Speaker Recognition Using the Resynthesized Speech via Spectrum Modeling

Xiang Zhang, Chuan Cao, Lin Yang, Hongbin Suo, Jianping Zhang, Yonghong Yan

ThinkIT Speech Lab, Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, P.R. China
{xzhang, ccao, lyang, hsuo, jzhang, yyan}@hccl.ioa.ac.cn

Abstract

Recently, using prosodic information such as pitch and energy for speaker recognition has attracted much attention. However, these kinds of systems yield performance much worse than the traditional cepstral-based systems. Limited performance improvement can be achieved when combining the two kinds of systems. In this paper, we present a new approach for speaker recognition, which uses the prosodic information calculated on the original speech to resynthesize the new speech data utilizing the spectrum modeling technique. The resynthesized data are modeled with sinusoids based on pitch, vibration amplitude and phase bias. We use the resynthesized speech data to extract cepstral features for speaker modeling and scoring in the same way as in traditional speaker recognition approaches. Then, we model these features using GMMs and compensate for speaker and channel variability effects using joint factor analysis. The experiments are carried out on the core condition of NIST 2008 speaker recognition evaluation data. The experimental results show that our proposed system achieves comparable performance to the state-of-the-art cepstral-based joint factor analysis systems which use the original data for speaker recognition. Besides, the fusion of the two kinds of systems can achieve significant performance improvement compared to the cepstral-based system alone.

Index Terms: Speaker recognition, prosodic features, joint factor analysis, spectrum modeling

1. Introduction

In the current framework of speaker recognition, a large majority of speaker recognition systems are based on short term cepstral features such as Mel Frequency Cepstral Coefficients (MFCCs) and Linear Predictive Cepstral Coefficients (LPCCs). Several speaker model techniques have been proposed for these kinds of features, such as Gaussian Mixture Models (GMM) [1] and Support Vector Machines (SVM) [2]. These systems have become the predominant approaches in the NIST speaker recognition evaluation (SRE). The performance of such systems is however relatively sensitive to the recording conditions.

Prosodic information characterizes the speaker’s intonation and speaking style. In recent years, interesting characteristics of prosodic features, such as pitch and energy contours of speech [3], have attracted much attention for their robustness to the channel distortion and their complementary advantage to the cepstral features in speaker recognition. However, using prosodic features alone achieves performance much worse than cepstral features. The fusion of these two kinds of features usually cannot significantly improve the performance of conventional cepstral-based speaker recognition systems, especially on the minimum Detection Cost Function (minDCF) [4]. Besides, prosodic systems yield little effectiveness when insufficient adaptation data are available to train speaker models. Thus, the prosodic speaker recognition systems were rarely used in the recent NIST SRE for most of the participants [5][6].

In this paper, we propose a novel approach for speaker recognition, which uses the prosodic information such as pitch, vibration amplitude and phase bias to resynthesize the harmonic structure of the speech signals using spectrum modeling technique [9]. A selected spectrum is applied to weaken the effects of the channel variability and noise, aiming to furthermore diminish spectrum difference between speech signals recorded under different channels. The harmonic structure represents the time-varying spectral characteristics of a speech sound, containing the most important information for speaker recognition, which is seen as the deterministic components of the signal and is modeled as the sum of a set of quasi-sinusoids in [9]. In this study, we use prosody-resynthesized speech data to extract traditional cepstral features for speaker modeling and scoring. This approach can be seen as a combination of prosodic and acoustic information. The resynthesized speech are generated based on the high level prosodic information. The cepstral features extracted from the resynthesized data reflect the low level acoustic characteristics of the speakers. We then model these features using GMMs and compensate for speaker and channel variability effects using joint factor analysis. The experimental results show that the proposed approach achieves comparable results to the state-of-the-art cepstral-based speaker recognition systems using original speech data. The fusion of the two types of systems results in an improved performance over the system using original speech data.

The structure of this paper is organized as follows: In Section 2, we summarize the joint factor analysis model. Section 3 describes our proposed prosodic-based approach for speaker recognition. Experiments and results are presented in Section 4. The fusion results are also given in this section. Section 5 presents the conclusion.

