CHINESE DIALECT IDENTIFICATION USING TONE FEATURES BASED ON PITCH FLUX

Bin MA, Donglai ZHU and Rong TONG

Institute for Infocomm Research, Singapore
{mabin, dzhu, tongrong}@i2r.a-star.edu.sg

ABSTRACT

This paper presents a method to extract tone relevant features based on pitch flux from continuous speech signal. The autocorrelations of two adjacent frames are calculated and the covariance between them is estimated to extract multi-dimensional pitch flux features. These features, together with MFCCs, are modeled in a 2-stream GMM models, and are tested in a 3-dialect identification task for Chinese. The pitch flux features have shown to be very effective in identifying tonal languages with short speech segments. For the test speech segments of 3 seconds, 2-stream model achieves more than 30% error reduction over MFCC-based model.

1. INTRODUCTION

Language identification is a process of determining the language identity of a given spoken query. It is an important technology in many speech processing applications. Besides the information from phonetics and phonotactics, pitch relevant features are also an important factor to discriminate the languages, especially the tonal languages. The pitch relevant features have been successfully used in speaker recognition [1,2], language recognition [3,4], and speech recognition of tonal language [5,6].

Dialect identification is a special case of language identification task. It is more difficult than the general language identification because dialects are highly confusable and with highly overlapping phonetic systems. In this paper, we study Chinese dialect identification [7,8] on the three Chinese dialects, Mandarin, Cantonese and Shanghainese, using Gaussian mixture models (GMM). The main focus here is the effect of pitch relevant information on the Chinese dialect identification task.

Chinese dialects not only share a common written script and vocabulary, but also use similar phonetic systems. However, they are all tonal languages with different numbers of intonations, and the patterns of their intonations are very different. Human perception experience also tells that the prosodic information is very important to tell one dialect from another. We believe that features extracted from pitch information provide discriminative ability among Chinese dialects.

Instead of calculating F0 features explicitly, we extract frame-based multi-dimensional pitch flux features in continuous speech signal. There is no need for the direct pitch detection and no need for voiced/unvoiced decision. It is straightforward to use these multi-dimensional feature vectors in a general speech recognition task of tonal languages, but in this paper we only study its contribution in Chinese dialect identification.

In section 2, we will introduce the motivation of pitch flux and describe the proposed feature extraction algorithm in detail. In section 3, we will show the experimental results on identification of three Chinese dialects and give the discussion in section 4.

2. FEATURE EXTRACTION

2.1. Pitch Flux Features

Assume that a frame of voiced speech signal is modeled as the summation of harmonics

$$x(n) = \sum_{i=1}^{I} \alpha_i \cos(\omega_i n + \varphi_i),\ n = 0,...,N-1\,, \tag{1}$$

where N is the number of samples in the frame, I is the number of harmonics, α_i , ω_i and φ_i are the amplitude, frequency and phase of the ith harmonics respectively.

The autocorrelation of $x(n)$ is

$$R(\tau) = E\{x(n)x(n+\tau)\} \propto \sum_i \alpha_i^2 \cos(\omega_i \tau),\ \tau = 0,...,N-1 \tag{2}$$

For a frame at time t, covariance of autocorrelation with its adjacent frame can be derived as the pitch flux features

$$c_t(d) = E\{[R_t(\tau) - \mu_t(\tau)][R_{t+1}(\tau+d) - \mu_{t+1}(\tau+d)]$$
$$\propto \sum_{\tau,i,j} \alpha_{t,i}^2 \alpha_{t+1,j}^2 \cos((\omega_{t,i} - \omega_{t+1,j})\tau + \omega_{t+1,j} d)\ , \tag{3}$$

where $\mu_t(\tau) = E\{R_t(\tau)\}$ and $d \in (N/2, N/2]$ is the index in the feature vector. The frame shift is usually between 10ms to 20ms. During this short interval, the extent of spectral and pitch flux are subject to the constraint of human's vocal movement. So Eq. (3) can further become

$$\tilde{c}_t(d,\Delta)=\sum_{\tau,i}\alpha_{t,i}^2\alpha_{t+1,i}^2\cos(\delta_i\tau+\omega_{t+1,i}d)\,, \tag{4}$$

where $\delta_i=\omega_{t,i}-\omega_{t+1,i}$, $\Delta=\{\delta_i,\ i=1,...,I\}$.

