ON THE USE OF GAUSSIAN MIXTURE MODEL FOR SPEAKER VARIABILITY ANALYSIS

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

Microsoft Research Asia
5F, Beijing Sigma Center, No. 49, Zhichun Road, Beijing 100080, China
*Department of Automation, Tsinghua University, Beijing 100084, China
{chentao, wang-jc}@proc.au.tsinghua.edu.cn; {chaoh, echang}@microsoft.com

ABSTRACT

Analysis and modeling of speaker variability is important to help understand in-depth inter-speaker variances and to enhance current speech/speaker recognition system. In this paper we introduce adapted Gaussian mixture model (GMM) based speaker representation for the task. Two powerful multivariate statistical analysis methods, principal component analysis (PCA) and independent component analysis (ICA), are used to extract the sources of dominant speaker variability. In addition, analysis of variance (ANOVA) is adopted to evaluate the dominance of a factor in a certain principal/independent component. Further, the generalization ability of our method is investigated by experiments.

1. INTRODUCTION

As the development of automatic speech recognition (ASR) system progresses, more and more researchers are aware of the importance of speaker variability. It is well known that performance of speaker-independent (SI) ASR system is generally 2–3 times worse than that of speaker-dependent (SD) one. In recent years, speaker adaptation techniques have emerged as one solution to alleviate the mismatch between SI model and a certain test speaker. Adaptation techniques also provide us in-depth understanding of speaker variability. On the other hand, SI system and speaker adaptation can be facilitated if principal speaker variability can be modeled and corresponding compensations can be made. Furthermore, recent adaptive training scheme [1] shows that explicit modeling of different sources of variability may improve adaptation performance. All these facts motivate us to conduct analysis of speaker variability.

One of the main difficulties in analysis of speaker variability is the complexity of speech model. In large vocabulary continuous ASR system, there are a huge number of parameters associated with a set of models. If SD HMM model is used as speaker representation, the problem of high dimension brings a big challenge to our analysis.

Fortunately there are some alternative ways to describe a speaker. In [2], 15 critical band spectrums are used as feature to investigate speech variability caused by phoneme, channel and speaker. While feature vectors are suitable to characterize the variability of phoneme, it is not appropriate to analyze inter-speaker variability, in which the effect of phoneme is expected to be removed. In our previous work [3], MLLR adaptation matrix [4] is adopted to represent a speaker. Though efficient, there are still a couple of concerns with MLLR representation. First, the procedure of SI model training and adaptation is very time-consuming. Second, the performance relies heavily on the selection of regression class and feature sets, which has to be determined empirically. On the other hand, Gaussian mixture model (GMM), although has achieved great success in speaker recognition [5], is seldom used to analyze speaker variability, to the best of our knowledge.

In model based speaker representations (such as MLLR and GMM), the problem of high dimension can be solved by statistical tools, such as principal component analysis (PCA) and independent component analysis (ICA) [6]. In recent years they have been widely applied in fast speaker adaptation [7] and speaker clustering [8]. PCA decomposes second-order correlation and extracts orthogonal principal components of variances, while ICA makes unknown linear mixtures of random variables as statistically independent as possible. Vector space extracted by ICA seems to capture essential structures of data in many applications, including feature extraction and blind signal separation.

In this paper we present GMM based speaker representation for speaker variability analysis. GMM of each speaker is adapted from a universal background model (UBM). In addition to extending our previous speaker representation method, we adopt analysis of variance (ANOVA) [2] to evaluate a certain factor’s dominance in a component extracted by PCA/ICA. Three sources of speaker variability (gender, accent and speaking rate) are analyzed. Experiments show that adapted GMM is more efficient to characterize a speaker. Further we find that the first two principal/independent components clearly present the characteristics of gender and accent, respectively. In addition, it is discovered that speaking rate is not dominant in the GMM based speaker representation. At last the generalization ability of our method is tested.

This paper is organized as follows. Section 2 describes GMM based speaker representation. Section 3 introduces dominance evaluation method of variability sources. In Section 4 experiments show the effectiveness of our method.

* Work carried out as visiting student at MSR Asia.
Conclusions are presented in Section 5.

