AUTOMATIC ACCENT IDENTIFICATION USING GAUSSIAN MIXTURE MODELS

Tao Chen[*,+], *Chao Huang, Eric Chang and Jingchun Wang*[*]

Microsoft Research China
5F, Sigma Center, No. 49, Zhichun Road, Beijing 100080, P.R.C
[*]Department of Automation, Tsinghua University
{chentao,wang-jc}@proc.au.tsinghua.edu.cn; {chaoh, echang}@microsoft.com

ABSTRACT

It is well known that speaker variability caused by accent is an important factor in speech recognition. Some major accents in China are so different as to make this problem very severe. In this paper, we propose a Gaussian mixture model (GMM) based Mandarin accent identification method. In this method, a number of GMMs are trained to identify the most likely accent given test utterances. The identified accent type can be used to select an accent-dependent model for speech recognition. A multi-accent Mandarin corpus was developed for the task, including 4 typical accents in China with 1,440 speakers (1,200 for training, 240 for testing). We explore experimentally the effect of the number of components in GMM on identification performance. We also investigate how many utterances per speaker are sufficient to reliably recognize his/her accent. Finally, we show the correlations among accents and provide some discussions.

1. INTRODUCTION

Speaker variability, such as gender, accent, age, speaking rate, and phones realizations, is one of the main difficulties in speech recognition task. It is shown in [1] that gender and accent are the two most important factors in speaker variability. Usually, gender-dependent model is used to deal with the gender variability problem.

In China, almost each province has its own dialect. When speaking Mandarin, the speaker's dialect greatly affects his/her accent. Some typical accents, such as Beijing, Shanghai, Guangdong and Taiwan, are quite different from each other in acoustic characteristics. Similar to gender variability, a simple method to deal with accent problem is to build multiple models of smaller accent variances, and then use a model selector for the adaptation. Cross accents experiments [2] show that performance of accent-independent systems is generally 30% worse than that of accent-dependent ones. Thus it is meaningful to develop an accent identification method with acceptable error rate.

Current accent identification research focuses on foreign accent problem. That is, identifying non-native accents. Teixeira et al. [3] proposed a Hidden Markov Model (HMM) based system to identify English with 6 foreign accents. A context independent HMM was used since the corpus consisted most of isolated words, which is not always the case in applications. Hansen and Arslan [4] also built HMM to classify foreign accent of American English. They analyzed some prosodic features' impact on classification performance and concluded that carefully selected prosodic features would improve the classification accuracy. Instead of phoneme-based HMM, Fung and Liu [5] used phoneme-class HMMs to differentiate Cantonese English from native English. Berkling et al. [6] added English syllable structure knowledge to help recognize 3 accented speaker groups of Australian English.

Although foreign accent identification is extensively explored, little has been done to domestic one, to the best of our knowledge. Actually, domestic accent identification is more challenging: 1) Some linguistic knowledge, such as syllable structure used in [6], is of little use since people seldom make such mistakes in their mother language; 2) Difference among domestic speakers is relatively smaller than that among foreign speakers. In our work, we want to identify different accent types spoken by people with the same mother language.

Most of current accent identification systems, as mentioned above, are built based on the HMM framework. Although HMM is effective in classifying accents, its training procedure is time-consuming. Also, using HMM to model every phoneme or phoneme-class is computationally expensive. Furthermore, HMM training is a supervised one: it needs phone transcriptions. The transcriptions are either manually labeled, or obtained from a speaker independent model, in which the word error rate will certainly degrade the identification performance.

In this paper, we propose a GMM based method for the identification of domestic speaker accent. 4 typical Mandarin accent types are explored. Since phoneme or phoneme class information are out of our concern, we just model accent characteristics of speech signals. GMM training is an unsupervised one: no transcriptions are needed. We train two GMMs for each accent: one for male, the other for female, as gender is the greatest speaker variability. Given test utterances, the speaker's gender and accent can be identified at the same time, compared with the two-stage method in [3]. The relationship between GMM parameter and recognition accuracy

[+] Work carried out as visiting student at MSR China.

is examined. We also investigate how many utterances per speaker are sufficient to reliably recognize his/her accent. We randomly select *N* utterances from each test speaker and average their log-likelihoods in each GMM. It is hoped that the more the averaged utterances, the more robust the identification results. Experiments show that with 4 test utterances per speaker, about 11.7% and 15.5% error rate in accent classification is achieved for female and male speakers, respectively. Finally, we show the correlations among accents and provide some discussions.

This paper is organized as follows. In Section 2, we will describe the multi-accent Mandarin corpus we collected for this task. GMM based accent identification system is presented in Section 3. Detailed experiments and result analysis are given in Section 4. Section 5 concludes with summary of our work and discussions on possible applications.