Differentiating with respect to δ_i, we get

$$\frac{\partial\tilde{c}_t(d,\Delta)}{\partial\delta_i}=-\sum_{\tau}\alpha_{t,i}^2\alpha_{t+1,i}^2\tau\sin(\delta_i\tau+\omega_{t+1,i}d)\,. \tag{5}$$

Since δ_i 's are small values between two adjacent frames, Eq. (5) can be approximated as

$$\frac{\partial\tilde{c}_t(d,\Delta)}{\partial\delta_i}=-\delta_i\sum_{\tau}\alpha_{t,i}^2\alpha_{t+1,i}^2\tau^2\cos(\omega_{t+1,i}d) - \sum_{\tau}\alpha_{t,i}^2\alpha_{t+1,i}^2\tau\sin(\omega_{t+1,i}d) \tag{6}$$

In the case of $d=0$, Eq. (6) can even become as simple as follows

$$\frac{\partial\tilde{c}_t(d,\Delta)}{\partial\delta_i}\propto-\delta_i\,. \tag{7}$$

It shows explicitly the relationship between the dynamics of pitch contour and pitch flux feature curve at $d=0$. If pitch frequency increases, i.e. $\delta_i<0$, the feature curve shows a positive slope, while if the pitch frequency decreases, the slope is negative.

In Eq. (4), we can also see that pitch flux features, at $d\neq0$, are the function of $\omega_{t+1,i}$. Since they do not depend upon the fundamental frequency only, but also depend on the other harmonics as well, we expect that it can work well for telephony speech where lower part of the spectra is normally cut severely due to the distorted bandwidth of telephony channel. The pitch flux feature extraction algorithm is as follows.

Given speech data of two adjacent frames

$x_t(n), x_{t+1}(n),\ n=0,...,N-1$,

the process of pitch flux feature extraction consists of the following steps:

Step 1: Calculate power density spectrum

$$P_t(k)=|DFT\{x_t(n)\}|^2,\ k=0,...,K-1\,. \tag{8}$$

Step 2: Make low frequency emphasis by passing through a low-pass filter

$$\hat{P}_t(k)=P_t(k)\cdot W(k)\,, \tag{9}$$

with

$$W(k)=1+\cos(2\pi k/K)\,,$$

so that the dominant harmonics can be enhanced.