2. Joint factor analysis

The joint factor analysis [8][10] approach has become the state-of-the-art in the field of speaker recognition during the last few years. This modeling addresses the problem of speaker and session variability in Gaussian Mixture Models (GMM) framework. It can be seen as an extension for conventional GMM-UBM approach, which joints the eigenvoices speaker modeling, eigenchannels variability compensation and relevance MAP adaptation into a unified framework. In this model, each speaker is represented by the means, covariance, and weights of a $C$ multivariate diagonal-covariance Gaussian densities defined in some continuous feature space of dimension $F$. The GMM for a target speaker is obtained by adapting the universal background model (UBM) mean parameters to the target
speaker data. In joint factor analysis, the basic assumption is that a speaker and channel dependent supervector of means $M$ can be decomposed into a sum of two supervectors: a speaker supervector $s$ and a channel supervector $c$

$$M = s + c \quad (1)$$

where $s$ and $c$ are normally distributed. In [11], Kenny et al. described in greater details how the speaker dependent supervector and channel dependent supervector can be represented in low dimensional space. The first term in the right hand side of (1) is modeled by assuming that if $s$ is the speaker supervector for a randomly chosen speaker then

$$s = m + V y + D z \quad (2)$$

where $m$ is the speaker and channel independent supervector (the mean supervector of UBM). $D$ is a diagonal matrix, $V$ is a rectangular matrix of low rank named eigenvoice space and $y$ and $z$ are independent random vectors having standard normal distributions. In other words, $s$ is assumed to be normally distributed with mean $m$ and covariance matrix $VV^T + DD^T$. The components of $y$ and $z$ are respectively the speaker and common factors.

The channel dependent supervector $c$, which represents the channel effect in an utterance, is assumed to be distributed according to

$$c = U x \quad (3)$$

where $U$ is a rectangular matrix of low rank named eigenchannel, $x$ is distributed with standard normal distribution. This is equivalent to saying that $c$ is normally distributed with zero mean and covariance $UU^T$. The components of $x$ are the channel factors. When both speaker and channel spaces are defined, the term joint factor analysis (JFA) is used to refer to this modeling.

The underlying task in JFA is to train the hyperparameters $U$, $V$ and $D$ on a large training set. In the Bayesian framework, posterior distribution of the factors (knowing their priors) can be computed using the enrollment data. The likelihood if the test utterance $\chi$ is then computed by integrating over the posterior distribution of $y$ and $z$, and the prior distribution of $x$ [12].

3. Proposed prosodic-based approach for speaker recognition

3.1. Sinusoidal modeling theory

Harmonic structure refers to the structure of the primary spectral partials of a speech signal, which subjects to a harmonic-related pattern, including information such as oscillating frequency, vibration amplitude and phase bias. An intuitive harmonic structure modeling method is the sinusoidal model proposed by Serra [9], in which the harmonic structure is seen as the deterministic components of the sound signal and is modeled as the sum of series of quasi-sinusoidal components (sinusoids with slowly amplitude and frequency). Each sinusoid models a narrowband component of the original sound and is described by an amplitude and a frequency function. For a given analysis frame, the deterministic components of the signal can be modeled by

$$d(n) = \sum_{r=1}^{R} \hat{A}_r \cos(2\pi \frac{N}{R} r f_r + \phi_r), n = 0, 1, ..., S - 1 \quad (4)$$

where $R$ is the total number of harmonic partials and $S$ is the length of the frame, $\hat{A}_r$ refers to the vibration amplitude of partial $r$, $f_r$ is the oscillating frequency and $\phi_r$ is the initial phase bias. Basically, $f_r$ in equation (4) equals to integer multiples of the fundamental frequency $f_0$.

