音声言語情報処理 42－7
（2002．7．13）

自発発話音声認識のための話者内変動を考慮した2段階MLLR適応

李 宝潔[1]　　広瀬啓吉[2]　　峯松信明[3]
[1]東京大学大学院・工学系研究科　[2]東京大学大学院・新領域創成科学研究科
[3]東京大学大学院・情報理工研究科

MLLRに基づく音響モデルの適応は、限られた適応データしか得られない場合でも有効な適応手法である。しかしながら、認識対象の音響空間の様子が、もとのモデル構築に利用したデータの音響空間のそれと大きく異なる場合、思ったような適応の効果が得られなくなる。このような状況は自発発話で起こると考えられる。自発発話では、読み上げと比べ、話者内変動が大きくなると考えられるからである。この様な大きな話者内変動に対処する手法として、2段階MLLR手法を提案した。この方法では、（読み上げ音声に対する）不特定話者モデルを、まず特定話者に適応して特定話者モデルを作成する。次に、特定話者の学習データを数個のカテゴリに分けた後、各カテゴリに対して、特定話者モデルを再適応し、特定カテゴリモデルを構築する。認識は、特定話者モデルと各特定カテゴリモデルを用いて個別かつ並列に行い、最大尤度を与えるものを最終的な認識結果とする。種々の感情の音声を対象とした認識実験行った結果、提案手法によって、従来のMLLRに基づく話者適応を超える性能が得られることが示された。

Two-level MLLR Adaptation for Intra-speaker Variation in Spontaneous Speech Recognition

Baojie LI[†]　　Keikichi HIROSE[‡]　　Nobuaki MINEMATSU[†]
[†] School of Engineering, University of Tokyo
[‡] School of Frontier Sciences, University of Tokyo

Acoustic model adaptation based on MLLR works well even when only a limited number of adaptation data are available. However, when the acoustic space of speech to be recognized is quite different from that of speech used for the training of the original models, its effect comes limited. This situation may occur when spontaneous speech with wide variation is dealt with. Intra-speaker variation comes larger in spontaneous speech than in read speech. In order to cope with this large intra-speaker variation, two-level MLLR adaptation method was proposed. A speaker independent model (for read speech) is first adapted to a specific speaker to generate a speaker dependent model. Then, after classifying the training data into several categories, the speaker dependent model is further adapted to each category using data classified to it (category dependent model). The recognition is done in parallel using the speaker dependent model and each category dependent model, and the result with the highest likelihood is selected as the final result. Recognition experiments conducted on speech with various emotions showed that the proposed method outperformed the conventional MLLR-based speaker adaptation.

1 Introduction

Various mismatches between the training and testing conditions considerably degrade the performance of speech recognition. Among them, inter-speaker and intra-speaker variations of speech are considered to be crucial. In order to cope with inter-speaker variation, speaker adaptation approaches have been broadly investigated, and have achieved a remarkable success even with a small amount of data from the speaker to be recognized. In general, these approaches can be divided into three types: maximum a posteriori probability (MAP) estimation[1], maximum likelihood linear regression (MLLR) estimation[2], and speaker clustering[3]. All of these adaptation ap-

proaches are designed to cope with speaker or environmental variations, and work well for read speech, whose features do not vary a lot in a speaker. However, in spontaneous speech or in other styles of speech, the intra-speaker variation of acoustic features comes larger, degrading the performance of speech recognition a lot. For example, Japanese speech recognition engine JULIUS, widely used for read speech with favorable results, can only achieve 65.64% of recognition accuracy for ATR/APP conversational speech even with well-trained CSRC-SI models (tri-phone models with 2,000 states, trained by 169,348 utterances from 4130 speakers)[4]. Even after the adaptation with sufficient data, acoustic models originally trained for read speech still perform poorly for conversational speech. The major reason is the wide variety of acoustic features in the conversational speech, which is considered to be smaller in the read speech.

To improve recognition performance for speech other than calmly read one, the issue of intra-speaker variation should be addressed. In the current paper, a two-level adaptation method is proposed to deal with both the inter- and intra-speaker variations in speech sounds. In the method, the original SI model trained for a large corpus is first adapted to a speaker by MLLR to obtain the SD model. (Strictly speaking, the obtained model should be called "SD-like" model). Then, the speech data of the speaker used for the adaptation are clustered into several categories according to their acoustic characteristics. The SD model is further adapted to each category also by MLLR to obtain category dependent (CD) models. The set of CD models is considered to have a good matching with the categories, and, therefore, to be able to deal with intra-speaker variations. Finally, the recognition process is run in parallel for each of CD models and the recognition result with maximum likelihood is selected as final output of the recognizer.

