ROBUST SPEECH RECOGNITION USING INTER-SPEAKER AND INTRA-SPEAKER ADAPTATION

Baojie Li, Keikichi Hirose and Nobuaki Minematsu

School of Engineering, University of Tokyo
Bunkyo-ku, Tokyo, 113-8656, Japan
{lbj, hirose, mine}@gavo.t.u-tokyo.ac.jp

ABSTRACT

Inter-speaker variation can be coped rather well in speech recognition by speaker adaptation techniques such as MLLR and MAP. However, when dealing with speech other than reading style, such as conversational speech, emotional speech and so on, current recognition systems cannot achieve a satisfactory performance even after speaker adaptation. In view of this situation, two-level adaptation method was newly proposed, where adaptation technique was applied in two levels to handle inter-speaker and intra-speaker variations. A speaker independent model 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 highest likelihood is selected as the final recognition result. Recognition experiments were conducted for speech with various emotions (emotion of input speech is unknown), and the results 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]. MAP method takes advantages of the prior knowledge on the original model (usually speaker independent (SI) model) parameters, and the model parameters show a good convergence to those of a speaker dependent model (SD model) when adaptation data size comes larger. While MLLR method obtains an SD model by applying a linear transformation to the original model parameters. The transformation is decided to have the maximum likelihood. The major advantage of MLLR method is that adaptation is even possible for phonemes without or with very limited adaptation data. Speaker clustering classifies speakers into several groups according to the similarity in acoustic features, and prepare a model for each group. For a new speaker, one of the models is selected. All of these adaptation approaches 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. (Although, strictly speaking, the obtained model is not the one trained directly using speech data of the speaker, and should be called "SD-like" model, it is called simply as SD model in this paper). 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: following to a brief explanation on MLLR in section 2, section 3 ex-

plains the proposed two-level adaptation method in detail. After checking that the CD models have some effects on recognition improvements in section 4, the method is evaluated through recognition experiments by comparing with the conventional MLLR-based speaker adaptation method. Section 6 concludes the paper.

2. BRIEF INTRODUCTION TO MLLR

In MLLR adaptation, when some adaptation data of a new speaker are given, the Gussian distribution mean $\boldsymbol{\mu}^{MLLR}$ for the new speaker can be obtained by applying a transformation to the mean $\boldsymbol{\mu}$ of SI model:

$$\boldsymbol{\mu}^{MLLR} = \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} = \boldsymbol{W}\boldsymbol{\xi}$$

where $\boldsymbol{A}$ is the transform matrix and $\boldsymbol{b}$ is the bias vector. $\boldsymbol{W} = [\boldsymbol{A}\ \boldsymbol{b}]$ is the extended transform matrix and $\boldsymbol{\xi} = [1\ \mu^T]^T$ is the extended mean vector. $\boldsymbol{W}$ is estimated by maximizing the likelihood of adaptation data, and is shared among components.

3. TWO-LEVEL ADAPTATION

3.1. 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 small.

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 mode 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.

3.2. 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, 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.

3.3. 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 the Euclidian distance in the acoustic feature space may be a probable answer as the measure for the clustering, 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 (neutral, anger, delight, etc.) attached to the data. However, we should note that the recognition for evaluation in section 5 was done without any information on the emotion of the test speech. Since we are dealing with the intra-speaker variation due to emotions, the current adaptation shall be called *emotion adaptation*.

4. 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

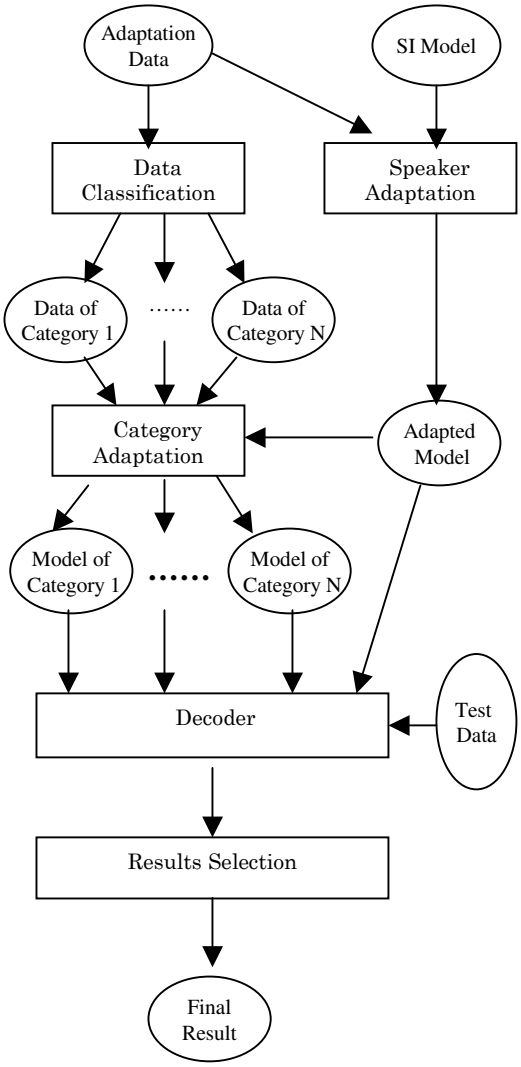

Figure 1. Block diagram of the proposed 2-level adaptation

we obtain the recognition rates $\{R_t^i\}$ (word *correct rate* or 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.

