2010 International Conference on Audio Language and Image Processing (ICALIP)
pp. 25-30, November 23-25, Shanghai, China

Multi-layered Features with SVM for Chinese Accent Identification

[1,2]Jue HOU, [1]Yi LIU, [1]Thomas Fang ZHENG, [3]Jesper OLSEN, [3]Jilei TIAN
[1]Center for Speech and Language Technologies, Division of Technology Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, Beijing, China
[2]Department of Computer Science and Technology, Tsinghua University, Beijing, China
[3]Nokia Research Center, Beijing, China
houj04@mails.tsinghua.edu.cn, {eeyliu, fzheng}@tsinghua.edu.cn
{jesper.olsen, jilei.tian}@nokia.com

Abstract

In this paper, we propose an approach of multi-layered feature combination associated with support vector machine (SVM) for Chinese accent identification. The multi-layered features include both segmental and suprasegmental information, such as MFCC and pitch contour, to capture the diversity of variations in Chinese accented speech. The pitch contour is estimated using cubic polynomial method to model the variant characters in different accents in Chinese. We train two GMM acoustic models in order to express the features of a certain accent. As the original criterion of the GMM model cannot deal with such multi-layered features, the SVM is utilized to make the decision. The effectiveness of the proposed approach was evaluated on the 863 Chinese accent corpus. Our approach yields a significant 10% relative error rate reduction compared with traditional approaches using sole feature at single level in Chinese accented speech identification.

1. Introduction

Automatic accent identification is the task of identifying the speaker's accent from a given spoken utterance. It can be used in automatic speech recognition (ASR) system to improve the recognition accuracy. In addition, if it is used in automated telephone helplines, analyzing accent and then directing callers to the appropriately accented response system may improve customer comfort and understanding [1].

Over the past years, many approaches have been applied in accent identification task. It is shown that most of these research works focused on English accents caused by foreign speakers or non-native speakers, such as American English versus Indian accented English [2]. On the other hand, due to special pronunciation structure in Mandarin as well as the diversity of accents in Chinese, such work using similar algorithms on Mandarin Chinese is not to be followed smoothly and efficiently.

There are many approaches for doing language, dialect and accent identification [1-5]. In general, the previous work focused on either using segmental information or using suprasegmental information solely to process accent identification. In the paper of [1] and [4], MFCC features were used to train different types of acoustic models to extract typical characters among accents for identification. In addition, [4] used GMM to model accent characteristics of speech signals, for GMM training is unsupervised and computationally inexpensive compare to Hidden Markov Model (HMM) [4]. In [3], suprasegmental information, such as pitch contour was extracted using Legendre polynomials approach. In [5], pitch flux was adopted as an important role to identify Chinese dialects.

However, there are still challenges in the above approaches, especially for Chinese accent identification. Accent is quite different from dialect. Most speakers of Mandarin Chinese learned Mandarin (Putonghua) as a second language, and their pronunciations are strongly influenced by their native regional language, which causes the accent problem. In addition, the previous approaches always used segmental and suprasegmental aspects independently, which cannot catch sufficient information to model the effects of variations in different accents. As a result, the differentiation of accent cannot reach high resolution. Moreover, the decision-making criterion cannot combine different layered features for accent identification.

In this paper, we propose a novel approach of multi-layered feature combination associated with SVM for Chinese accent identification. The pitch contour was estimated using cubic polynomial method to model the variant characters in different accents in Chinese. Meanwhile, the conventional MFCC method is used, we train a GMM acoustic model in order to express the segmental features of a certain accent. As the original criterion of GMM model cannot deal with such multi-layered features, the SVM is utilized to make the decision. Compared to previous methods, our approach makes full use of features of the accented speech. Therefore, we expect a decrease in error rate compare to those previous methods using sole feature information.

The paper is organized as follows: Section 2 generally describes the Chinese accents and difficulties in Chinese accent identification, Section 3 shows how we exploit the pitch contour as a feature, SVM related decision method is in Section 4, the experimental configurations and results are shown in Section 5. We conclude our work in Section 6.

2. Chinese accents

2.1. Chinese accents for accent identification

Chinese has seven major accent regions all over China [6], and there are great differences among them. Some of the accents (e.g., Guangdong accent) are quite different from standard Mandarin.