2. MULTI-ACCENT MANDARIN CORPUS

The multi-accent Mandarin corpus, consisting of 1,440 speakers, is part of 7 corpora for speech recognition research collected by Microsoft Research China. There are 4 accents: Beijing (BJ, including 3 channels, that is, collection venues: BJ, EW, FL), Shanghai (SH, including 2 channels: SH, JD), Guangdong (GD) and Taiwan (TW). All waveforms were recorded at a sampling rate of 16 kHz, except that the TW ones were 22 kHz. Most of the data were from students and staff at universities in Beijing, Shanghai, Guangdong and Taiwan, with ages varying from 18 to 40. In training corpus, there are 150 female and 150 male speakers of each accent, with 2 utterances per speaker. In test corpus, there are 30 female and 30 male speakers of each accent, with 50 utterances per speaker. Most of the utterances last about 3~5 seconds each, forming about 16 hours' speech data of the whole corpus. There is no overlap between training and test corpus. That is, all the 1,440 speakers are different.

The speaker distribution of the multi-accent Mandarin corpus is listed in Table 1.

Accent	Channel	Gender	Training Corpus		Test Corpus	
BJ	BJ	F	50	300	10	60
		M	50		10	
	EW	F	50		10	
		M	50		10	
	FL	F	50		10	
		M	50		10	
SH	SH	F	75	300	15	60
		M	75		15	
	JD	F	75		15	
		M	75		15	
GD	GD	F	150	300	30	60
		M	150		30	
TW	TW	F	150	300	30	60
		M	150		30	
ALL			**1,200**		**240**	

Table 1. Speaker distribution of the multi-accent Mandarin corpus.

3. ACCENT IDNETIFICATION SYSTEM

Since gender and accent are important factors of speaker variability, the probability density functions of distorted features caused by different gender and accent are different. As a result, we can use a set of GMMs to estimate the probability that the observed utterance comes from a particular gender and accent.

In our work, *M* GMMs, $\{\Lambda_k\}_{k=1}^{M}$, are independently trained using the speech produced by the corresponding gender and accent. That is, model Λ_k is trained to maximize the log-likelihood function

$$\log \prod_{t=1}^{T} p(x(t)\,|\,\Lambda_k) = \sum_{t=1}^{T} \log p(x(t)\,|\,\Lambda_k), k = 1,...,M, \qquad (1)$$

where the speech feature is denoted by $x(t)$. T is the number of speech frames in the utterance and M is twice (two genders) the total number of accent types. The GMM parameters are estimated by the expectation maximization (EM) algorithm [7]. During identification, an utterance is fed to all the GMMs. The most likely gender and accent type is identified according to

$$\hat{k} = \arg \max_{k=1}^{M} \sum_{t=1}^{T} \log\ p(x(t)\,|\,\Lambda_k). \qquad (2)$$

4. EXPERIMENTS

4.1. Experiments Setup

As described in Section 2, there are 8 subsets (accent plus gender) in the training corpora. In each subset, 2 utterances per speaker, altogether 300 utterances per subset, are used to train the GMMs. Since the 300 utterances in a subset are from 150 speakers with different ages, speaking rates and even recording channels, speaker variability caused by these factors is averaged. Thus we hope to represent effectively the specific gender and accent by this method. The speech data is pre-emphasized with $H(z)=1-0.97z^{-1}$, windowed to 25-ms frames with 10-ms frame shift, and parameterized into 39 order MFCCs, consisting of 12 cepstral coefficients, energy, and their first and second order differences. Cepstral mean subtraction is performed within each utterance to remove the effect of channels. When training GMMs, their parameters are initialized and reestimated once. Data preparation and training procedures are performed using the HTK 3.0 toolkit [8]. In the first experiment, we investigate the relationship between the number of components in GMM and the identification accuracy.

50 utterances of each speaker are used for test. In the second experiment, we study how the number of utterances affects the performance of our method.

4.2. Number of Components in GMM

In this experiment, we examine the relationship between the number of components in GMMs and the identification accuracy.

Since our objective is to classify the unknown utterances to a specific subset, and the eight subsets are labeled with gender and accent, our method can identify the speaker's gender and accent at the same time. When calculating the error rate of gender, we just concern with speakers whose identified gender is different with the labeled one. Similarly, when calculating the

error rate of accent, we just concern with speakers whose identified accent is error.

Table 2 and Fig. 1 show the gender and accent identification error rate respectively, varying the number of components in GMMs. The experiment is based on 1 utterance per test speaker. The relative error reduction is calculated when regarding GMM with 8 components as the baseline.

# of Components	8	16	32	64
Error Rate (%)	8.5	4.5	3.4	3.0
Rel. Error Reduction (%)	-	47.1	60.0	64.7

Table 2. Gender identification error rate with different number of components in GMM.