3.2. Details of harmonic structure resynthesis

The primary issue for harmonic structure resynthesis is to estimate related prosodic information, such as the frequencies, amplitudes and phase biases of partials. Fig. 1 shows the framework of the harmonic extraction and synthesis process used in our study. Firstly, fundamental frequencies are estimated frame by frame, and then only frames with valid pitch values (50Hz-500Hz) are taken into account. And harmonic structure coefficients are estimated by using the fundamental frequency information. Finally, the deterministic part of the speech signal is resynthesized with the sinusoidal model as in subsection 3.1 frame-by-frame, to be the concise representation of the signal.

Due to the imperfection of vocal vibration, harmonic partials may not appear at the frequency of integer multiples of the fundamental frequency. Thus, in our implementation, the spectrum bins with the maximum magnitude in a small range near integer multiples of the fundamental frequency are considered as the target partials. Instantaneous frequency, amplitude and phase of the partials’ bins are estimated to represent the harmonic structure. Smooth strategies are also applied to avoid discontinuity between adjacent frames.

3.3. Feature extraction and modeling

After the process of synthesizing the deterministic of the speech signals, the primary information of the voiced are retained in the resynthesized data. We use these data for speaker recognition in this study. The Mel Frequency Cepstral Coefficients (MFCCs)
are extracted from the resynthesized signals as in the state-of-the-art cepstral-based systems. We then model these features using GMMs and compensate for speaker and channel variability effects using joint factor analysis (JFA) for further improvement of the recognition performance. We refer this system as JFA-GMM.

Besides, we also use the speaker supervisor \( s = m + V y + D \) derived in JFA modeling process as the feature for support vector machine (SVM) based speaker recognition. The kernel inner product used is

\[
K(g_a, g_b) = \sum_{i=1}^{N} (\sqrt{\lambda_i} \Sigma_i^{-\frac{1}{2}} s_{a,i})^T \sum_{i=1}^{N} (\sqrt{\lambda_i} \Sigma_i^{-\frac{1}{2}} s_{b,i})
\]

where \( s_{a,i} \) are the speaker mean vectors, \( \lambda_i \) are the mixture weights of the UBM, and \( \Sigma_i \) are the UBM covariances. SVMs are trained using SVMLight. This system is denoted by JFA-SVM.

4. Experiments

4.1. Experimental setup

The experiments for different systems based on the two kinds of speech data (original data and resynthesized data) and the fused systems are carried out on the NIST 2008 core condition (short2-short3). The NIST 2008 evaluation tasks are distinguished by including in the training and test conditions not only conversational telephone speech but also interview speech recorded with different microphones involving an interview scenario. The primary evaluation condition is short2-short3, in which the training condition is called short2, and the test condition short3. The short2 training data includes conversational telephone speech and interview speech. The short3 testing data consists of not only the same two types of speech as short2 but also the telephone speech recorded on an ancillary microphone channel. NIST has separated the trials according to their channel type. In this study, we carry out the experiments on three types of trials: telephone-telephone, and interview-interview and interview-telephone. We report the performance in terms of equal error rate (EER), minimum decision cost function value (minDCF) [4] and DET curves.

The features used for recognition are MFCCs, which consist of 12 cepstral coefficients and their first and second derivatives to produce a 36 dimensional feature vector. An energy-based speech detector is applied to discard vectors from low-energy frames. To mitigate channel effects, feature warping is applied to features.

The gender dependent UBM models with 1024 components are trained using the NIST SRE 2004 Isidre training corpus. We use the P2 and P3 of Switchboard II and Switchboard Cellular to train the eigenvoice space \( V \) with 300 speaker factors. And the SRE 2004 corpus is used to train the diagonal matrix \( D \). The NIST SRE 2004, 2005, 2006 corpus and the MIX5 interview development data are used to train three types of 100-rank eigencorrelation spaces. One space is trained by only telephone data, used for the telephone-telephone condition. Another space is trained by microphone and the interview data that used for the interview-interview condition. The third one is trained by all the data above, which is used for the telephone-interview trial condition. For telephone-telephone condition, Tnorm and Znorm utterances are from the NIST SRE 2006 corpus, and the telephone utterances from NIST SRE 2004 are used as the negative data for SVM training. For interview-interview conditions, Tnorm and Znorm utterances are taken from NIST SRE 2006 auxiliary microphone data and the MIX5 interview data, and the NIST SRE 2005 auxiliary microphone utterances are used as the negative examples for SVM training. For the cross-channel interview-telephone condition, Tnorm and negative data are the same with interview-interview condition, and the Znorm utterances are the same with telephone-telephone condition. The fusion method we used is a linear fusion.