Accent is severe problem for Chinese speakers since the written Chinese character is ideographic and independent of its pronunciation [6]. All the accents share the same phoneme system and the same grammar as well as sentence structures.

Unlike some of the European languages, such as English and German, Chinese is a tone rich language with multiple intonations. The intonations are important information for people to understand the spoken Chinese [5]. In addition, the main difference among Chinese accents is the prosodic features, such as tone or intonation. Therefore, to utilize the prosodic features is the key challenge to raise the accuracy of accent identification of Chinese.

2.2. Tone information in accent identification

In order to obtain the initial results of Chinese accent identification using multi-layered features integrated with tone information, we selected several typical accents to validate our approach. We firstly tried Guangdong and Chongqing. Chongqing accent is relatively similar to standard Mandarin, while Guangdong accent is greatly influenced by the dialect of Cantonese. Our prior experiment showed the accuracy of identifying Guangdong accent and Chongqing accent was about 98% using solely 26-dimensional MFCC features on 64-component GMM models without any pitch information. Therefore, we finally chose Shanghai instead of Guangdong in the corpus, so that the mixed accents were not so different.

One of the main differences between Shanghai accent and Chongqing accent is the tone. Standard Mandarin has four tones, while Shanghai accent has five. The Chongqing accent has four tones, the same as standard Mandarin. However, they are not identical to Mandarin. In Chongqing accent, the first tone is a high rising tone, the same as standard Mandarin's second tone. The second tone is not present in standard Mandarin and the third and fourth tones are the reverse of standard Mandarin.

Except for the tones, the Chongqing accent and shanghai accent is highly confusable, especially in segmental aspects. The two accents share some differences from standard Mandarin, such as both of them do not distinguish between alveolar nasal and velar nasal, retroflex affricate and retroflex fricative. Therefore, the traditional MFCC based method cannot get as high accuracy as those experiments performed on Guangdong.

3. Multi-layered features

3.1. MFCC for accent identification

Mel Frequency Cepstral Coefficients (MFCCs) provide an efficient means of representing the frequency components of the speech waveform [1]. It has been widely used in automatic speech recognition (ASR) and speaker identification. As shown in [1] and [4], MFCC is also useful in accent identification. We use MFCC as one kind of the features. The detailed configurations are shown in Section 5.3.

3.2. Pitch contour extraction

We use the tool of Praat [7] to extract the pitch contour features. Figure 1 shows an example of extracted pitch contour (the real lines on the spectrum). As the Chinese language is character based, normally one character is exactly one syllable, so it is clear that most of the time the contour is already segmented. Some connected contours, such as the segment on the right most, shows how the pitch changes during continuous speech. Thus, we did not need to do any further segmentation work.

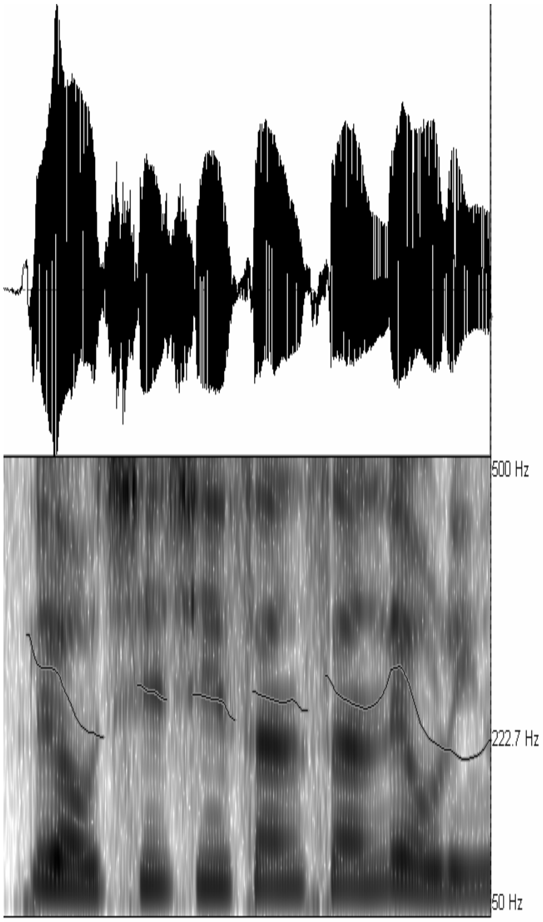

Figure 1.The pitch contour extracted using Praat

3.3. Cubic polynomial for pitch contour approximation

It is clear in Figure 1 that not all of the lengths of the segments are identical. In order to utilize the segments as features, we use cubic polynomials to approximate the contours. Further, use the coefficients to form a feature vector. The algorithm is as follows.