Step 3: Normalize the power density spectrum

$$\bar{P}_t(k)=\frac{\hat{P}_t(k)}{\sum_{k=0}^{K-1}\hat{P}_t(k)}\,. \tag{10}$$

Step 4: Calculate autocorrelation by applying inverse DFT

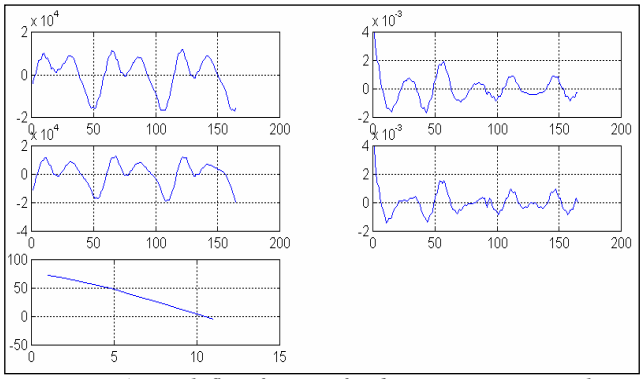

Figure 1a Pitch flux features for decreasing tone speech

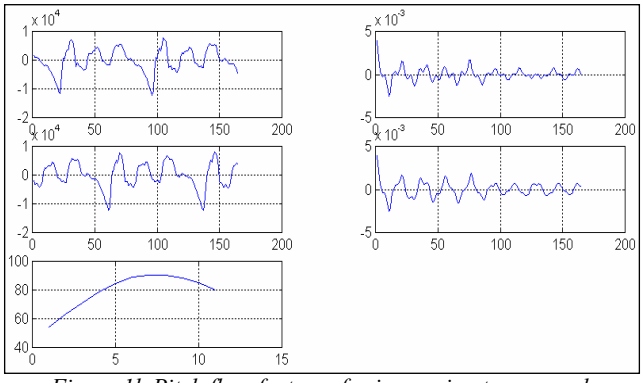

Figure 1b Pitch flux features for increasing tone speech

$$R_t(k)=DFT^{-1}\{\bar{P}_t(k)\}\,. \tag{11}$$

Step 5: Finally define the pitch flux features as follows

$$c(d)=\frac{C}{K-d}\Big[\sum_k R_t(k)\cdot R_{t+1}(k+d) - \frac{1}{K-d}\sum_k R_t(k)\cdot\frac{1}{K-d}\sum_k R_{t+1}(k+d)\Big] \tag{12}$$

where $-D\le d\le D$ is the index in the feature vector, and the dimension of the feature vector is (2D+1); C is a constant to normalize the range of the features, chosen as 10^8.

2.2. Examples

The features introduced in section 2.1 have a clear physical meaning of pitch flux in the speech. Figure 1a and Figure 1b give examples of the feature curves for typical types of voiced speech signals in Mandarin.

Figure 1a shows the 11-dimension features (bottom graph) for the case of decreasing tone speech in Mandarin. D in Eq. (12) is set to 5 and the feature indexes in the graph correspond with d from -5 to 5. The graph at top left corner shows the speech signal of previous frame, and the graph under it shows the speech signal of current frame. The right two graphs show the autocorrelations of the corresponding speech signal on the left side. It can be seen that because the period of the speech signal increases (while pitch decreasing), the covariance of the autocorrelations increase

when the current frame is shifted to left and decrease when the current frame is shifted to right. Figure 1b shows a typical pattern of pitch flux features for the speech with increasing tone in Mandarin.

3. CHINESE DIALECT IDENTIFICATION

We evaluate the effect of the above-mentioned pitch flux features on the identification task of three Chinese dialects, Mandarin, Cantonese and Shanghainese. Chinese dialects are tonal languages and it is generally agreed that there are 5 lexical tones in Mandarin, 9 lexical tones in Cantonese and 5 lexical tones in Shanghainese. These dialects have different patterns of intonations and provide us additional information to discriminate from each other. Our Chinese dialect identification task is designed to make the identification on the three dialects by using GMM modeling with the fusion of pitch flux features and MFCC features. Different durations of testing segments and different amounts of training data for GMM modeling are used to inspect the efficiency of pitch flux features.

3.1. Feature Fusion

For each speech frame, we extract two streams of feature vectors. The first stream is a 39-dimensional feature vector that consists of 12 MFCCs and normalized energy, plus their first and second order derivatives. Sentence-based cepstral mean subtraction is applied to acoustic normalization both in the training and testing. The second stream is an 11-dimension pitch flux feature vector obtained by setting $D=5$ in Eq. (12). Each of the two streams is modeled with a GMM model separately. The total likelihood score is obtained by the fusion of two feature streams.

Let λ_{MFCC} and λ_{PFlux} be GMM models for MFCC feature stream and pitch flux feature stream respectively, the fused log likelihood score of a speech segment $X:\{x_t\}, 1\le t\le T$ is calculated by

$$l=\sum_t [w\log(p(x_t|\lambda_{MFCC}))+(1-w)\log(p(x_t|\lambda_{PFlux}))] \quad (13)$$

where $0\le w\le 1$ is the weighting factor for the two feature streams and can be estimated from development speech data which are different from training data.

3.2. Experiment Setup

The speech data of Mandarin, Cantonese and Shanghainese are collected from telephone line. There are about 10 hours of speech data in each of the three dialects as the training data. In order to have a clear picture on the relationship between the amount of training data and identification accuracy, the GMM models are trained with different amount of training data, 1, 2, 4, 6, 8, and 10 hours. For the evaluation, 1000 speech segments for each of the four durations, 3, 6, 9, 15 seconds respectively are used as test data. For each of the four durations, additional development data of 100 speech segments in each dialect are used to estimate the weighting factor *w*. There is no speaker overlap among the training, development and testing data.

