Using Cepstral and Prosodic Features for Chinese Accent Identification

Jue HOU, Yi LIU, Thomas Fang ZHENG
Center for Speech and Language Technologies
Division of Technology Innovation and Development,
Tsinghua National Laboratory for Information Science and Technology, Beijing
houj04@mails.tsinghua.edu.cn, {eeyliu, fzheng}@tsinghua.edu.cn

Jesper OLSEN, Jilei TIAN
Nokia Research Center, Beijing
{jesper.olsen, jilei.tian}@nokia.com

Abstract—**In this paper, we propose an approach for Chinese accent identification using both cepstral and prosodic features with gender-dependent model. We exploit a combination of conventional Shifted Delta Cepstrum (SDC) features and pitch contour features as an example of segmental and suprasegmental features, to capture the characteristics in Chinese accents. We use cubic polynomials to estimate the pitch contour segments in order to model the differences within accents. We train gender-dependent GMM acoustic models to express the features in order to deal with the gender variation. Since conventional criterion of the GMM assumption cannot solve those multi-feature problems, we use the support vector machine (SVM) to make the decision. We evaluated the effectiveness of the proposed approach on the 863 Chinese accent database. The result shows that our approach yields a 15.5% relative error rate reduction compared to conventional approaches of using only SDC features.**

Keywords-Chinese accent identification, SVM, multi-layered features, gender-dependent model

I. INTRODUCTION

Automatic accent identification is the task of identifying the speaker’s accent from a given spoken utterance. It is known that the accent causes severe accuracy loss in automatic speech recognition (ASR) systems. A typical method is to build multiple models of smaller accent variances, together with a model selector in the front-end part [1]. Accent identification is able to be used in automated telephone helplines, and first analyzing accent and then directing callers to the appropriately accented response system is able to improve customer comfort and understanding [2].

There are a lot of works on language, dialect and accent identification reported in recent years [1-9]. Segmental and suprasegmental information were commonly used in accent identification. For segmental features, as MFCC is a short-time feature based on frames, [5] proposed the Shifted Delta Cepstrum (SDC) features to provide additional tempo information, and it was used widely in automatic language identification or dialect identification [6-8]. In addition, [1, 8] used the GMM to model accent characteristics of speech due to the fact that GMM training is unsupervised and less computationally expensive than the Hidden Markov Model (HMM) [1]. For the use of suprasegmental information, pitch contour was extracted and estimated using cubic Legendre polynomials in [3]. Also, the pitch flux was adopted as an important role to identify Chinese dialects in [4].

However, there are still challenges in the above approaches of using these two information sources separately, especially for Mandarin Chinese. Most speakers of Mandarin Chinese learned Mandarin (Putonghua) as a second language, and their pronunciations are strongly influenced by their native regional language, or Chinese dialects. Therefore, modeling the segmental and suprasegmental information independently is not sufficient to capture the full diversity of variations. In [7], the cepstral feature and prosodic feature were combined by simply concatenating those two kinds of features to form a vector. Since unvoiced frames do not have pitch information, so that this approach may not perform well on those unvoiced frames for some pronunciations with this type of initials.

In addition, the GMM based decision-making criterion cannot deal with those combined features at different layers. Moreover, as accent and gender are two of the most important factors in speaker variability [1], the gender has a side effect on the accuracy of accent identification. Thus, training gender-dependent models is an effective way in reducing the variability, further leads to a boost on the overall accuracy.

In this paper, we propose an approach of multi-layered feature combination associated with SVM for Chinese accent identification. The conventional SDC based method is associated with pitch contour information as a second feature. Meanwhile, the models are trained separately for each gender so as to reduce the effect of gender variation. We perform the accent identification on all models simultaneously without gender detection to avoiding detection errors. As the original criterion of GMM model cannot deal with such multi-layered features, we utilize the SVM to make the decision. Compared to previous methods, our approach makes full use of features of the accented speech, and it is especially efficient in Chinese accents.

