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

2004－SLP－54（35）
2004／12／21

雑音環境における複数モデルを用いた十分統計量に基づく教師なし話者適応

Randy Gomez† 李　晃伸† 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@naist.ac.jp

あらまし　音声認識において，話者ごとに異なる話者の声の特性を考慮して，音韻モデルの話者適応の研究が行われている．一方で，性別や年齢層などの話者クラスごとに学習したクラス依存音韻モデルを用いることで，不特定話者モデルよりも認識精度は向上する．本研究では，多様な音声データベースが整備されつつある現状を背景に，HMM 十分統計量に基づく教師なし話者適応を複数のデータベースおよび複数の初期モデルに拡張する．従来法では単一の不特定話者モデルから適応を行っていたが，提案手法では年齢層や性別などの複数のクラス依存音韻モデルを元に適応を行うことで初期モデルの改善を図る．まず，入力音声に対して GMM から最も音響的特徴の近い話者集合を抽出する．その際に，そのリスト中の近傍話者の属するクラスから，入力音声に最も近いクラス依存音韻モデルを選択する．その後，それを基準モデルとして，そのクラスに対応する近傍話者の十分統計量から音韻モデルを再構築する．JNAS 成人および高齢者のデータベースを用い，オフィス・人混み・展示会場ブース・車室内の各雑音環境において評価を行ったところ，従来手法に比べて精度が向上することが確かめられた．さらに，教師あり適応の MLLR 法と比較したところ，10 文章による教師あり適応よりもよい精度が得られることが示された．

キーワード　教師なし話者適応、HMM 十分統計量、クラス依存音韻モデル

Unsupervised Speaker Adaptation Based on HMM Sufficient Statistics Using Multiple Acoustic Models Under Noisy Environment

Randy GOMEZ†, Akinobu LEE†, Hiroshi SARUWATARI†, and Kiyohiro SHIKANO†

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

Abstract　Speaker adaptation in speech recognition is necessary to achieve a high accuracy for wide varieties of speakers. On the other hand, using class-dependent (CD) acoustic model for specific gender/age class can result to a better accuracy than a single speaker-independent (SI) model. In this research, we extend the unsupervised speaker adaptation based on HMM Sufficient Statistics (HMM-SS) for multiple database and multiple initial models, given a wide varieties of speech database. As opposed to the conventional approach which utilizes only a single SI model as a base model, the proposed method makes use of multiple CD models to push up the performance of initial model before adaptation. A speaker's class is estimated from the N-best neighbor speakers by Gaussian Mixture Models (GMM) on the way of speaker selection, and the corresponding CD model is adopted as a base model. Then, the unsupervised speaker adaptation is performed by constructing HMM from HMM-SS of the selected speakers. Experiments were carried out on two database namely, adults and senior people by JNAS, and we performed testing under noisy environment conditions such as office, crowd, booth and car noise with 20dB SNR. Recognition results show that the proposed method based on multiple model outperforms the conventional approach. Moreover, comparison with the Maximum Likelihood Linear Regression (MLLR) adaptation with 10 supervised utterance confirms that our method perfroms better with only a single utterance input.

Key words　Unsupervised Adaptation, Noise Robustness, HMM Sufficient Statistics

1. Introduction

Model mismatch causes a problem of speaker variability in speech recognition which degrades the performance of the recognizer [1]. Mismatch also arises under noisy environment conditions where noise contributes to the degradation of the recognition performance. There exist various speakers in a database and different types of noise. Models are trained using this collected database and noise conditions which do not often match with the actual testdata [2].

It is neccessary to employ adaptation technique to match the model with various speakers. Existing adaptation technique available like the MLLR [3] can considerably improve the recognition rate as opposed to using only an SI model. Speaker adaptation based on HMM Sufficient Statistics HMM-SS is an unsupervised speaker adaptation that requires an arbitrary utterance from the speaker and makes use of the pre-generated Speaker-Dependent (SD) HMM-SS [4]. Although existing speaker adaptation technique such as MLLR performs better than HMM-SS, the latter is more preferrable for practical application since it only requires one utterance to carry out the adaptation process. Adaptation based on HMM-SS has been applied in wide speaker varieties of database where two database are involved namely, JNAS adult (260 speakers) and SENIOR database (301 speakers)