3.3. Experiment Results

GMM modeling with MFCC features for the language identification has been well studied and has been proven to be efficient. Although pitch flux features do not contain as much discriminant information as MFCCs do, they still can provide useful information for the language recognition task, especially for Chinese dialects. In the experiments of this paper, we are interested in how much the proposed features can help at the basis of MFCC features. First we run the experiments on MFCC feature stream and pitch flux feature stream separately to see their individual performance on this dialect identification task. The followed experiments with fusion of the two feature streams will provide us the sufficient knowledge about the efficiency of the pitch flux features at the basis of MFCC features. Figures 2a-2d in the next page show the details. In each the figures, six accuracy curves indicate the different amount of training data.

Figure 2a shows the identification accuracy of three Chinese dialects with GMM modeling on MFCC features only. It shows clearly the relationship between the accuracy and amount of training data, and the relationship between the accuracy and durations of test speech segments. Along the increase of test speech segment durations from 3 seconds to 15 seconds, the accuracy increases almost linearly. More training data can also bring a better accuracy on the identification.

Figure 2b shows the identification accuracy of three Chinese dialects by using pitch flux features only. These tone relevant features provide discriminat information on the three dialect recognition even with a short speech segment of 3 seconds, but the accuraciy is low compared with the accuracy on MFCCs. It also shows that there is no big gain with more training data and longer test speech segments. We can only use such features as useful auxiliary information to help MFFC features to achieve a better performance.

Figure 2c shows the results of adding the pitch flux feature stream at the basis of MFCC stream, and Figure 2d shows relative error reduction, compared with those using MFCC feature stream only. More than 30% error reduction can be obtained for the 3-second test segments with 8 or 10 hours training data. The pitch flux features indeed help a lot on the Chinese dialect identification when the test speech segments are short. When longer testing speech segments are available, the benefit from MFCCs becomes the dominant factor and the gain from pitch flux features is quite limited.