The paper is organized as follows: Section II describes the Chinese accents and challenges in Chinese accent identification. Section III shows how we exploit the pitch contour as a feature. Gender-dependent model and SVM related decision method is

ISBN 978-1-4244-6245-2/10/$26.00 ©2010 IEEE

shown in Section IV. The experimental results are shown in Section V. We conclude our work in Section VI.

II. Chinese Accents

A. *Chinese Accents for Accent Identification*

Chinese characters are ideographic and independent of their pronunciations [10]. The same character may have different pronunciations in different accents. As a result, accent is a severe problem in Chinese, and some of the accents (e.g., Guangdong accent) are quite different in pronunciation from standard Mandarin, which can be regarded as another language. In addition, Chinese is a tone rich language with multiple intonations while some of the European languages are not, such as English and German. The intonations are important information for people to understand the spoken Chinese [4]. It is shown that the main differences among Chinese dialects (also accents) are the prosodic features, such as tone or intonation [4]. Therefore, to utilize the prosodic features is the key issue to raise the accuracy of Chinese accent identification.

B. *Tonal Information in Accent Identification*

We selected several typical accents to establish the baseline system using SDC features. We firstly tried Guangdong and Chongqing as an example. Chongqing accent is relatively similar to standard Mandarin, while Guangdong accent is not, for it is greatly affected by the dialect of Cantonese. Our prior experiment showed the accuracy of identifying Guangdong accent from a mixture of Guangdong, Shanghai, Chongqing and Xiamen accents was over 90%. Moreover, the accuracy of identifying Guangdong accent from Chongqing accent was over 96% using SDC features without any pitch or gender information. Therefore, we finally chose Shanghai instead of Guangdong in the experiments, so that the mixed accents were not remarkably different.

One of the main differences between Shanghai and Chongqing accent is the tone. The Standard Mandarin has four tones, while Shanghai accent has five. The Chongqing accent has four tones, but not identical to those in Mandarin. In Chongqing accent, the first tone is 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.

On segmental features, the Chongqing accent and shanghai accent are highly confusable. The two accents share some differences from standard Mandarin. Both of the two accents do not distinguish alveolar nasal from velar nasal, and retroflex affricate from alveolar affricate. So the conventional MFCC (includes SDC) feature based approaches cannot obtain as high accuracy as that in Guangdong accent.

III. Multi-layered Features

A. *SDC for Accent Identification*

MFCC coefficients provide an efficient means of representing the frequency components of the speech waveform [2]. It has been widely used in ASR and speaker identification applications. The SDC is an extension of the conventional MFCC features, and it is widely used in language, dialect and accent identification [5-9].The SDC feature contains four parameters, N, d, P and k. N is the number of MFCC coefficients of a single frame, d is the time shift for delta calculation. P is the distance between two consecutive blocks, and k is the total number of blocks. The feature vector of time t is combined by concatenating all the blocks of delta vectors. The total length of the feature vector is Nk, as shown in Fig. 1 [6][8].

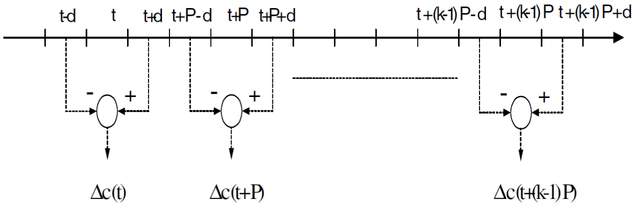

Figure 1. Computation of the SDC feature vector at frame t for parameters N-d-P-k

B. *Pitch Contour Extraction*

We use the tool of Praat [11] to extract the pitch contour of a given spoken utterance. Fig. 2 shows an example of the waveform and the corresponding spectrum with the extracted pitch contour (the real lines on the spectrum). Since the Chinese language is character based and one character is exactly one syllable in most cases, and only voiced frames contain pitch information, so that the pitch contour usually has several segments instead of a continuous curve as a whole. Some connected segments such as the linked segments on the right most, shows how the pitch changes on continuous speech. As a result, it is no need to do any work on partitioning the pitch contour segments. Pitch Contour Estimation using Cubic Polynomials

It is shown in Fig. 2 that the lengths of the segments are not identical. In order to form a feature vector, we use cubic polynomials to approximate those segments. The algorithm is the least square method [12] shown as follows.