In this paper we extended the conventional HMM-SS method using single SI model to using multiple acoustic models unsupervised speaker HMM-SS adaptation. Furthermore, we carried out the experiments under noisy environment conditions using office, crowd, booth and car noise. A conventional HMM-SS approach from one acoustic model is applied to combine these two database. However, this implementation does not take into consideration the speaker variabilities both in the contexts of age and gender. The very wide varieties of speakers are very difficult to represent by just a single SI model. In short, one model is not enough to capture the statistics which is unique to a class of speakers. In order to reduce the effects of these problems, we extend the unsupervised HMM-SS to using CD multiple acoustic models to cover both age and gender variabilities. In this work we created a set of CD multiple acoustic models based on gender

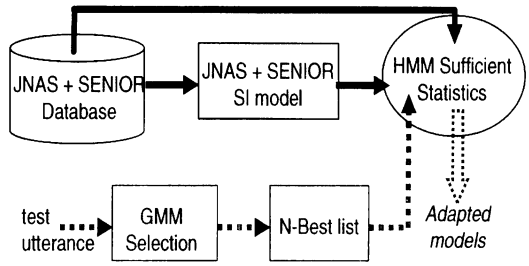

Fig. 1 Conventional HMM-SS speaker adaptation

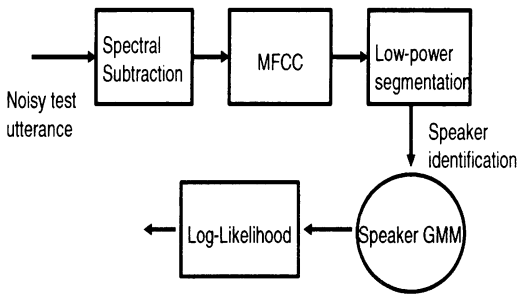

Fig. 2 GMM selection using the noisy test utterance

and age. The proposed method is evaluated through speech recognition experiments under 20dB noisy environment (office, crowd, booth and car) and showed improvements compared to the conventional HMM-SS method from one acoustic model.

In the later part of this paper, we will be showing the detail description of our approach and its recognition performance. Also, we will discuss our current work which aims to minimize adaptation time in implementing the multiple acoustic models HMM-SS method.

2. System Implementation

In this section, the difference between the conventional and the proposed HMM-SS speaker adaptation will be discussed when using a huge database that consists of 60K-utterance from 301 male and female speakers (SENIOR database) and 52K-utterance from 260 male and female speakers (JNAS database) [5].

2.1 Conventional HMM-SS Method

Figure 1 shows a conventional method which is a trivial approach wherein a single SI model is created from the unified database namely JNAS (J) which consists of adult speakers and SENIOR (S) which consists of elderly speakers. HMM-SS is created for every speaker in the unified database. HMM-SS contains model information such as means, variance and EM counts. In the actual adaptation, N-best speakers close to the test

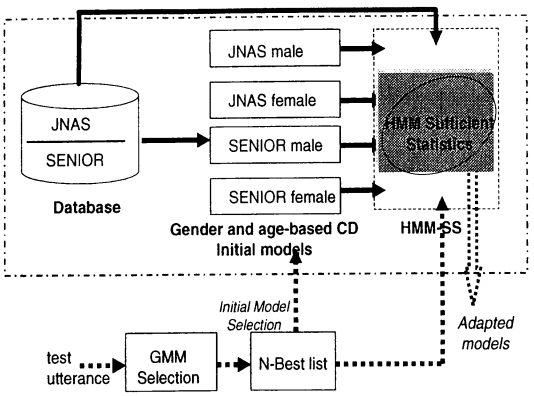

Fig. 3 Proposed Method : HMM-SS speaker adaptation using gender and age-based multiple acoustic models

utterance are selected for HMM-SS adaptation based on speakers' Gaussian Mixture Model (GMM).

Each speaker in the database has a corresponding phoneme-independent GMM and trained from clean data. The actual GMM selection is shown in Figure 2 where the noisy test utterance is being denoised using SS and then parameterized (MFCC). In order to minimize the effect of the residual noise that is present in the silenced or unvoiced region of the speech utterance, we remove the low power part and retain only the MFCCs that have high energy. This however does not have any detrimental effects in the speaker selection process since the GMM are phoneme-independent. Consequently, we generate log-likelihood among the speakers given the the parameterized test utterance.

In carrying the actual adaptation of the conventional method, the system does not have the option to choose any initial model in creating the HMM-SS as there is only one initial model.

2.2 Proposed HMM-SS Method

In our proposed method, we take advantage of using the CD models in creating different classess of HMM-SS to have a better discrimination among speaker's acoustical characteristics. By giving the system a more degree of freedom in choosing the approriate HMM-SS that is close to the test utterance, speaker variability will be reduced.