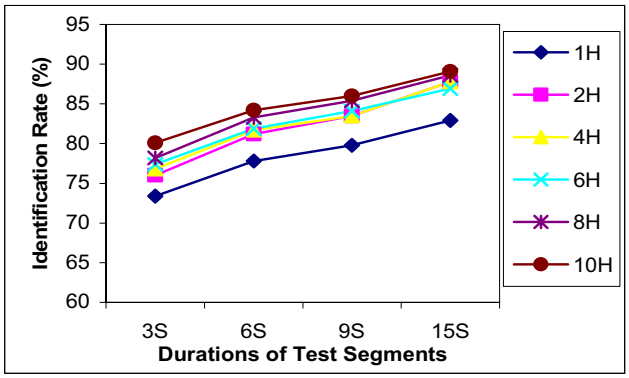

Figure 2a Dialect Identification Rate (%) with MFCC Features Only

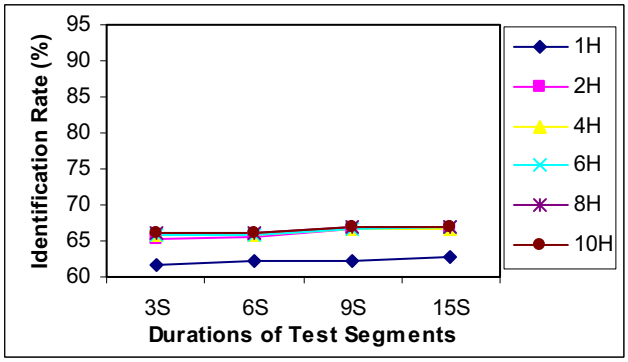

Figure 2b Dialect Identification Rate (%) with Pitch Flux Features Only

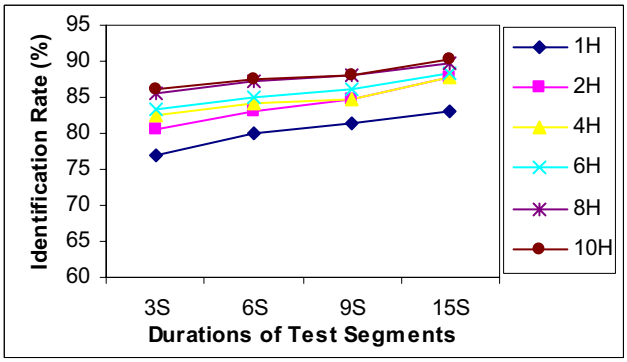

Figure 2c Dialect Identification Rate (%) with both MFCC and Pitch Flux Features

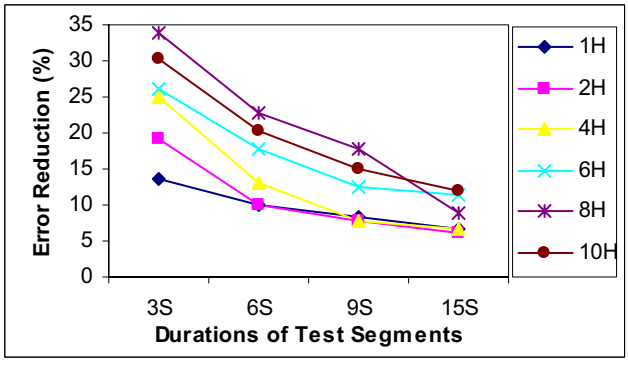

Figure 2d Dialect Identification Error Reduction (%) by Adding Pitch Flux Features at the basis of MFCC Features

4. DISCUSSION

Chinese are tone rich language with multiple intonations. The intonations are important information for people to understand the spoken Chinese. Different Chinese dialects have different numbers of intonations and different patterns of intonations. Better performance on Chinese dialect identification can be achieved by making good use of such kind of discrimination information.

Instead of calculating F0 features explicitly, we extract frame-based multi-dimensional tone relevant features based on the pitch flux in continuous speech signal. Covariance coefficients between the autocorrelations of two adjacent frames are estimated to serve as such features.

These pitch flux features are applied as a separated feature stream to provide additional discriminative information at the basis of MFCC feature stream. Each of two streams is modeled by GMM models of 512 Gaussian mixtures. By fusing the pitch flux feature stream with the MFCC stream, the error rate is reduced by more than 30%, compared with those using MFCC feature stream only, when the test speech segments are as short as 3 seconds. It also shows that the improvement by using pitch flux features will be limited when the test speech segments are long enough.

5. REFERENCES

[1] F. Farahani, P. G. Georgious, and S. S. Narayanan, "Speaker Identification Using Supra-Segmental Pitch Pattern Dynamics," *Proc ICASSP 2004.*

[2] D. A. Reynolds, W. Campbell, T. Gleason, C. Quillen, D. Sturim, P. Torres-Carrasquillo, and A. Adami, "The 2004 MIT Lincoln Laboratory Speaker Recognition System" *Proc ICASSP* 2005.

[3] C.-Y. Lin and H.-C. Wang, "Language Identification Using Pitch Contour Information," *Proc ICASSP 2005.*

[4] T, J, Hazen and V. W. Zue, "Segment-Based Automatic Language Idntification," *Journal of Acoustic Society of America*, 101(4):2323-2331, Apr. 1997.

[5] C.-H. Huang and F. Seide, "Pitch Tracking and Tone Features for Mandarin Speech Recognition," *Proc ICASSP 2000.*

[6] T. Lee, W. Lau, Y. W. Wong and P. C. Ching, "Using Tone Information in Cantonese Continuous Speech Recognition," *ACM Trans. on Asia Language Processing*, Vol. 1, No. 1, March 2002.

[7] W. W. Chang and W. H. Tsai, "Chinese Dialect Identification Using Segmental and Prosodic Features," *Journal of Acoustic Society of America*, 108(4):1906-13, Oct. 2000.

[8] B. P. Lim, H. Li and B. Ma, "Using Local and Global Phonotactic Features in Chinese Dialect Identification", *Proc ICASSP,* 2005.