The advantage of the proposed method over the conventional speaker adaptation methods may be clearer for the speech data with larger variations. From this viewpoint, we selected speech uttered with several emotions for the experiments, which surely has wide variations. The following part of the paper is constructed as follows: Section 2 explains the proposed two-level adaptation method in detail. After checking that the CD models have some effects on recognition improvements in section 3, the method is evaluated through recognition experiments by comparing with the conventional MLLR-based speaker adaptation method in section 4. Section 5 concludes the paper.

2 Two-Level Adaptation

2.1 Objective and motivation

When the intra-speaker variation comes large as in the case of conversational speech or emotional speech, speech recognition systems cannot achieve a satisfactory performance only by a conventional speaker adaptation process.

To solve this problem, we cluster the speech data of a speaker into several categories according to their acoustic characteristics. The variation within a category will be much smaller.

An utterance to be recognized is regarded as belonging to one of these categories. If we can construct a category dependent (CD) model for this category by adaptation, then we will do a more accurate recognition using this CD model, than using the SD model. This is the idea of our two-level adaptation.

Since the factors causing intra-speaker variation such as emotion and speaking rate, may change largely across utterances. In recognition, we have to deal with each utterance individually, assigning a suitable CD model to each individual utterance according to its acoustic characteristic, to match it accurately.

2.2 Adaptation strategy

Given an SI model and adaptation data, the SI model is first adapted to the speaker of the data

to generate an SD model. This step is the conventional speaker adaptation and aims to alleviate the inter-speaker variability. Then, the SD model is further adapted to alleviate the intra-speaker variability. As for this process, unsupervised online adaptation will be a candidate. It can be conducted as follows: (1) Recognize an utterance using the SD model. (2) Recognized words with high confidence are selected and used to adapt the SD model. Adaptation is done so that the adapted model best-fits to the acoustic features of the utterance. (3) Re-recognize the utterance using the adapted model to have more reliable recognition result. However, we can obtain only a few words from one utterance for adaptation. And because different utterances may belong to different categories, the words from other utterances will not be suitable for adapting the model for this utterance. Moreover, it is not guaranteed that the selected words are correctly recognized in the first step. In this occasion, obviously a reliable adaptation cannot be done.

To avoid this unfavorable situation, we developed the following adaptation strategy: (1) Cluster the whole adaptation data into several categories based on their acoustic characteristics. (2) Adapt the SD model to each category to obtain CD models using the data of each category.

However, in recognition stage, we do not know which category the input utterance belongs to, and consequently we cannot assign a proper CD model to it. Hence we do the recognition in parallel using all the CD models and the SD model. So we will get several recognition candidates and the one with highest likelihood score will be selected as the final recognition result.

Figure 1 gives a block diagram for this process.

2.3 Emotional speech

The validity of the proposed method may come clearer when the intra-speaker variation is larger and the recognition performance by the conventional method is poorer. In the current paper, we selected emotional speech as the data for the experiment. Although many features were proposed to describe or cluster emotional speech[6], there may be other candidates depending on the data, which we are handling with. As shown in the next section, in the current experiment, we classified the data according to the emotion labels (anger, delight, etc.) attached to the data. Since we are dealing with the intra-speaker variation due to emotions, the current adaptation shall be called *emotion adaptation.*

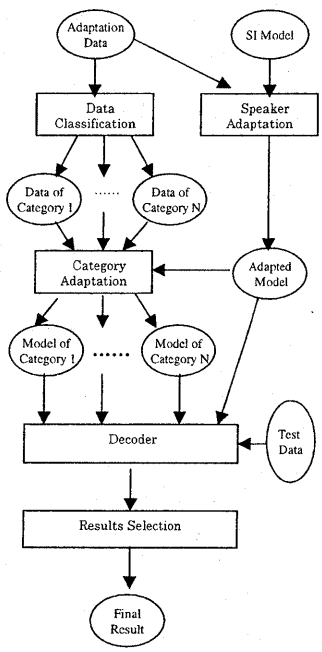

図 1: Blockdiam of the proposed 2-level adaptation

3 Preliminary experiments

The basic assumption of our proposed method is that the model of one category will perform better when tested on data that belong to this category, than when tested on data that belong to the other categories. For example, a model adapted with *angry* data will performs better on *angry* data, but worse on *sad, delighted* or any

other emotional data. It will be demonstrated in this section.

We exploit four types of emotions in our experiments, named *anger, delight, sympathy,* and *sadness.* Additionally *neutral* is also regarded as one type of emotion. All the experiments are conducted on these five types of emotions. We represent the emotions as $\{E_i\}$, where $i = 1, \ldots, 5$.

First we prepare two data sets for each type of emotion E_k: one is for adaptation, named D_a^k and the other for testing, named D_t^k . Then we adapt the SD model with D_a^k, to generate a new emotion dependent (ED) model M^k for emotion E_k.