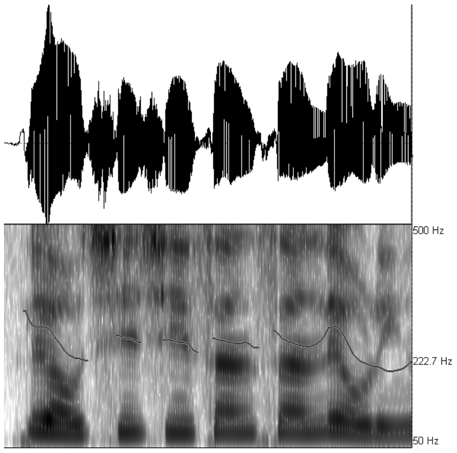

Figure 2. The waveform and pitch contour extracted using Praat

For each given pitch contour segment, we use $\{(x_i, y_i), i = 0, 1, \ldots, m\}$ to denote the pitch values. Here $y_i = f(x_i)$ for $i = 0, 1, \ldots, m$. We try to use a function $y = S^*(x)$ to do the approximation. If we denote the errors on each data point by $\delta_i = S^*(x_i) - y_i$ for $i = 0, 1, \ldots, m$ and form them into a vector $\boldsymbol{\delta} = (\delta_0, \delta_1, \ldots, \delta_m)^T$, then the problem is to find an optimal cubic polynomial $S^*(x)$ in $\psi = span\{1, x, x^2, x^3\}$ so that the square error is minimized, or

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

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 [12]

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

Here $\boldsymbol{a} = (a_0, a_1, a_2, a_3)$, $\boldsymbol{d} = (d_0, d_1, d_2, d_3)^T$ and $\boldsymbol{G} = \begin{bmatrix} (\psi_0,\psi_0) & (\psi_0,\psi_1) & (\psi_0,\psi_2) & (\psi_0,\psi_3) \\ (\psi_1,\psi_0) & (\psi_1,\psi_1) & (\psi_1,\psi_2) & (\psi_1,\psi_3) \\ (\psi_2,\psi_0) & (\psi_2,\psi_1) & (\psi_2,\psi_2) & (\psi_2,\psi_3) \\ (\psi_3,\psi_0) & (\psi_3,\psi_1) & (\psi_3,\psi_2) & (\psi_3,\psi_3) \end{bmatrix}$.

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

IV. GENDER-DEPENDENT MODEL AND DECISION RULES USING SVM

A. Decision Criterion of the GMM Model

In the conventional GMM based approach, suppose that the given test vector is x and the trained GMM model is G, the likelihood score can be calculated through the following function:

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

Here ω_i is the weight for each Gaussian component, and $p_i(x)$ is the likelihood score of the *i*-th Gaussian component, which can be expressed as

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

The decision is to select the model which has the largest likelihood score as the result.

B. The SVM

The support vector machine (SVM) is an efficient tool for data classification. It has been used successfully on a wide variety of tasks such as identifying Arabic and Indian accents of English in [2]. The general process of training and testing in SVM is as follows.

Suppose we're given a training set of instance-label pairs (x_i, y_i) for $i = 1, \ldots, l$, where $x_i \in R^n$ and $y \in \{1, -1\}^l$. On the training process, the training vectors are mapped into a higher dimensional space. The SVM finds a linear hyperplane with the maximal margin in this higher dimensional space. On the testing process, test vectors are then mapped into that same space, and to see on which side of the gap they fall on.

C. Decision on Multi-layered Features using Gender-dependent Models

Since both SDC and pitch features are used in the system, the criterion described in Section IV.A cannot be used directly. The gender variation has a reverse effect on accent identification, we are able to train an SDC feature based GMM model as well as a pitch contour feature based GMM model separately for a certain accent on each gender. After that, for each item of labeled training data, we perform a likelihood score calculation using function (5) on all of the GMM models. As the ranges of scores may vary on different GMMs, we normalize the scores of each GMM model to the interval of $[-1, 1]$, and form the scores into a vector. Then the SVM is trained using those score vectors.