Table 2 shows that the gender identification error rate decreases significantly when components increase from 8 to 32. However, only small improvement is gained by using 64 components compared with 32 ones. It can be concluded that GMM with 32 components is capable of effectively modeling gender variability of speech signals.

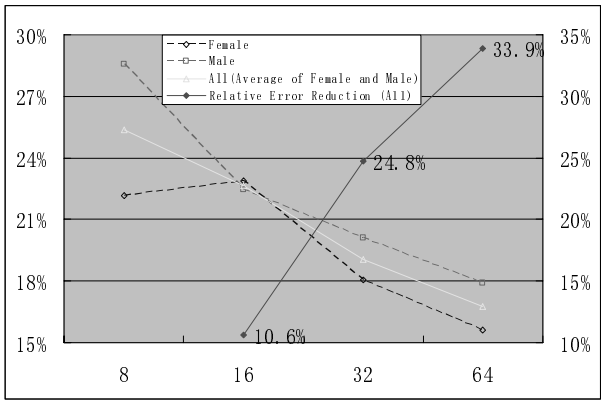

Fig. 1: Accent identification error rate with different number of components. The horizontal axis is the number of components in GMMs. The left vertical axis is the identification error rate; the right vertical axis is the relative error reduction of "All".

Fig. 1 shows the similar trend with Table 2. It is clear that the number of components in GMMs greatly affects the accent identification performance. Different to the gender experiment, in accent, GMMs with 64 components still gain some improvement over 32 ones (Error rate decreases from 19.1% to 16.8%). Since the accent variability in speech signals is more complicated and not as significant as gender, 64 components are better while describing the detail variances among accent types.

However, it is well known that to train a GMM with more components is much more time-consuming and requires more training data to obtain reliable estimation of the parameters. Concerning the trade-off between accuracy and costs, using GMMs with 32 components is a good choice.

4.3. Number of Utterances per Speaker

Sometimes it is hard even for linguistic experts to tell a specific accent type given only one utterance. Thus making use of more than one utterance in accent identification is acceptable in most applications. We want to know how many utterances are sufficient to reliably classify accent types. In experiment, we randomly select N (N<=50) utterances for each test speaker and average their log-likelihoods in each GMM. The test speaker is classified into the subset with the largest averaged log-likelihood. The random selection is repeated for 10 times. Thus 2,400 tests are performed in each experiment. This will guarantee to achieve reliable results. According to Section 3.2, 32 components for each GMM are used.

Table 3 and Fig. 2 show the gender and accent identification error rate respectively, varying the number of utterances. When averaging the log-likelihoods of all 50 utterances of a speaker, it is no need to perform random selection. The relative error reduction is calculated when regarding 1 utterance as the baseline.

# of Utterances	1	2	3	4	5	10	20	50
Error Rate (%)	3.4	2.8	2.5	2.2	2.3	1.9	2.0	1.2
Rel. Error Reduction (%)	-	18	26	35	32	44	41	65

Table 3. Gender identification error rate with different number of utterances.

Table 3 shows that it is more reliable to tell a speaker's gender by using more utterances. When the number of utterances increases from 1 to 4, the gender identification accuracy improves greatly. Still considerable improvement is observed when using more than 10 utterances. However, in some applications, it is not applicable to collect so much data just to identify the speaker's gender. Also, the results of 3~5 utterances are good enough in most situations.

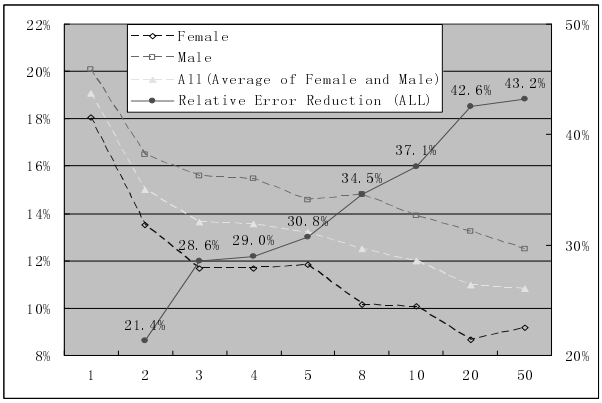

Fig. 2: Accent identification error rate with different number of utterances. The horizontal axis is the number of utterances for averaging. The left vertical axis is the identification error rate; the right vertical axis is the relative error reduction of "All".

It is clear from Fig.2 that increasing the number of utterances improves identification performance. This is consistent with our idea that more utterances of a speaker, thus more information, help recognize his/her accent better. Considering the trade-off between accuracy and costs, using 3~5 utterances is a good choice, with error rate 13.6%~13.2%.

4.4. Discussions on Inter-Gender and Inter-Accent Results

It can be noticed from Fig. 1 and Fig.2 that the accent

identification results are different between male and female. In experiments we also discovered different pattern of identification accuracy among 4 accent types. In this subsection, we will try to give some explanations.