Model M^k is then used to conduct recognition on all the five sets of testing data $\{D_t^i\}$, where $i = 1, \ldots, 5$. Then we obtain the recognition rates $\{R_t^i\}$ (*word accuracy*) for $\{D_t^i\}$. If we have $R_t^k > \{R_t^i\}$, where $i = 1, \ldots, 5$ and $i \neq k$, then our assumption is valid.

3.1 Description of test conditions

The emotional data are recorded in our laboratory. Three lists of text, list 1, list 2 and list 3 are read by four male actors (named *M1, M2, M3* and *M4*. Both list 1 and list 2 consist of 8 sentences, and each sentence is read once with each of the five types of emotion . The utterance set is called D_1). List 2 is read twice (called D_{21} and D_{22} respectively). List 3 consists of 26 sentences. It is designed as a dialogue. Two actors read it alternatively with emotions that change with the contexts (called D_3). So we have four data sets for each speaker: $\{D_1^i\}$, $\{D_{21}^i\}$, $\{D_{22}^i\}$ and D_3, where $i = 1, \ldots, 5$, represent five types of emotion.

We used mono-phone models as the SI model. They are provided by Information-technology Promotion Agency, Japan and called IPA-SI models. Each state of a mono-phone HMM consists of 16 mixture components. They are trained with the *ASJ Continuous Speech Corpus for Research* and *Japanese newspaper article sentences,* totally 20k sentences uttered by 132 speakers. The parameter vector is 25-dimensional containing 12th order $MFCC$s, $\Delta MFCC$s and Δpower). The dictionary consists of 130 words, and no language model is used.

3.2 Experimental results

Utterances of speaker $M1$ were used to conduct the preliminary experiments. The experiments are run twice: At the first time, $\{D_1^i\}$ and $\{D_{21}^i\}$ are used as adaptation data, and $\{D_{22}^i\}$ are used for testing. At the second time, $\{D_1^i\}$ and $\{D_{22}^i\}$ are used as adaptation data, and $\{D_{21}^i\}$ are used for testing.

HEAdapt is used to conduct an MLLR adaptation and *HVite* is used as the recognizer. Both *HEAdapt* and *HVite* are tools provided by *HTK3.1*[5].

Figure 2 displays the recognition results of each emotion model tested on every emotion data set individually, where M_{neu}, M_{ang}, M_{del}, M_{sad} and M_{sym} represent the adapted models for emotions *neutral, anger, delight, sadness* and *sympathy* respectively. D_{neu}, D_{del}, D_{ang}, D_{sad} and D_{sym} represent the test data sets of emotions *neutral, anger, delight, sadness* and *sympathy* respectively.

As shown in the figure, each emotion model achieves the highest recognition rate on the data set that belongs to the same emotion. These results give us the basic support to our two-level adaptation approach.

4 Evaluations

4.1 Evaluations on emotional data

In a real recognition task, we have no idea about the emotion type of each utterance to be recognized. Therefore all the recognition tests in this section, are done without emotion labels assigned to testing utterances.

The experiments are conducted on four speakers, *M1, M2, M3,* and *M4* respectively, and are run twice: At the first time, $\{D_1^i\}$ and $\{D_{21}^i\}$ are used as adaptation data, to generate five ED mod-

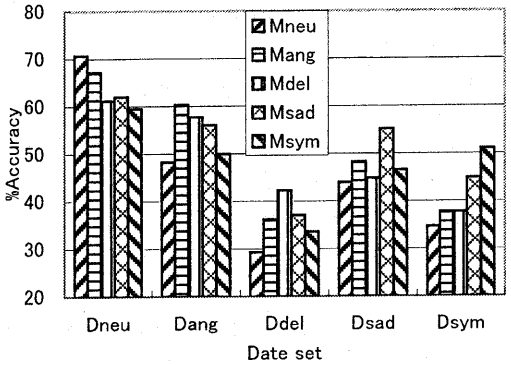

図 2: Recognition word accuray for emotional speech using emotion models

Models	Percentage of word accuracy			
	M1	M2	M3	M4
IPA-SI	42.79	49.67	54.31	46.94
MLLR	62.16	56.87	67.41	67.03
MLLR+ Emotion adaptation	68.11	61.09	69.83	66.81
Increased Percentage	5.95	4.22	3.59	1.39

表 1: Comparison of 2-level and MLLR adaptation method on emotional data

Models	Percentage of word accuracy			
	M1	M2	M3	M4
MLLR	41.08	44.86	61.08	42.7
MLLR+ Emotion adaptation	44.86	45.41	62.16	41.62
Increased Percentage	3.78	0.55	1.08	-1.08

表 2: Comparison of 2-level and MLLR adaptation method on conversational data

els for five types of emotion. The remaining five data sets $\{D_{22}^i\}$ are used for testing. Their emotion types are not presented to the recognizer.