2.2.1 Gender and Age-based Multiple Acoustic Models HMM-SS

In Figure 3, four different CD (gender and age-based) HMM models are created namely: JNAS Male (JM), JNAS Female (JF), SENIOR Male (SM) and SENIOR Female (SF). Note that the CD models in this case are both age and gender dependent. Consequently, four classes of HMM-SS for each speaker are created which correspond to these CD models. The actual HMM-SS adaptation procedure are as follows:

1) N-best speakers close to the test utterance are selected using the GMM speaker dependent models. This process gives us a list of log-likelihood among all the speakers in the GMM model as can be seen in Figure 2

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

3)From the N-best speaker list we count the number of class hits for the 4 different classes from the speaker labels. It should be noted that since the N-best list contains the speaker ID then it would be easy to check for the particular class it belongs.

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

5)The system will go through the pre-trained initial models and HMM-SS and select the appropriate HMM-SS that correspond to the selected initial model for adaptation.

3. Experimental Results

The performance using the conventional HMM-SS adaptation is compared with the proposed multiple acoustic models in creating HMM-SS (gender and age-based) under 20dB noisy environment. In addition to that, speaker adaptation using MLLR trained with 10 and 50 utterrances is also compared. Results of class counting using N-best speakers close to the test utterance for selecting the class of HMM-SS to be used is also discussed

3.1 Experimental Conditions

The language model is provided by the IPA dictation toolkit [6]. Phonetically Tied Models (PTM) [7] is used which is modelled by superimposing 25dB office noise to the database in creating the CD models [8]. Figure 4 shows the overall block diagram of the system. In the acoustic modelling part office noise is superimposed to the clean speech from the database that results to

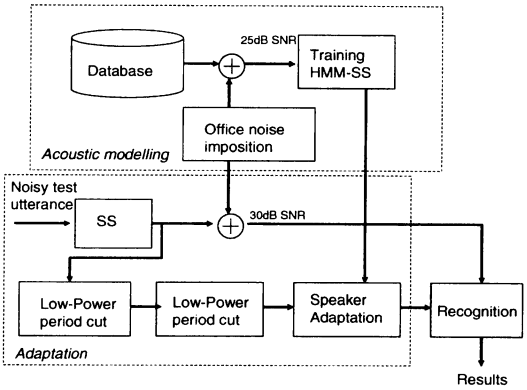

Fig. 4 Training of acoustic model

Table 1 Conditions used in acoustic modelling

Model	PTM 2000 tied-states, 64Mix
Sampling Freq/ADC	16kHz/16bit
Window Shift	10msec
Parameters	MFCC(12 dim), Δ MFCC, Δ Power

Table 2 Actual number of PTM models

	Conventional	Proposed
Models	1 Speaker-Independent	4 Class-Dependent
HMM-SS	1 set	4 sets

25dB SNR which is used in training. In the adaptation part, A noisy test utterance is denoised with SS which is used for speaker adaptation (see GMM selection). Lastly, for the actual recognition test, the SS-denoised test utterances are superimposed with 30dB office noise prior to recognition. The significance of using a single office noise acoustic model of this type instead of noise-matched models is the fact that in real application there exist so many types of noise and it would be impractical to create a matched model for each of these noise. The noise robust speech recognition algorithm with SS and noise superimposition of office noise 30dB is robust enough against various noise conditions [8]. With this, we will be able to 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. A summary of the basic experimental conditions used is given in Tables 1 and 2.

The testsets are grouped into four classes according to database (JNAS or SENIOR) and according to gender (male or female) namely JM, JF, SM and SF. Each testset is of 100 utterances from 23 speakers which are taken outside from the training data. Office, crowd, booth and car noise are superimposed which results to 20dB SNR respectively. Figure 5 shows the spectral plot of the four types of noise that are superimposed in the test data. 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 [8]. The 30dB office noise-superimposed denoised utterance is tested for recognition performance using JULIUS with 20K-word on Japanese newspaper dictation task from JNAS and SENIOR.

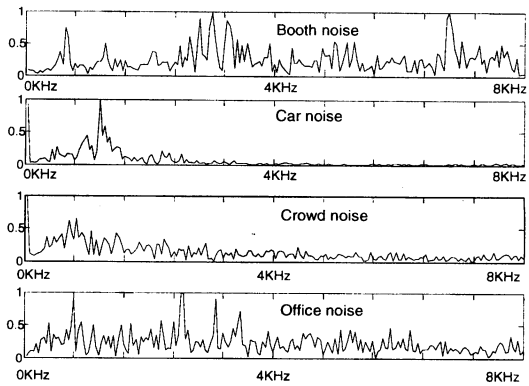

Fig. 5 Spectrum of the superimposed noise

3.2 Recognition Results

Figure 6 is an average result over all noise conditions. This graph shows the recognition performance using single model SI with and without HMM-SS adaptation and using the proposed method multiple acoustic models with HMM-SS adaptation. Figure 7 shows the detailed recognition performance on all noise conditions of the proposed method using gender and age-based CD (multiple models HMM-SS adaptation). It is apparent that it outperforms the conventional method in all noisy conditions and in all different classes of testsets under 20dB. We also tested the cluster-based approach and the recognition performance improved considerably compared to the conventional approach in the same trend as the gender and age-based approach.