For each pitch contour segment, the data is given as $\{(x_i, y_i), i = 0, 1, \ldots, m\}$. We try to use a function $y = S^*(x)$ to approximate the data. If we denote the approximation error $\delta_i = S^*(x_i) - y_i$ for $i = 0, 1, \ldots, m$ and $\boldsymbol{\delta} = (\delta_0, \delta_1, \ldots, \delta_m)^T$, then the problem is to find a $S^*(x)$ in $\psi = span\{1, x, x^2, x^3\}$ so that

$$\begin{aligned}\|\boldsymbol{\delta}\|_2^2 &= \sum_{i=0}^{m} \delta_i^2 = \sum_{i=0}^{m} [S^*(x_i) - y_i]^2 \\ &= \min_{S(x)\in\psi} \sum_{i=0}^{m} [S(x_i) - y_i]^2\end{aligned} \tag{1}$$

Here $S(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$.
If we define

$$(\psi_j, \psi_k) = \sum_{i=0}^{m} \psi_j(x_i)\psi_k(x_i) \tag{2}$$

$$(f, \psi_k) = \sum_{i=0}^{m} f(x_i)\psi_k(x_i) \equiv d_k \tag{3}$$

Then the coefficients we need are the solution to the equation [8]

$$\boldsymbol{G}\boldsymbol{a} = \boldsymbol{d} \tag{4}$$

Here

$$\begin{aligned}\boldsymbol{a} &= (a_0, a_1, a_2, a_3) \\ \boldsymbol{d} &= (d_0, d_1, d_2, d_3)^T\end{aligned}$$

$$\boldsymbol{G} = \begin{bmatrix} (\psi_0, \psi_0) & (\psi_0, \psi_1) & (\psi_0, \psi_2) & (\psi_0, \psi_3) \\ (\psi_1, \psi_0) & (\psi_1, \psi_1) & (\psi_1, \psi_2) & (\psi_1, \psi_3) \\ (\psi_2, \psi_0) & (\psi_2, \psi_1) & (\psi_2, \psi_2) & (\psi_2, \psi_3) \\ (\psi_3, \psi_0) & (\psi_3, \psi_1) & (\psi_3, \psi_2) & (\psi_3, \psi_3) \end{bmatrix}$$

According to [3], we use the length of the pitch contour segment, the coefficients a_1 and a_2 to form the 3-dimensional feature.

4. Decision criterion using SVM

4.1. The SVM

The SVM is an efficient algorithm for data classification [9]. It has been used successfully on a wide variety of tasks including text and image classification. It is used in identifying Arabic and Indian accents of English in [1].

Given a training set of instance-label pairs (X_i, y_i) for $i = 1, \ldots, l$ where $X_i \in R^n$ and $y \in \{1, -1\}^l$. The training vectors are mapped into a higher (maybe infinite) dimensional space [9]. The SVM finds a linear hyperplane with the maximal margin in this higher dimensional space. It is suitable for classification on low-dimensional features. On the test process, test vectors are then mapped into that same space, and to see on which side of the gap they fall on.

4.2. Decision rules of single-layered feature

In the traditional GMM based identification method, suppose that the given test vector is λ and the GMM is G, then the likelihood score can be calculated by the following function:

$$p(x|\lambda) = \sum_{i}^{M} \omega_i p_i(x) \qquad (5)$$

Here ω_i is the weight for each GMM component, and $p_i(x)$ is the *i*-th Gaussian component. It can be expressed as

$$p_i(x) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|} \exp\{-\frac{1}{2}(x - \mu_i)^{'} \Sigma_i^{-1} (x - \mu_i)\} \qquad (6)$$

The rule of the decision is to select the model that has the largest score as the result.

