数減少に伴うデータ量不足を不特定話者に対する十分統計量を用いて補うことで、認識性能の劣化を防ぐ。実験的評価結果から、高い認識性能を維持したまま約５０％の適応時間削減が可能であることを示す。また、他の適応手法(VTLN, MLLR, MAP)との比較結果や、様々な雑音環境下における評価結果についても報告する。

キーワード　高速教師なし話者適応、HMM十分統計量、線形補間、対雑音性

Evaluating Rapid Unsupervised Speaker Adaptation Using Linear Interpolation of HMM-Sufficient Statistics

Randy GOMEZ†, Tomoki TODA†, Hiroshi SARUWATARI†, and Kiyohiro SHIKANO†

† 8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0101
E-mail: †randy-g@is.naist.jp

Abstract　Speaker adaptation techniques minimize the effect of speaker variability. It is neccessary to carry out speaker adaptation rapidly using a minimum amount of adaptation data in real-time application. We propose to improve the unsupervised speaker adaptation based on HMM-Sufficient Statistics using linear interpolation. This adaptation technique uses a single arbitrary utterance to provide data for adaptation by means of selecting N-best speakers' Sufficient Statistics. Reducing the selected N-best speakers implies reduction in adaptation time. However, recognition performance is degraded due to insufficiency of data needed to robustly adapt the model. We introduce linear interpolation of the global HMM-Sufficient Statistics to offset the negative effect of reducing N-best. We achieved a 50% reduction in adaptation time without recognition performance degradation. In our experiment, we have reduced the adaptation time from 10 sec to 5 sec without degrading the recognition performance. Furthermore we compared our method with Vocal Tract Length Normalization (VTLN), Maximum A Posteriori (MAP) and Maximum Likelihood Linear Regression. Moreover, we tested the performance of our approach in office, car, crowd and booth noise environments in 10 dB, 15 dB, 20 dB and 25 dB SNRs.

Key words　Rapid Unsupervised Speaker Adaptation, Noise Robustness, HMM Sufficient Statistics

1. INTRODUCTION

Mismatch due to different classes of age-group and gender results in speaker variability problem which degrades the performance of the recognizer [1]. There are several methods in addressing this problem, like training multiple classes of acoustic models with smaller variance [2]. Normalization of the vocal tract such as VTLN [3] has also been proposed. Model adaptations such as MLLR [4] and MAP [5] for example had proven to be very effective. Transformation and combination of HMMs [6] is also proposed. 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 which requires only one adaptation utterance with 10 seconds adaptation time [7] [8]. Relevant works in rapid adaptation includes linear combination of rank-one matrices [9]and the very fast compact context-dependent eigenvoice model adaptation [10]. In this paper we extend the conventional unsupervised HMM Sufficient Statistics speaker adaptation using linear interpolation to further reduce the adaptation time. The proposed method carries out adaptation in 5 sec 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. Conventional 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. Model adaptation by means of HMM 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 :

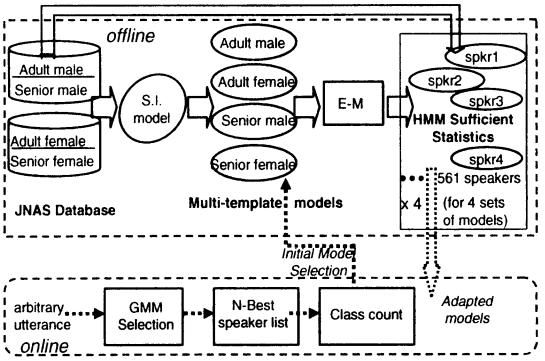

Fig. 1 Block diagram of the conventional HMM-Sufficient Statistics adaptation.