We select one experiment in subsection 4.3 as an example to illustrate the two problems. Here GMMs are built with 32 components. 4 utterances of each speaker are used to calculate the averaged log-likelihood to recognize his/her accent. The inter-gender result is listed in Table 4. Table 5 shows the accent identification confusion matrix.

Error Rate (%)	BJ	SH	GD	TW	All Accents
Female	17.3	11.4	15.2	2.7	11.7
Male	27.7	26.3	7.6	0.3	15.5

Table 4. Inter-gender accent identification error rate.

We can see from Table 4 that Beijing (BJ) and Shanghai (SH) female speakers are much better recognized than corresponding male speakers, which causes the overall better performance for female. This is consistent with speech recognition results. Experiments in [2] show better recognition accuracy for female than for male in Beijing and Shanghai, while reverse result for Guangdong and Taiwan.

Recognized As	Testing Utterances From			
	BJ	SH	GD	TW
BJ	**0.775**	0.081	0.037	0.001
SH	0.120	**0.812**	0.076	0.014
GD	0.105	0.105	**0.886**	0.000
TW	0.000	0.002	0.001	**0.985**

Table 5. Accent identification confusion matrix.

Table 5 shows clearly different performance among accents. We provide some discussions below.

- Compared with Beijing and Taiwan, Shanghai and Guangdong are most likely to be recognized to each other, except to themselves. In fact, Shanghai and Guangdong both belong to southern language tree in phonology and share some common characteristics. For example, they do not differentiate front nasal and back nasal.
- The excellent result of Taiwan speakers may lie in two reasons. Firstly, as Taiwan civilians communicate with the Mainland relatively infrequently and their language environment is unique, their speech style is relatively easy to be recognized. Secondly, limited by the recording condition, there is a certain portion of noise in the waveform of Taiwan corpus (both training and test), which makes them more distinctive.
- The reason of relatively low accuracy of Beijing possibly lies in its corpus's channel variations. It is shown in Table 1 there are 3 channels in Beijing corpus. Greater variations lead to a more general model, which is not so specific for the accent and may degrade the performance.
- Channel effect may be a considerable factor to the GMM based accent identification system. From Beijing, Shanghai and Guangdong, accuracy increases when the number of channels decreases. Further work is needed to solve this problem.

5. CONCLUSION

In this paper, we proposed a Gaussian mixture model (GMM) based Mandarin accent identification method. GMM method can avoid building model for phoneme or phoneme-class, which is not economic for many applications. Furthermore, GMM training is an unsupervised one: no transcriptions are needed, compared with the supervised HMM training. Other than the conventionally studied foreign accent, we explored the more challenging domestic accent identification problem. 4 typical Mandarin accent types are investigated. Experiments show that properly selected GMM parameter (number of components) can result in good identification performance.

We also investigated how many utterances per speaker are sufficient to identify his/her accent reliably. Usually more utterances will guarantee better accuracy, while in many applications we cannot obtain so much data. Fortunately, our experiments show that 3~5 utterances are good enough to be used to recognize a speaker's accent by the proposed method. With 4 test utterances, about 11.7% and 15.5% error rate in accent classification was achieved for female and male speakers, respectively. Finally, we showed the correlations among accents and provide some discussions.

The accent identification system can be directly used to select accent-dependent model for speaker adaptation. Future work of its applications in speech recognition is undergoing.

6. ACKNOWLEDGEMENT

The authors would like to thank Xiaohan Li in Microsoft Research China, for discussions about Gaussian mixture model.

7. REFERENCES

[1] C. Huang, T. Chen, S. Li, E. Chang and J.L. Zhou, "Analysis of Speaker Variability," in *Proc. Eurospeech'2001*, vol.2, pp.1377-1380, 2001.

[2] C. Huang, "Accent issue in large vocabulary continuous speech recognition," *Microsoft Research Technical Report*, MSR-TR-2001-69, 2001.

[3] C. Teixeira, I. Trancoso and A. Serralheiro, "Accent Identification," in *Proc. ICSLP'96*, vol.3, pp. 1784-1787, 1996.

[4] J.H.L. Hansen and L.M. Arslan, "Foreign Accent Classification Using Source Generator Based Prosodic Features," in *Proc. ICASSP'95*, vol.1, pp. 836-839, 1995.

[5] P. Fung and W.K. Liu, "Fast Accent Identification and Accented Speech Recognition," in *Proc. ICASSP'99*, vol.1, pp. 221-224, 1999.

[6] K. Berkling, M. Zissman, J. Vonwiller and C. Cleirigh, "Improving Accent Identification Through Knowledge of English Syllable Structure," in *Proc. ICSLP'98*, vol.2, pp. 89-92, 1998.

[7] A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, vol.39, pp. 1-38, 1977.

[8] http://htk.eng.cam.ac.uk.