At the second time, $\{D_1^i\}$ and $\{D_{22}^i\}$ are used as adaptation data, and $\{D_{21}^i\}$ are used for testing.

To generate CD models, the whole adaptation data should be clustered automatically into different categories. But in this paper, for simplicity, the data are simply divided into different emotion types according to the emotion labels attached to the data.

IPA model is used as the SI model. The SD model is generated by applying MLLR on it, using all the adaptation data (consisting of five types of emotion). Each ED model is generated by applying MLLR on the SD model, using only the adaptation data of that type of emotion.

The recognitions are done using the SD and the five ED models in parallel. The final result is selected from their outputs by the likelihood score of the whole utterance.

In Table 1, the row labeled "*MLLR + Emotion adaptation*" lists the recognition results of the above experiments for each of the four speakers. For comparison, the row labeled "*IPA-SI*" lists the results obtained using IPA-SI model. And the row labeled "*MLLR*" lists the results obtained using the SD model.

The results show increases in *word accuracies* after intra-speaker emotion adaptation. This demonstrated the effectiveness of the two-level adaptation in emotional speech recognition.

4.2 Evaluations on conversational data

We have evaluated the effectiveness of the proposed method on emotional data. But these data were intentionally uttered with distinct emotions. In real tasks, the emotion of speech data changes continuously rather than discretely. The models for intermediate emotions may be obtained by interpolation. But in this paper, we just simply conduct the experiments the same way as we did in the last subsection, on data set D_3. The results are listed in Table 2.

As expected, we got improvements in *word ac-*

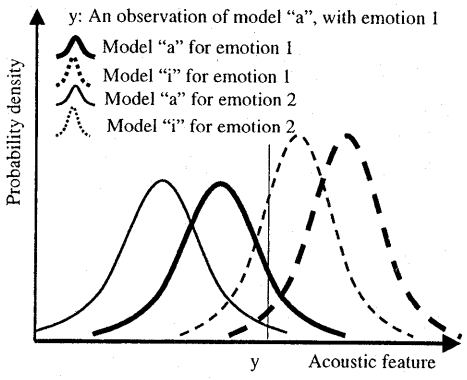

図 3: Image of the reason for mis-selection

curacies for speakers *M1*, *M2* and *M3*. But it came worse for speaker *M4*. One of the reasons is considered to lie in our criterion for selecting the final recognition result from five candidates. In the above experiments, we selected the one with the highest likelihood score as the final result. But in the cases like Figure 3, a mis-selection will be done. In Figure 3, assume the "true" model for observation y is model "a" of emotion 1 (denoted by M_a^{e1}). When recognizing y using two model sets of two emotions in parallel, M_a^{e1} and M_i^{e1} will be outputted by model sets of emotion 1 and emotion 2 respectively. And M_i^{e2} will be mis-selected due to its higher likelihood score.

Such error words will lead to a mis-selection when the recognition *word accuracy* of each model set is too low (in the above experiments, the *word accuracies* range from 40% to 60%). Further investigation about this assumption should be done. And the criterion for selection should also be reconsidered.

5 Conclusion

A method of two-level adaptation was proposed to cope with intra-speaker variation in speech recognition, which comes larger in speech other than the reading style, such as conversational speech, emotional speech and so on. This method performs an intra-speaker adaptation after the normal speaker adaptation.

Speech recognition was conducted for intentionally uttered emotional speech, and the result clearly indicated the advantage of the scheme over the conventional speaker adaptation.

For real conversational speech rich in emotions, a problem in the criterion for selecting the final result was observed. Further investigation is needed.

Additionally, in the above task, since the amount of adaptation data was too small, it was difficult to find out a criterion to classify them. Therefore the classification of the adaptation data was done simply according to the types of emotion attached to the data. When there are sufficient adaptation data, classification should be done using a clustering scheme.

参考文献

[1] C.H. Lee, C.H. Lin, B.H. Juang, "A study on speaker adaptation of the parameters of continuous density Hidden Markov Models", *IEEE Trans. on Speech and Audio Processing* , Vol. 39, No. 4 pp. 806-814, 1991

[2] C.J Leggetter, P. C. Woodland, "Maximum Likelihood Linear Regression for speaker adaptation of continuous density hidden Markov models", *Computer Speech and Language*, pp. 171-185,September 1995.

[3] L. Mathan, L. Miclet, "Speaker hierarchical clustering for improving speaker-independent HMM Word Recognition", *Proc. IEEE International Conference on Acoustics, Speech and Signal Process ing*, pp. 149-152, 1990

[4] *http://www.itakura.nuee.nagoya-u.ac.jp/ takeda/IPA*

[5] *http://htk.eng.cam.ac.uk/index.shtml* , 2000

[6] Marc SchroDer, "Emotional speech synthesis: A review", *Eurospeech*, pp. 561-564, 2001