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

$$\boldsymbol{\mu}_{im}^{adp} = \frac{\sum_{s=1}^{S} \boldsymbol{m}_{im}^{s}}{\sum_{s=1}^{S} L_{im}^{s}}, \tag{2}$$

$$\boldsymbol{\Sigma}_{im}^{adp} = \frac{\sum_{s=1}^{S} \boldsymbol{v}_{im}^{s}}{\sum_{s=1}^{S} L_{im}^{s}} - \boldsymbol{\mu}_{im}^{adp}\boldsymbol{\mu}_{im}^{adpT}, \tag{3}$$

$$a_{ij}^{adp} = \frac{\sum_{s=1}^{S} L_{i\to j}^{s}}{\sum_{s=1}^{S}\sum_{j=1}^{J} L_{i\to j}^{s}}, \tag{4}$$

where C_{im}^{adp} , $\boldsymbol{\mu}_{im}^{adp}$ $\boldsymbol{\Sigma}_{im}^{adp}$, a_{ij}^{adp} are the updated mixture, mean weight, covariance matrix and updated transition probability respectively. $L_{i\to j}^{s}$ is the accumulated probability of the state occupancy from state i to state j and S denotes the number of selected speakers. The construction process is facilitated by a model selection which will be explained in later sections.

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 is consist of 200 utterances. 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

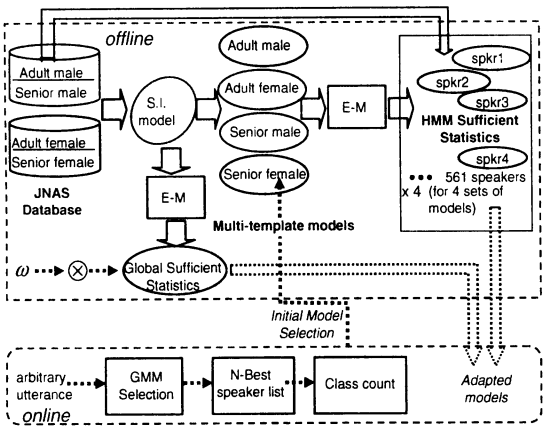

Fig. 2 Block diagram of HMM-Sufficient statistics multiple models adaptation.

(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]. Further reducing S results in a reduction of adaptation time with a trade-off of the recognition performance. This is attributed to the fact that further decreasing S results to insufficient data necessary to robustly estimate the target speaker's HMMs.

3. 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 :

$$C_{im}^{adp_{new}} = \frac{\sum_{s=1}^{S} L_{im}^{s} + \omega L_{im}^{global}}{\sum_{m=1}^{M}(\sum_{s=1}^{S} L_{im}^{s} + \omega L_{im}^{global})}, \quad (5)$$

$$\boldsymbol{\mu}_{im}^{adp_{new}} = \frac{\sum_{s=1}^{S} \boldsymbol{m}_{im}^{s} + \omega \boldsymbol{m}_{im}^{global}}{\sum_{s=1}^{S} L_{im}^{s} + \omega L_{im}^{global}}, \quad (6)$$

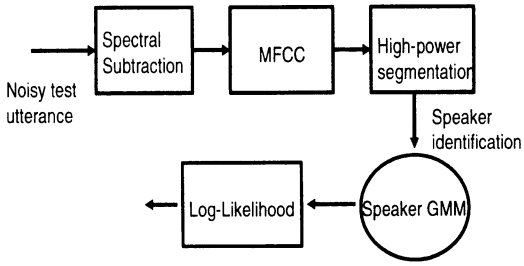

Fig. 3 GMM selection using the noisy test utterance.

$$\boldsymbol{\Sigma}_{im}^{adp_{new}} = \frac{\sum_{s=1}^{S} \boldsymbol{v}_{im}^{s} + \omega \boldsymbol{v}_{im}^{global}}{\sum_{s=1}^{S} L_{im}^{s} + \omega L_{im}^{global}} - \boldsymbol{\mu}_{im}^{adp} \boldsymbol{\mu}_{im}^{adp^T}, \quad (7)$$

$$a_{ij}^{adp_{new}} = \frac{\sum_{s=1}^{S} L_{i\to j}^{s} + \omega L_{i\to j}^{global}}{\sum_{j=1}^{J}(\sum_{s=1}^{S} L_{i\to j}^{s} + \omega L_{i\to j}^{global})}, \quad (8)$$

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 :

$$\omega = \tau_1, \quad (9)$$

$$\omega = \frac{\tau_2}{\tau_2 + L_{i\to j}^{global}}, \quad (10)$$

where in eqn (9) we used a multiplying constant τ_1 and in eqn (10), the weighting factor ω is normalized by the accumulated probability of the state occupancy, $L_{i\to j}^{global}$.

