A Bag-of-phonemes Model for Homeplace Classification of Mandarin Speakers

Hanqing Zhao[1,2](✉), Zengchang Qin[1], Yiyu Wang[2], and Yuxiao Wang[3]

[1] Intelligent Computing and Machine Learning Lab School of ASEE,
Beihang University, Beijing, People's Republic of China
zhq@gmx.com

[2] École d'Ingénieur Généraliste École Centrale de Pékin,
Beihang University, Beijing, People's Republic of China

[3] Department of Chinese Literature College of Liberal Arts,
Fu Jen Catholic University, New Taipei City, Republic of China

Abstract. Mandarin, also known as Standard Chinese is the official language of China and Singapore, there are certain differences when mandarin is spoken by people from different homeplaces. The homeplace classification is important in speech recognition and machine translation. In this paper, we proposed a novel model named Bag-of-phonemes (BOP) for homeplace classification of mandarin speakers, which follows the conceptually similar idea of the Bag-of-words (BOW) model in text processing. The low-level Mel-frequency cepstral coefficients (MFCC) speach features of each homeplace are clustered into a set of codewords referred to as phonemes. With this codebook, each speech signal can be represented by a feature vector of distribution on phonemes. Classical classifiers such as support vector machine (SVM) can be applied for classification. This model is tested by RASC863 database, empirical studies show that the new model has a better performance on the RASC863 database comparing to previous works [1].

Keywords: Bag-of-words · Bag-of-phonemes · Mandarin accents

1 Introduction

The homeplace recognition and classification is for identifying speaker's homeplace by detecting characteristics of their voice (voice biometrics). The elements which decide the characteristics of the speakers' voices include acoustic features (cepstrum), lexical features, prosodic features, languages, channel information, and accents information. According to one's accent, we can judge where he comes from, which means his homeplace. With the rapid growth and development of the society, the needs for homeplace identification spread to several areas. For example, in expert testimony and even in military field, the homeplace identification technology helps change "manpower"into "intelligence". However, existing homeplace identification technology mainly concentrates on the machines which recognize the homeplace mainly by dialects in China. As the popularization of

R. Paredes et al. (Eds.): IbPRIA 2015, LNCS 9117, pp. 683–690, 2015.
DOI: 10.1007/978-3-319-19390-8_76

Mandarin, which is the official language of China and Singapore, accents of Mandarin becomes another way to judge one's original place.

For the subject of the identification of homeplace Gu *et al.* [1] used Gaussian mixture models (GMM) and n-gram language models to produce a global language feature, and makes decision using clustered support vector machine. Hou *et al.* [2] proposed an approach for homeplace identification using both cepstral and prosodic features with gender-dependent model. The identification of homeplace plays an important role not only in China. Malhotra and Khosla [3] discussed text independent and identifies accent among four regional Indian groups spoken Hindi. Teixeira *et al.* [4] proposed an approach using a parallel set of ergodic nets with context independent HMM units to identify six English-spoken countries in Europe. There was also the recognition of hometown based on the prosody of accents. In Gholipour *et al.* [5], prosodic features are used for recognition that including rhythm-related features, global statistics on pitch contour, energy contour and their derivatives.

In this paper, we proposed a novel model named Bag-of-phonemes (BOP) for homeplace identification of mandarin speakers. We deal with sound files in .wav format which is well used in the real-world. The structure of the remainder of the paper is as follows: in Sect. 2, we introduced the basic feature for homeplace identification. In Sect. 3, we fully described the Bag-of-phonemes in details. In Sect. 4, we designed experiments and tested the model. Finally, the conclusions were given in Sect. 5.

2 Voice Features Extraction

In sound processing, the Mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. Mel Frequency Cepstrum Coefficients (MFCC) are coefficients that collectively make up an MFC and such features have been widely used in speaker recognition [11]. The following are the main steps to extract the coefficients from voice or sound.

2.1 Pre-emphasis

The first step of MFCC extracting is to enhance the energy in high frequencies which has been suppressed by certain parts of the human vocal system, such as lips. This step makes the voice signal more smoothly, and improve the accuracy of the model. A high-pass filter is used in this step:

$$S_2(n) = S(n) - \alpha \times S(n-1) \tag{1}$$

where $S(n)$ is the input signal in time domain. And α is a coefficient, in general, we have $\alpha = 0.95$.

2.2 Fast Fourier Transform (FFT)

The voice spectrum changes rapidly in time domain. In order to obtain stable voice features, we extract features from a small window. Each window is determined by three following parameters: the width of window, the offset between successive windows and the shape of window. The piece extracted from a window is called a frame. The frame size β is a number of milliseconds, generally between 10 ms to 30 ms. And the number of milliseconds between the left edge of successive windows is called frame shift, in our experiment, we use $\beta/4$ to avoid information losses.

In order to calculate the energy contains by the signal in different frequency bands, We use FFT to extract the spectral information. The FFT is a improved algorithm of Discrete Fourier Transform (DFT). For each sequence of N complex numbers $x_0, x_1, ..., x_{N-1}$, we transform them into an N-periodic sequence of complex numbers $X_0, X_1, ..., X_{N-1}$, by using this following DFT formula:

$$X_k = \sum_{n=0}^{N-1} x_n \exp(-\frac{2\pi i}{N} nk) \qquad k = 1, 2, ..., N-1 \tag{2}$$

where $x_n$is a windowed signal as the input of DFT, and the output X_K is a complex number representing the magnitude and phase of that frequency component in the input signal. N is the sample length of analysis window, and the sinusoid's frequency is k cycles per N samples, and i is the imaginary unit of complex number.