The selection of what particular class of HMM-SS to be used in the online adaptation given a test utterance

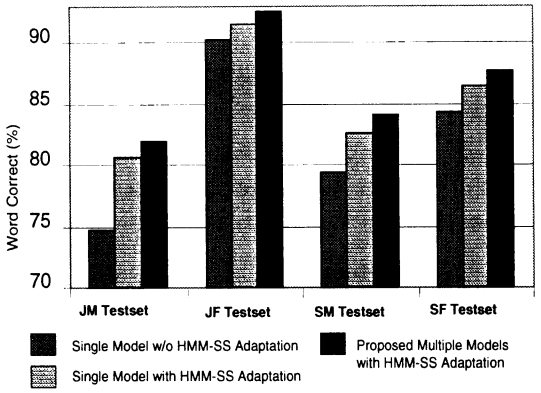

Fig. 6 Comparison of the averaged recognition performance over different noise conditions

Table 3 Class counting using the N-best speakers for the proposed gender and age-based method

Input Testset	JMclass	JFclass	SMclass	SFclass
Jmale	82%	0%	18%	0%
Jfemale	2%	76%	0%	22%
Smale	20%	0%	80%	0%
female	0%	0%	0%	100%

is dependent on the N-best label count as outlined in step 2 of section 2.3. Table 1 shows an actual data of a class count during testing. It can be noted that there are minor misclassifications in identifying the classes but these do not affect the performance of the proposed method as can be pointed to the fact that in our separate experiment using cluster-based approach which also employs multiple acoustic models in creating HMM-SS is having a consistent result in improving the recognition rate.

3.3 Recognition Results Using MLLR

MLLR results for the gender and age-based multiple acoustic models are given in Figure 8 using 10 and 50 utterances respectively. The initial models used for MLLR are the gender and age-based CD initial models. This result is the average of all the different noisy conditions. It should be noted that in both 10 and 50 utterances used for supervised MLLR adaptation, the proposed multiple acoustic models HMM-SS unsupervised adaptation only needs 1 arbitrary utterance.

In Figure 8, it is shown that the proposed unsupervised method works better than that of the supervised MLLR when using 10 utterances for training. Braces in the graph shows the difference between the

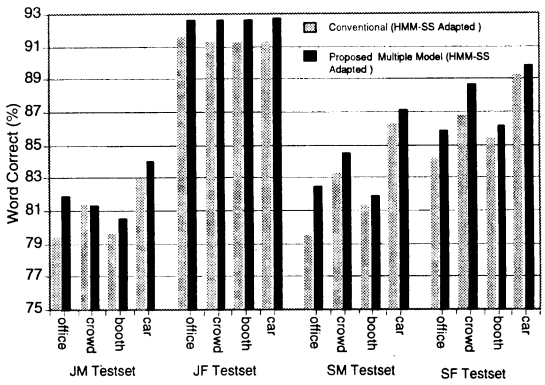

Fig. 7 Recognition performance of the proposed multiple models HMM-SS adaptation

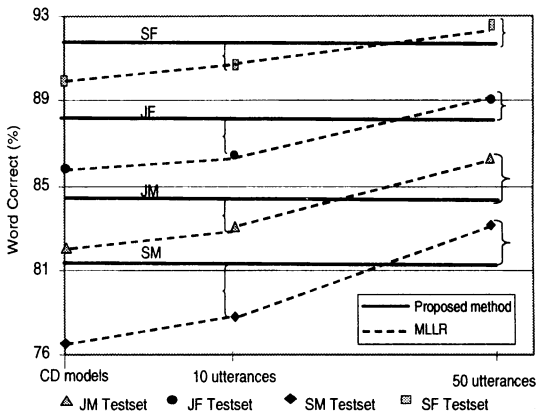

Fig. 8 Comparison of recognition performance using MLLR for the gender and age-based approach and the proposed method

proposed multiple acoustic model HMM-SS adaptation and MLLR. The recognition rate using the CD models that are gender and age-dependent of a particular class without adaptation is obviously having the least recognition rate as compared to the proposed method and MLLR.