3.1 Speaker and Template Selection

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

1) The arbitrary noisy test utterance is denoised as shown in Figure 3 using Spectral Subtraction (SS) and then parameterized (MFCC). To reduce the effects of the residual noise that is present in the silence or unvoiced

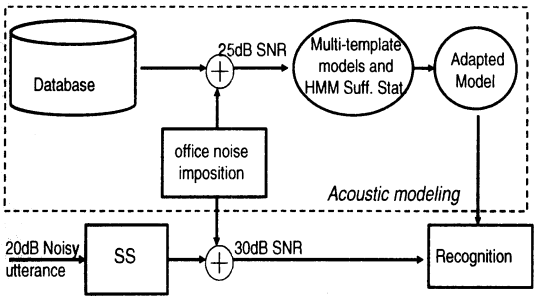

Fig. 4 Block diagram of the overall system implementation.

region of the speech utterance, the low power parts are removed and only the MFCCs that have high energy are retained for speaker selection.

2) We find the log-likelihood scores given the arbitrary test utterance and the GMM speaker dependent models. This process returns a list of log-likelihood scores among all 561 speaker-dependent GMMs from JNAS adult and senior database.

3) From the log-likelihood scores, 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.

4) From the N-speakers list, a class count is performed for the 4 different templates (Adult male, Adult female, Senior male, Senior female). The class counting is carried out using the speaker labels (that are present in the speaker IDs). Each of the speakers in the list will be calssified into 4 classes mentioned above.

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

5) Template model, N-best HMM-Sufficient Statistics and the weighted global HMM-Sufficients Statistics are prepared 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. Figure 4 shows the overall block diagram of the system. 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].

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

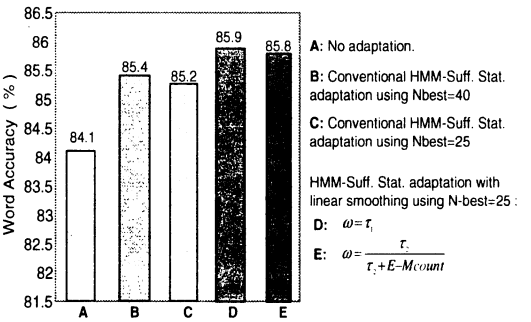

Fig. 5 Average recognition performance under four noisy environment conditions

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. A summary of the basic experimental condition parameters used in this set-up is provided in Table 1. In the case of the number of selected speakers S used in adapting the model parameters in equations (2)-(4), we found $S = 40$ which is the optimal value $S_{optimal}$ of the N-best which is sufficient to construct a robust model from the Sufficient Statistics. Weighting factors given in equations (9)-(10) achieved best results when $0 < \tau_1 < 0.2$ and $1 \leqq \tau_2 \leqq 2$. In particular we used $\tau_1 = 0.015$ and $\tau_2 = 2$.

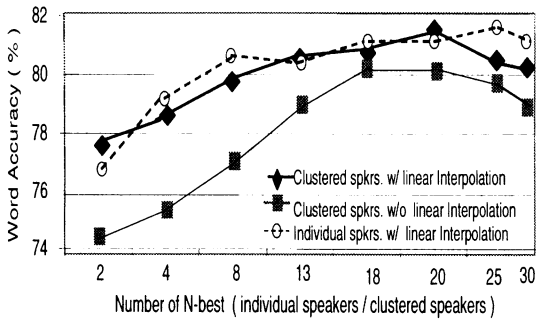

Fig. 6 Performance of the clustered speakers' HMM-Sufficient Statistics adaptation with linear interpolation.

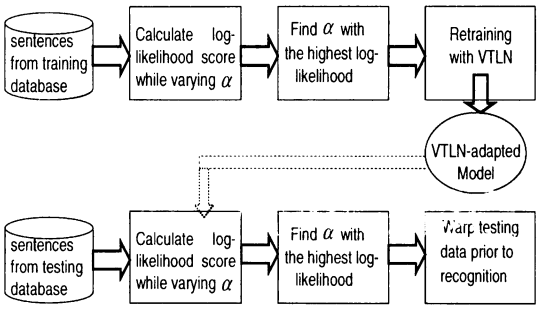

Fig. 7 Block diagram of the supervised VTLN adaptation in finding for the optimum α.