4.3. Decision on multi-layered features

As we are using two kinds of features (MFCC and pitch contour), we could not directly use the algorithm described in Section 4.2. We utilize the SVM to evaluate the likelihood among different kinds of models. The process of training and testing with an SVM is shown as follows:

We firstly train an MFCC feature based GMM model and a pitch contour feature based GMM model for each accent. For each item of labeled training data, perform a likelihood score calculation using function (1) on all of the GMM models of its accent, and form the scores into a vector. Then the SVM is trained using the score vectors. On the test process, we firstly extract the features of a given test utterance, perform score calculation on all of the GMMs. Then use the SVM trained as above to give out the identification results.

5. Experimental results

5.1. System Overview

Our system contains the training part and the testing part. The procedure has been described in Section 4.3. Figure 2 shows the prototype of our system.

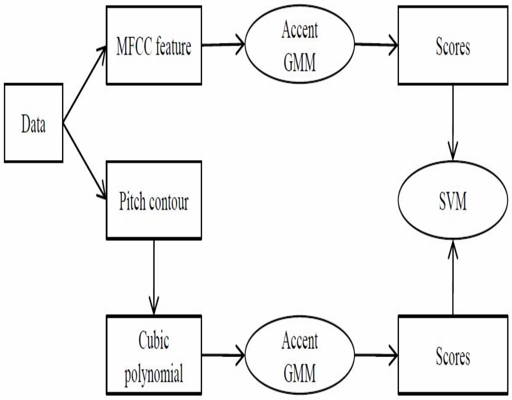

Figure 2. A prototype of the proposed system

5.2. Resources

We used the 863 Chinese accent corpus to evaluate the proposed method. The selected accent regions were Shanghai and Chongqing. All speech data files were provided in 16 kHz, 16 bit, WAVE format.

Each of the regions contained 100 male speakers and 100 female speakers, about 150 sentences for each person. The contents of the utterances varied from single word to long sentences, and the lengths were from 3 seconds to 30 seconds.

For each accent region, the training set contained 170 speakers, 85 male and 85 female. Each speaker had about 120 utterances that were randomly selected from the original 150 utterances. The testing set contained 15 male speakers and 15 female speakers with about 80 utterances for each speaker. The speakers in the test and training sets did not overlap.

5.3. Configurations of experiments

We used the HTK toolkit [10] to extract the 12 MFCC features with the energy and their first order derivatives or MFCC_E_D in HTK format. The frame length was set to 20 ms and the window step was 10 ms. The pre-emphasis coefficient was set to 0.97.

We further used Praat to extract the pitch contours. The algorithm chosen in Praat was the autocorrelation method. The parameters were same as those used in [3].

We used the LibSVM [9] to be our SVM training and testing tool. We exploited LibSVM's built-in function to determine and optimize the parameters as well as kernel function used in training an SVM.

5.4. Analysis of experimental results

5.4.1. The baseline

As described in Section 4.3, we used HTK to train a GMM model to identify the accents same as [4] to be our baseline system. According to conclusions of some existing approaches [3][4], the number of GMM components was set to 64. As there was no pitch contour information, the decision was simply made by the GMM assumption: select the model that has the bigger likelihood score to be the identification result. Table 1 shows the result and the accuracy. The result shows the Chongqing accent and the shanghai accent can be distinguished, the overall accuracy is 74.5%.

Table 1. Accuracy of the baseline system

Accent	Correct	Incorrect
Chongqing	74.5%	25.5%
Shanghai	74.4%	25.6%
Overall	**74.5%**	**25.5%**

5.4.2. Using only pitch contour features

Another approach in accent identification was using only pitch contour information as in [3]. We extracted 3-dimensional features using the algorithm described in Section 3.3, and then used those features to train a 64-component GMM model. Table 2 shows the results. It can be seen that the two accents can be distinguished, and the overall accuracy of the identification is 63.4%, which is worse than using MFCC features.