4. Conclusion

We successfully extended the conventional unsupervised HMM-SS speaker adaptation using a large database into using multiple acoustic models HMM-SS unsupervised adaptation. The proposed method considerably improves the performance of the system. Also, the proposed method proves to be noise-robust and works consistently under various kinds of noise conditions. This result suggests that by using multiple

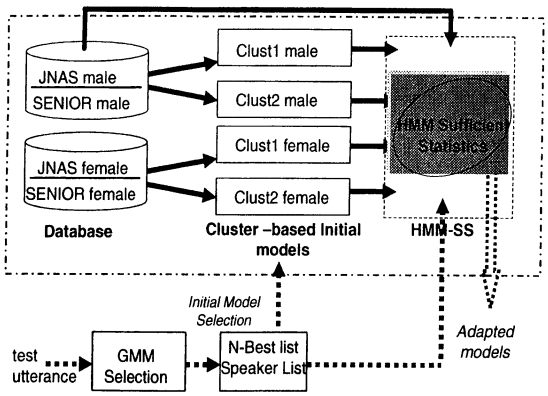

Fig. 9 HMM-SS speaker adaptation using cluster-based multiple acoustic models

initial models (CD initial models) that contain the statistical information such as age and gender of a particluar class work well than using a single initial model that has no distinction between age and gender in the implementation of HMM-SS speaker adaptation. This is true, especially when dealing with huge database.

5. Current Work

In this research we have also performed multiple models adaptation using clustering where the initial models and HMM-SS are cluster-based in creating the CD models. K-means clustering technique is used in creating two different clusters in each gender. Four different CD (cluster-based) models are created as follows, Clust1 male and Clust2 male for the combined male speakers. In the case of female, we have Clust1 female and Clust2 female, respectively. The HMM models created in cluster-based are dependent on both cluster and gender but independent of age. The actual HMM-SS adaptation for the cluster-based approach is the same as that of the gender and age-based previously outlined.

Although, the results we have achieved are consistent to those of the proposed method, the gender and age-based approach, it would be worthy to do more experiments and further investigate the benefits of the cluster-based HMM-SS unsupervised apeaker adaptation shown in Figure 9. One interesting benefit could be the reduction of calculation time in performing the actual HMM-SS. At the moment we use 40 speakers in the adaptation process using the 561 speakers both in the adult and senior database. The intuition is that, if we could reduce these number of speakers through clustering then we might be able to reduce the N-best speakers for adaptation thus reducing the computation time.

In carrying out this, we should experiment different types of clustering techniques and distance measures that would be approriate to HMM-SS adaptation.

6. Acknowledgment

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

Reference

[1] Akira Baba, S. Yoshizawa, M. Yamada, A. Lee, K. Shikano "Elderly Acoustic Model for Large Vocabulary Continuous Speech Recognition" *In Proceedings of EUROSPEECH*, pp. 1657-1660, 2001.

[2] R. Gomez, et al. " Robust Speech Recognition with Spectral Subtraction in Low SNR" *In Proceedings of International Conference on Speech and Language Processing ICSLP*, 2003.

[3] 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

[4] S. Yoshizawa, A. Baba, Y. Mera, M. Yamada, A. Lee, K. Shikano " Evaluation on Unsupervised Speaker Adaptation Based on Sufficient HMM Statistics of Selected Speakers" *In Proceedings of International Conference on Acoustics, Speech and Signal Processing ICASSP*, 2001.

[5] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, S.Itahashi "JNAS: Japanese Speech Corpus for Large Vocabulary Continuous Speech Recognition Research", The Journal of the Acoustical Society of Japan (E), Vol.20, pp.199-206, 1999

[6] T. Kawahara et al, "'Free Software Toolkit for Japanese Large Vocabulary Continuous Speech Recognition', *In Proceedings of International Conference on Speech and Language Processing ICSLP*, pp. IV-476-479, 2000.

[7] A. Lee, T.Kawahara, K. Takeda, K. Shikano "A New Phonetic Tied Mixture Model For Efficient Decoding", *In Proceedings of International Conference on Acoustics, Speech and Signal Processing ICASSP*, pp. 1269–1272, 2000.

[8] Y. Shingo, K. Matsunami, A. Baba, A. Lee, H. Saruwatari, K. Shikano "Spectral Subtraction In Noisy Environments Applied To Speaker Adaptation Based on HMM-SS" *In Proceedings of International Conference on Speech and Language Processing ICSLP*, pp. I-1045-1048, 2000.