4.1 General Results

In Figure 5, the word accuracy (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 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 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.

4.2 Clustered speakers' HMM-Sufficient Statistics

We extended the proposed adaptation method by clustering the speakers in the database shown in. 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 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 6 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. 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 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.

4.3 Recognition Results Using VTLN, MAP and MLLR

We implemented VTLN to normalize the effect of vocal tract's size of the the different speaker classess. Figure 7 shows the set-up of implementing VTLN adaptation. First, we search for the optimum α that would maximize the log-likelihood score of the training database. This particular value of α is then used to adapt the model. The process is repeated in the case of the testing database utterances but this time using the VTLN-adapted model. Consequently an optimal α is

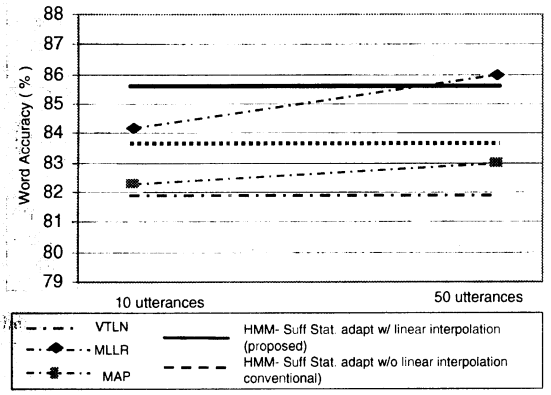

Fig. 8 Recognition Performance of the proposed method compared to VTLN, MAP and MLLR.

found that maximizes the log-likehood score of the testing utterances given the VTLN-adapted model. This α is then used to check for the recognition performance of the test data using VTLN adaptation, which is to be compared to the proposed method. The fact that α is optimized in both the training and testing data which is used in evaluating the recognition performance serves as the upperlimit of the VTLN.

Figure 8 is the result for MAP, MLLR and VTLN experiments. In the abscissa, the labels 10 and 50 utterances correspond to the adaptation data for MAP and MLLR. It is apparent that the proposed HMM-Sufficient Statistics adaptation which uses only one arbitrary utterance outperforms MAP and MLLR (10 adaptation utterances). However, when the adaptation utterance for MLLR is increased to 50 then it slightly peforms better than the proposed method. It is also very clear that the proposed method is better than the VTLN adaptation. Lastly, there is a significant improvement when using the proposed linear interpolation in HMM-Sufficient Statistics adadaptation as compared the conventional approach without linear adaptation.

5. Conclusion

We have succesfully reduced the adaptation time from 10 sec to 5 sec with linear interpolation of the global HMM-Sufficient Statistics as shown in Figure 9. The reduction in adaptation time is achieved without degrading the recognition performance. In the future, we will experiment on different noisy environment conditions and different SNRs.

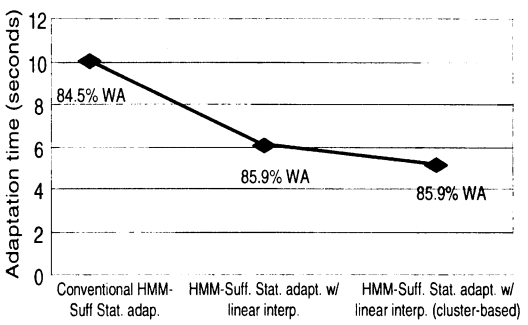

Fig. 9 Summary of adaptation time reduction.

6. Acknowledgment

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

Reference

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

[2] C. Huang, T. Chen, S. Li, JL. Zhou, "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,G. Evermann,D. Kershaw,G. Moore,J. Odell,D. Ollason,D. Povey,V. Valtchev,P. Woodland , "The HTK Book"

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

[7] S. Yoshizawa, A. Baba, K. Matsunami, Y. Mera, M. Yamada, K. Shikano, "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, K. Matsunami, A. Baba, A. Lee, H. Saruwatari, K. Shikano , "Spectral Subtraction In Noisy Environments

[12] R. Gomez, et al., "Speaker-Class Reduction for HMM-Sufficient Statistics Adaptation Using Multiple Acoustic Models