On the testing process, we extract the two kinds of features of a given testing utterance, perform score calculation on all of the pre-trained GMMs and the normalized scores are used to form a vector. (We do not use a gender detection process in order to avoid the detection error, even the gender detection error can be less than 5% according to our prior experiments.) Then use the SVM trained as above to get the accent identification results.

V. EXPERIMENTAL RESULTS

A. System Overview

Our system contains the training and testing parts as shown in Fig. 3. The detailed procedure was described in Section IV.C.

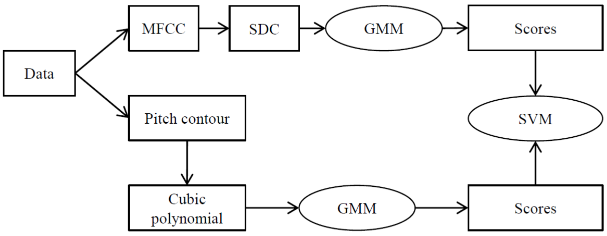

Figure 3. A prototype of the proposed system

B. Resources

We used the 863 Chinese accent corpus to evaluate the effectiveness of the proposed method. The corpus contains 10 regional accents. Each regional accent set contains 100 male speakers and 100 female speakers. Each speaker has 150 utterances. The contents of the utterances vary from isolated words to long sentences, and the lengths are from 3 seconds to over 30 seconds. All speech data are sampled at 16 kHz on 16 bit, WAVE format.

The selected regional accents were Chongqing and Shanghai. The training set contained 170 speakers, 85 male and 85 female for each accent. Each speaker had about randomly selected 120 utterances. The training set was about 86 hours. The testing set contained the remaining 15 male and 15 female speakers with about 80 utterances for each speaker. The four SDC parameters were set to 7, 1, 3 and 7 as in [9]. The frame length was set to 20 ms with 10 ms window step. The pre-emphasis coefficient was set to 0.97. We firstly used the HTK toolkit [13] to extract the 7 dimensional MFCC features with their first derivatives (MFCC_D in HTK format), then folded the delta data to form a 49 (=7*7) dimensional feature vector for each frame.

We further used Praat to extract the pitch contours. The algorithm chosen in Praat was the autocorrelation method. We used the LibSVM [14] for our SVM training and testing tool. We exploited the built-in feature of LibSVM to optimize the parameters as well as kernel function used in training the SVM.

C. Analysis of Experimental Results

1) The baseline

We used HTK to train a GMM model to identify the accents. The baseline system used only SDC features. The decision was simply made by the GMM assumption as described in Section IV.A. The number of GMM components was set to 64 to achieve a good balance between model complexity and accuracy.

The baseline accuracy for Chongqing accent and Shanghai accent are 80.2% and 71.8%, respectively. The overall accuracy is 75.9%.

2) Using only pitch contour features

In this part, we use the 3-dimensonal pitch contour features to identify the two accents. The accuracy of Chongqing accent is 75.8%, while it is 66.6% for Shanghai accent. The overall accuracy is 71.1%.

3) The proposed method

In order to show the effectiveness of our proposed method, we listed the accuracy on all kinds of features, as shown in Table I.

As a result, the overall accuracy of the proposed method is 79.6%, so that the absolute error rate reduction is 3.74% (or15.5% relative) to the baseline system. It is shown from Table I that the integration of pitch information is able to improve the accuracy, and the pitch information is an important issue for accent identification in accented Chinese.