Table 2. Accuracy of accent identification using only pitch contour information

Accent	Correct	Incorrect
Chongqing	65.7%	34.3%
Shanghai	61.5%	38.5%
Overall	**63.4%**	**36.6%**

5.4.3. Using both features and SVM in decision

In this part, we show the results of the proposed method. We used both MFCC features and pitch contour features to form a multi-layered feature, and then utilized the SVM for decision-making. The algorithm was described in detail in Section 4.3. Table 3 shows the experimental results.

Table 3. Error rate reduction for using both features and SVM compared to conventional sole feature at single level

Accent	Correct	Incorrect	Error rate reduction	
			Only MFCC	Only pitch
Chongqing	76.8%	23.2%	9.1%	32.5%
Shanghai	77.3%	22.7%	11.0%	40.9%
Overall	**77.1%**	**22.9%**	**10.1%**	**37.3%**

As a result, the overall accuracy of the proposed method is 77.1%. Therefore, the error rate is decreased by 10.1% relative to the traditional MFCC approach, and 37.3% relative to using only pitch information.

6. Conclusions

In this paper, we propose multi-layered features including MFCC and pitch contour to model the diversity of variations in accented speech. We use cubic polynomials to estimate the pitch contour segments. The use of multi-layered features is able to capture both the segmental and suprasegmental characteristic features of accented speech. We also exploit SVM to integrate the two kinds of features to make the decision, so that we do not need to involve confidence measure issues.

The effectiveness of the proposed approaches was evaluated on Chongqing and Shanghai accents of the

863 accented Chinese corpus. The experimental results showed that our approach gained a higher accuracy than those use only MFCC features or pitch contour features. We have achieved a magnificent 10.1% reduction in error rate compared to only MFCCs and 37.3% to pitch contours.

Due to the diversity of variations in Chinese accented speech, the future work includes using more effective approaches to describe the pitch contour accurately, as well as involving other typical features (e.g., rhythm) in the feature vectors for accent identification on more Chinese accents.

Acknowledgements

This work was supported by Natural Science Foundation of China (60975018), the joint research grant of Nokia-Tsinghua Joint Funding 2008-2010, and New Teacher Grant of Ministry of Education of China (20090002120012).

References

[1] C. Pedersen and J. Diederich, "Accent Classification Using Support Vector Machines", *6th IEEE/ACIS International Conference on Computer and Information Science* (ICIS 2007), Melbourne, Australia, 11-13 July 2007, pp.444-449.

[2] S. Deshpande, S. Chikkerur and V. Govindaraju, "Accent Classification in Speech", *Fourth IEEE Workshop on Automatic Identification Advanced Technologies* (AutoID'05), Buffalo, New York, 17-18 October 2005, pp.139-143.

[3] C.Y. Lin and H.C. Wang, "Language Identification using Pitch Contour Information", *International Conference on Acoustics, Speech, and Signal Processing* (ICASSP 2005), Philadelphia, Pennsylvania, 18-23 March 2005, pp.601-604.

[4] T. Chen, C. Huang, E. Chang and J. Wang, "Automatic Accent Identification Using Gaussian Mixture Models", *IEEE Workshop on Automatic Speech Recognition and Understanding* (ASRU 2001), Madonna di Campiglio, Italy, 9-13 December, 2001, pp.343-346.

[5] B. Ma, D. Zhu and R. Tong, "Chinese Dialect Identification Using Tone Features Based on Pitch Flux", *International Conference on Acoustics, Speech and Signal Processing* (ICASSP 2006), Toulouse, France, 14-19 May 2006, pp.1029-1032.

[6] Y. Liu and P. Fung, "Partial change accent models for accented mandarin speech recognition", *IEEE Workshop on Automatic Speech Recognition and Understanding* (ASRU 2003), Virgin Islands, 30 November – 3 December 2003, pp.111-116.

[7] P. Boersma and D. Weenink, "Praat: doing phonetics by computer", *http://www.praat.org/*

[8] "Approximation of functions and curve fitting", *http://class.htu.cn/shuzhifenxi/chapter3/No3.htm*(in Chinese)

[9] C.C. Chang and C.J. Lin, LIBSVM: a library for support vector machines, 2001. Software available at *http://www.csie.ntu.edu.tw/~cjlin/libsvm*

[10] "HTK - Hidden Markov Model Toolkit", *http://htk.eng.cam.ac.uk/*