4.2. Experimental results

Table 1: JFA-GMM systems based on different types of speech utterances. The upper row in each table cell is the EER (%). The lower row is the minDCF value.

<table>
<thead>
<tr>
<th>task (train-test)</th>
<th>original data</th>
<th>resynthesized data</th>
<th>fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>tel-tel</td>
<td>5.92</td>
<td>7.24</td>
<td>5.23</td>
</tr>
<tr>
<td>intmic-intmic</td>
<td>2.15</td>
<td>2.28</td>
<td>1.40</td>
</tr>
<tr>
<td>intmic-tel</td>
<td>5.88</td>
<td>5.64</td>
<td>4.61</td>
</tr>
</tbody>
</table>

Table 2: JFA-SVM systems based on different types of speech utterances. The upper row in each table cell is the EER (%). The lower row is the minDCF value.

<table>
<thead>
<tr>
<th>task (train-test)</th>
<th>original data</th>
<th>resynthesized data</th>
<th>fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>tel-tel</td>
<td>6.36</td>
<td>6.48</td>
<td>5.66</td>
</tr>
<tr>
<td>intmic-intmic</td>
<td>2.47</td>
<td>2.94</td>
<td>1.80</td>
</tr>
<tr>
<td>intmic-tel</td>
<td>7.57</td>
<td>5.75</td>
<td>5.25</td>
</tr>
</tbody>
</table>

Table 3 lists the fusion of the JFA-GMM and JFA-SVM systems based on the different kinds of speech data. The final fusion results on the test conditions are also given. From the results, we can see that the resynthesized speech data own some complementary capacity to the original data. It achieves 9.1%, 29.6%, 31.5% improvements on the minDCF on the tel-tel, intmic-intmic and intmic-tel conditions, respectively. Our
Table 3: The combination of JFA-GMM and JFA-SVM based on different types of speech utterances. The upper row in each table cell is the EER (%). The lower row is the minDCF value.

<table>
<thead>
<tr>
<th>task (train-test)</th>
<th>original data</th>
<th>resynthesized data</th>
<th>fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>tel-tel</td>
<td>5.51</td>
<td>5.84</td>
<td>5.11</td>
</tr>
<tr>
<td></td>
<td>0.0264</td>
<td>0.0266</td>
<td>0.0240</td>
</tr>
<tr>
<td>intmic-intmic</td>
<td>1.85</td>
<td>2.13</td>
<td>1.34</td>
</tr>
<tr>
<td></td>
<td>0.098</td>
<td>0.0109</td>
<td>0.0069</td>
</tr>
<tr>
<td>intmic-tel</td>
<td>5.87</td>
<td>4.46</td>
<td>3.83</td>
</tr>
<tr>
<td></td>
<td>0.0235</td>
<td>0.0190</td>
<td>0.0161</td>
</tr>
</tbody>
</table>

In this paper, we proposed a novel prosodic information based approach for speaker recognition. This approach uses the prosodic information such as pitch, vibration amplitude and phase bias to resynthesize the harmonic structure of the speech data using the sinusoidal modeling theory, which is robust to the channel distortion and yields complementary advantage to the traditional cepstral systems based on the original speech signals. Our experiments on the NIST SRE 2008 core condition show that our proposed approach yields comparable performance to the conventional systems and can significantly improve the speaker recognition performance when it is fused with the state-of-the-art systems.

6. Acknowledgements

This work is partially supported by The National Science & Technology Pillar Program (2008BA150B03), National Natural Science Foundation of China (No. 10925419, 90920302, 10874203, 60875014).

7. References