4.1. Description of test conditions

The emotional data are recorded in our laboratory. Two lists of text are read by three male actors (named $M1, M2$, and $M3$). Each list consists of 8 sentences, and each sentence is read with five types of emotion. List 1 is read once (the utterance set is called D_1), and List 2 is read twice (called D_{21} and D_{22} respectively).

So we have three data sets in our experiments: $\{D_1^i\}$, $\{D_{21}^i\}$ and $\{D_{22}^i\}$, 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.

4.2. Experimental results

Speaker $M1$ was 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 a MLLR adaptation and *HVite* is used as the recognizer. Both *HEAdapt* and *HVite* are tools provided by *HTK3.1*[5].

Figure 2 (for word *correct rate)* and Figure 3 (for word *accuracy*) display 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. And 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 figures, each emotion model achieves the highest recognition rate on the data set that belongs to it, except that the word *correct rate* obtained by model M_{sad} when tested on D_{sad}, is slightly lower than that obtained by M_{ang}. These results give us the basic support to our two-level adaptation approach.

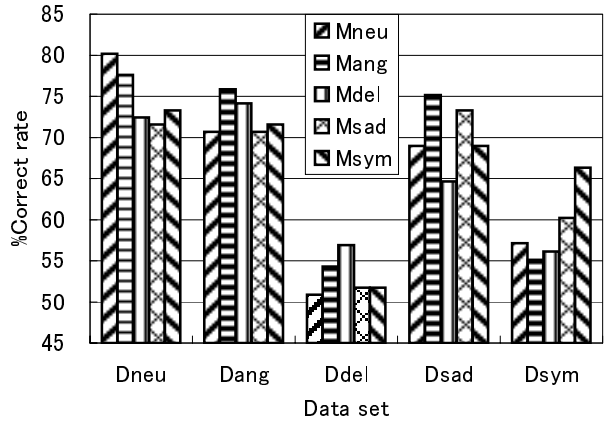

Figure 2. Recognition word correct rate for emotional speech using emotion models

5. EVALUATIONS OF THE PROPOSED METHODS

5.1. Adaptation and recognition

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

The experiments are conducted on three speakers, *M1, M2*, and *M3* 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 models for five types of emotion. The

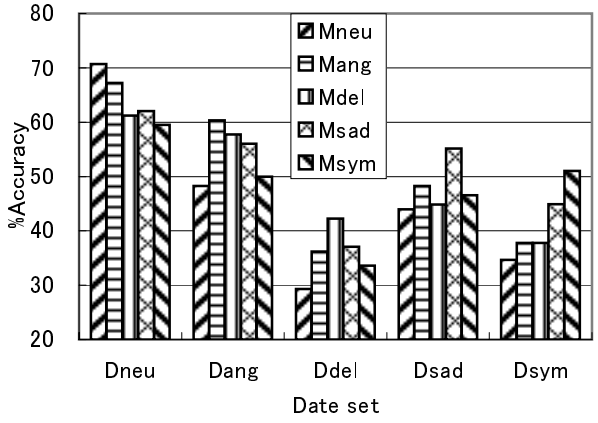

Figure 3. Recognition word accuracy for emotional speech using emotion models

Models	Percentage of recognition rate					
	M1		M2		M3	
	Corr.	Acc.	Corr.	Acc.	Corr.	Acc.
IPA-SI	63.04	42.79	60.33	49.67	63.86	54.31
MLLR	75.19	62.16	66.08	56.87	76.25	67.41
MLLR+ Emotion adaptation	79.09	68.11	67.12	61.09	78.02	69.83
Increased Percentage	3.90	5.95	1.04	4.22	1.77	3.59

Table 1. Recognition results compared with MLLR

remaining five data sets $\{D_{22}^i\}$ are used for testing. But 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 just divided into different emotion types by the emotion in which the speaker intended to utter when recording.

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.

5.2. Results

In Table 1, the row labeled "*MLLR + Emotion adaptation*" lists the recognition results of the above experiments for each of the three 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, which is generated from IPA-SI model by MLLR adaptation.

The results show that both the word *correct rate* and word *accuracy* are increased after intra-speaker emotion adaptation. Especially the increase in word *accuracy* is obvious. This demonstrated the effectiveness of the two-level adaptation in emotional speech.

6. CONCLUSION

A method of two-level adaptation was newly 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 emotional speech, and the result clearly indicated the advantage of the scheme over the conventional speaker adaptation. In the current experiment, the classification of the adaptation data was done simply according to the types of emotion attached to the data. We are now trying to evaluate the method for conversational speech, where classification should be done using a clustering scheme.

REFERENCES

[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] S.Young, D. Kershaw, J.Odell, D.Ollason, V.Valtchev and P.Woodland, "HTK-Hidden Markov Model Toolkit", *http://htk.eng.cam.ac.uk/index.shtml* , 2000