TABLE I. ACCURACY OF ACCENT IDENTIFICATION ON VARIOUS KINDS OF FEATURES

Accent	**Accuracy on various kinds of features (%)**		
	SDC	Pitch	SDC + Pitch
Chongqing	80.2	75.8	78.4
Shanghai	71.8	66.6	80.8
Overall	**75.9**	**71.1**	**79.6**

We can also conclude from Table I that the accuracy of using both features for Chongqing accent is even lower than using SDC feature only. The reason is we are just using the GMM likelihood scores in fusion, which contain only 'distance' measures. In other words, two utterances which have same GMM scores are not necessarily identical in the GMM space. Moreover, the SVM is seeking for an optimal hyperplane to divide the training utterances into two parts at a highest overall accuracy, so the result is to sacrifice the accuracy of Chongqing accent (about 2%) to gain a greater boost (about 9%) on Shanghai accent.

VI. CONCLUSIONS

In this paper, we presented multi-layered features including SDC and pitch contour to model the diversity of variations in accented speech. Compared to conventional accent identification approaches, the use of multi-layered features is able to capture both the segmental and suprasegmental characteristic features of accented speech. The approach of SVM is also applied to combine such two features efficiently without using confidence measure. Furthermore, gender-dependent model is used to alleviate the effect of variations within genders. The effectiveness of the proposed approach was evaluated on Chongqing and Shanghai accents of the 863 accented Chinese corpus. The results showed that our approach achieved a significant 3.74% absolute error rate reduction (15.5% relative) compared to SDC features and 8.54% (29.5% relative) to pitch contour features, respectively.

Our future work includes using more effective approaches to describe the pitch contour accurately, as well as involving other typical features such as rhythm in feature vector, and generating more complicated decision rules.

ACKNOWLEDGMENT

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

REFERENCES

[1] Tao, C., Chao, H., Eric, C. and Jingchun, W., "Automatic accent identification using Gaussian mixture models", IEEE Workshop on Automatic Speech Recognition and Understanding, Madonna di Campiglio, 343-346, 2001.

[2] Carol, P. and Joachim, D., "Accent classification using Support Vector Machines", 6th IEEE/ACIS International Conference on Computer and Information Science, Melbourne, 444-449, 2007.

[3] Chi-Yueh, L. and Hsiao-Chuan, W., “Language identification using pitch contour information”, International Conference on Acoustics, Speech, and Signal Processing, Philadelphia, 601-604, 2005.

[4] Bin, M., Donglai, Z. and Rong, T., “Chinese dialect identification using tone features based on pitch flux”, International Conference on Acoustics, Speech and Signal Processing, Toulouse, 1029-1032, 2006.

[5] B. Bielefeld, "Language identification using shifted delta cepstrum," In Fourteenth Annual Speech Research Symposium, 1994.

[6] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller, Jr., “Approaches to language identification using Gaussian mixture models and shifted delta cepstral features,” in International Conference on Spoken Language Processing, 89-92, 2002.

[7] Bo Yin, Ambikairajah, E., Fang Chen, “Combining cepstral and prosodic features in language identification”, 18th International Conference on Pattern Recognition, 254 – 257, 2006.

[8] Jonathan Lareau, “Application of shifted delta cepstral features for GMM language identification”, Master thesis, Rochester Institute of Technology, October 2006.

[9] Campbell, W. M., Singer, E., Torres-Carrasquillo, P. A., and Reynolds, D. A., “Language recognition with Support Vector Machines,” Proc. Odyssey 2004.

[10] LIU, Y. and Pascale, F., “Partial change accent models for accented mandarin speech recognition”, IEEE Workshop on Automatic Speech Recognition and Understanding, Virgin Islands, 111-116, 2003.

[11] Paul, B. and David, W., “Praat: doing phonetics by computer”, Online: http://www.praat.org/, accessed on 20 July 2010.

[12] “Approximation of functions and curve fitting”, Online: http://class.htu.cn/shuzhifenxi/chapter3/No3.htm (in Chinese), accessed on 20 July 2010.

[13] “HTK - Hidden Markov Model Toolkit”, Online: http://htk.eng.cam.ac.uk/, accessed on 20 July 2010.

[14] Chih-Chung, C. and Chih-Jen, L., “LIBSVM: a library for support vector machines”, Online: http://www.csie.ntu.edu.tw/~cjlin/libsvm, accessed on 20 July 2010.