2. SPEAKER REPRESENTATION

2.1 Adapted GMM Based Speaker Representation

Various speaker representations are used for analysis of speech/speaker variability in the literature. Kajarekar et al [2] adopted 15 critical band spectrums as the feature, while focusing on the variability in phoneme. In our previous work [3], each speaker is represented by an MLLR matrix. It is assumed that MLLR matrix provides sufficient speaker characteristics given enough adaptation data. While analysis result proved the efficiency of MLLR representation, there are still two issues: selecting appropriate regression class and acoustic feature, which greatly affect analysis result. In this paper, adapted GMM is used to describe the characteristics of a speaker. GMM training is faster and requires less data compared with MLLR adaptation from SI model. In addition, no regression class selection is necessary. Experiments will show that GMM is robust under different feature sets.

Instead of being trained on each speaker’s own data, speaker-dependent GMMs are adapted from a UBM. The UBM is a general GMM trained on speech data from a large number of speakers to represent speaker-independent distribution of features. It is shown in [5] that the GMM-UBM scheme is more efficient than independently trained GMM in speaker verification.

2.2 Dominant Variability Factors Extracted by PCA/ICA

After GMM is obtained for a speaker, we align all the mean vectors of that speaker into a large supervector and concatenate supervectors of all speakers to a matrix. Then PCA/ICA is applied to determine the principal/independent components of the speaker space. It is expected that a certain source of speaker variability is dominant in one of the components of the speaker space. It is straightforward to extend equation (2) when considering unequal sample size of $d$. The contribution of factor $i$ to the total variability is measured as $\text{SSR} (\text{SS Ratio)}: \frac{\text{SS}(i)}{\text{SS(Total)}} \times 100\%$. (3)

4. EXPERIMENTS

4.1 Experiment Setups

We evaluate our method on a Mandarin corpus collected by Microsoft Research Asia, consisting of 748 speakers labeled with gender (female or male), accent (Beijing, Shanghai or Guangdong), and speaking rate set (slow, normal and fast). Except in Section 4.4, about half of the speakers (380 speakers) are used for the analysis. In all experiments, UBM is trained on these 380 speakers with 1 utterance each. 380 GMMs are adapted from the UBM using 50 utterances per speaker. Mean vectors of GMMs are aligned to a large supervector. Then 380 supervectors are concatenated to a matrix to which PCA/ICA is applied. ANOVA is performed to evaluate each factor’s effect on the principal/independent components.

We try to find the most efficient speaker representation in Section 4.2, and then illustrate analysis result in Section 4.3. In Section 4.4 we evaluate the generalization ability of the method. The supervectors of the remaining 368 speakers are projected onto the independent space learned from previously used 380 speakers. Similar analysis result will be obtained if the generalization property is guaranteed.

4.2 Efficient Speaker Representation

This subsection investigates efficient speaker representation under different configurations of feature and GMM parameters. As is expected that gender factor is related to the
first principal component, the variability of gender in this component is used as criterion.

- **Feature Sets**

Table 1 shows SSR of gender in the 1st principal component with different feature sets: static (12-D), static plus delta (24-D) and static plus delta and delta-delta MFCC (36-D). It can be concluded that the GMM based speaker representation is robust to feature sets. While static feature lacks the ability to capture the dynamic property of speech signal, incorporating 2nd order feature does not bring additional improvement. In all the following experiments, 24-D MFCC is used for its best performance among the three feature sets.

<table>
<thead>
<tr>
<th>Feature Set (MFCC)</th>
<th>12-D</th>
<th>24-D</th>
<th>36-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSR of Gender (%)</td>
<td>80.3</td>
<td>85.8</td>
<td>85.5</td>
</tr>
</tbody>
</table>

*Table 1: SSR of gender in the 1st principal component under different set of features using GMM with 16 mixtures.*

- **Number of Gaussian Mixtures**

Table 2 shows SSR under different number of Gaussian mixtures. While 8 mixtures are not able to describe a speaker sufficiently, GMM with more than 16 mixtures brings worse results. This phenomenon may be explained by inherent property of PCA. More mixtures can describe detailed and local characteristics of a speaker, which may make it more difficult for PCA to extract dominant sources of variability.

<table>
<thead>
<tr>
<th>Number of Mixtures</th>
<th>8</th>
<th>16</th>
<th>32</th>
<th>64</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSR of Gender (%)</td>
<td>83.1</td>
<td>85.8</td>
<td>82.2</td>
<td>81.2</td>
</tr>
</tbody>
</table>

*Table 2: SSR of gender in the 1st principal component under different number of Gaussian mixtures using 24-D features.*

Results shown above are suggestive to make a trade-off when selecting appropriate parameters (features and GMM mixtures) in applications, such as speaker clustering in eigenvoice [8]. In all the following experiments, GMM with 16 mixtures is used.

- **GMM vs. MLLR**

Comparison between MLLR speaker representation and GMM one is listed in Table 3. Optimal regression class and feature set are used for MLLR. It is clear that gender property is more evident in GMM. In addition, concerning that MLLR requires selecting regression class and the performance relies heavily on feature selection, GMM representation is a better choice for speaker variability analysis.

<table>
<thead>
<tr>
<th>Speaker Representation</th>
<th>MLLR</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSR of Gender (%)</td>
<td>76.7</td>
<td>85.8</td>
</tr>
</tbody>
</table>

*Table 3: Comparison of two speaker representations using SSR of gender in the 1st principal component. The SSR shown is the best case within each representation.*

### 4.3 Analysis Result

In this subsection we extend the analysis to three factors, gender, accent and speaking rate, and investigate their contributions to the total speaker variability in principal/independent components extracted from GMM based speaker representation.

Figure 1 illustrates SSR of the three factors (gender, accent and speaking rate) in top 5 principal/independent components.

- Variability due to gender and accent is very dominant across speakers. It is obvious that the first two principal/independent components have strong correlations with gender and accent, respectively.
- Variability due to speaking rate is not evident, which may be due to the insufficient ability of GMM to model duration information. Incorporating duration information may help distinguish speakers. On the contrary, GMM is insensitive to unwanted speaking rate variability caused by emotion or health problem. More experiments are needed to investigate the impact of speaking rate on current speaker recognition systems.
- ICA is more efficient than PCA to extract dominant variability factors. Although similar trend is observed in both methods, ICA is more effective in emphasizing a certain variability factor in one component while
suppressing that in others.

1st Independent Component
2nd Independent Component
BJ-F SH-F GD-F BJ-M SH-M GD-M

Figure 2: Projection of all speakers on the first 2 independent components. Speakers are labeled with gender types (Female: F; Male: M) and three accents (Beijing: BJ; Shanghai: SH; Guangdong: GD).

In Figure 2, projection of all speakers on the first two independent components shows that major information about gender and accent is efficiently represented by ICA space. If applied in clustering method, ICA space may produce clusters consistent with dominant speaker variability characteristics.

4.4 Generalization Ability

In this subsection we evaluate the generalization ability of our analysis method. First an ICA space is created by the 380 speakers described above. Supervectors (obtained by the same method in Section 4.3) of the rest 368 speakers in the corpus are projected to the ICA space. Then top 5 independent components are used for analysis.

![Figure 3: SSR of three factors by projecting unseen 368 speakers to the formerly created ICA space.](image)

The results illustrated in Figure 3 shows the same trend as that in Figure 1. Since the remaining 368 speakers are unseen when either training UBM or creating the ICA space, the method is proved to have good generalization property.

5. CONCLUSIONS

In this paper we investigate the inter-speaker variability based on adapted GMM speaker representation. We suggest extracting key factors of speaker variability by PCA/ICA and evaluate the dominance of these factors by analysis of variance. Three sources of speaker variability, gender, accent and speaking rate, are analyzed. Experiments show that adapted GMM is more efficient to characterize a speaker than our formerly used MLLR. Further we find that the first two principal/independent components clearly present the characteristics of gender and accent respectively, which can be used in speaker clustering. In addition, it is discovered that speaking rate is not dominant in the GMM based speaker representation even with dynamic features. We suggest performing further experiments to explore in-depth impact of speaking rate on speaker recognition system. At last the generalization ability of the proposed method is demonstrated.

Currently we are working on applying our analysis results to speaker selection training in large vocabulary speech recognition.

6. ACKNOWLEDGEMENT

The authors would like to thank Ming Liu and Chengyuan Ma at Microsoft Research Asia for discussions about adapted GMM.

7. REFERENCES