2.3 Mel Filter Bank

The mapping between frequency in Hertz and Mel scale is linear in low-frequency and logarithmic in high frequency. The Mel scale can be computed from sound frequency by using this following formula:

$$Mel = 2595 \log_{10}(1 + \frac{f}{700}) \tag{3}$$

During MFCC computation, a bank of filter is created to collect energy from each frequency band. The filter bank contains 12 filters which spread linearly in low-frequency and spread logarithmically in high-frequency. As shown in Fig. 1. Finally, we calculate the log of each Mel spectrum value.

2.4 Cepstrum

The last step of extracting the coefficients is to calculate the cepstrum of the log of the Mel spectrum values. We can use Discrete Cosine Transform (DCT) to get the following:

$$Mel_k = \sum_{n=0}^{N-1} x_n \cos[k(n + 0.5)\frac{\pi}{N}]; \quad (0 \leq k \leq N-1) \tag{4}$$

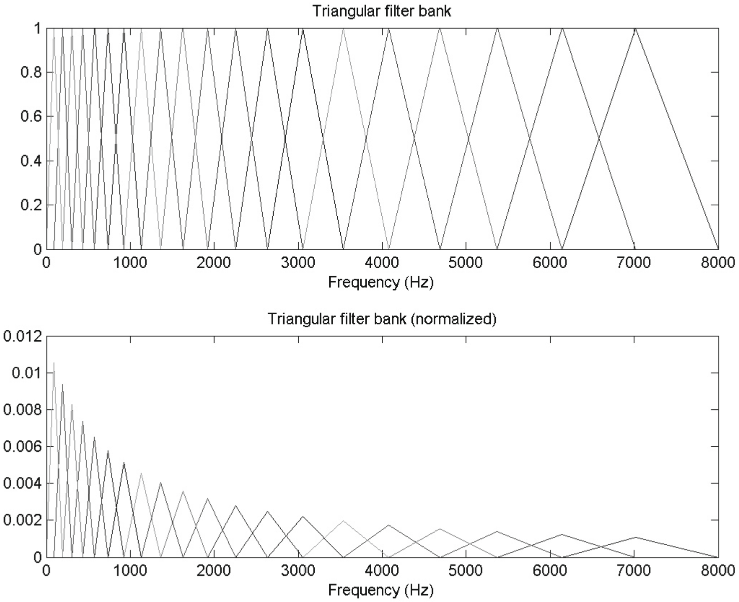

Fig. 1. An example of Mel filter bank.

where N is the number of filters in the filter bank, k is the number of cesptral coefficients. and x_n is formulated as the "log-energy"output of the n-th filter.

3 Bag-of-phonemes Model

The Bag-of-words (BOW) model is a simplified assumption which is well-used in statistical natural language processing and information retrieval (IR) [6]. In this model, a text is regarded as a bag of unordered words, disregarding grammar. The Bag-of-words model is commonly used in document classification, where the (frequency of) occurrence of each word is used as a feature for training a classifier [7,8]. In computer vision, the BOW model is also borrowed and re-invented as the Bag-of-features model, it can be applied to image classification [9,10] by treating image features as "visual words". In document classification, a bag of words is a sparse vector of occurrence counts of words; that is, a sparse histogram over the vocabulary. In computer vision, a bag of visual words is a sparse vector of occurrence counts of a vocabulary of local image features. Some other variants of BOW model were also proposed in music genre classification which named Bag-of tones [11].

Following the similar idea, we propose the Bag-of-phonemes model, in which we treat voice as a document, and the term "words" need to be defined. To achieve this, all files of voice are transformed into a high dimensional space of low-level features (i.e., MFCC in this paper) of each type are respectively clustered to obtain some significant basic units such as topics in text processing and visual words in image processing, in this paper, we call them the phonemes. Each voice file is then can be represented as a distribution of the phonemes. The details are described as the following.

3.1 Feature Description

MFCC feature has been widely used in voice recognition. Because of the good performance of MFCC in speaker recognition, it is employed to transform each voice file into a 12-dimensional matrix. It is like to have 12 channels where the length is related to the original voice file L_v, the sliding window frame length L_f and the window frame increment n. In all, each voice file can be transformed into a $12 \times \frac{L_v}{L_f \times n}$ dimensions matrix.

3.2 Codebook Generation

Each vector represented voice files is named codewords in Bag-of-phonemes (BOP) model, and listed"codewords" is named codebook. In our experiment the codebook is constituted by eight types which are the male and female speakers with homeplaces from four chinese cities: Chongqing, Shanghai, Xiamen and Guangzhou.

In clustering, all the voice files in the training set are mapped into a 12-dimensional space, where each voice is represented as a point in this space. Given the size of the codebook (number of clusters), K-means is used to cluster codewords of 8 types respectively, training codewords of each type is clustered into $k/8$ clusters, where k is the size of codebook and each clusters can be regarded as a codeword or phoneme. Given a voice file in the .wav format, each voice is classified to either of these basic phonemes based on nearest Euclidean distance in this space. The distribution of a voice file on phonemes can be simply calculated using frequency counting. As shown in Fig. 2.

3.3 Classification

Once descriptors has been assigned to cluster to form the feature vectors, we reduce the problem to a multi-classes supervised learning. The classifier performs two separate steps to predict the class of the unlabeled speakers: training and testing. During the training operation, labeled spoken clips are sent to classifier and used to adapt a statistical division procedure to distinguishing categories. In this paper, we chose the SVM classifier to test the Bag-of-phonemes model, and made a comparison of performance with pervious work [1]. The complete process of the Bag-of-phoneme model is shown in Fig. 2.

4 Experimental Studies

In order to verify the performance of the novel model, we use a speaker database named RASC863[1], including 800 speech clips of 800 speakers (400 male speakers, 400 female speakers) with homeplace of four Chinese cities: Chongqing,

[1] http://www.chineseldc.org/doc/CLDC-SPC-2004-003/intro.htm

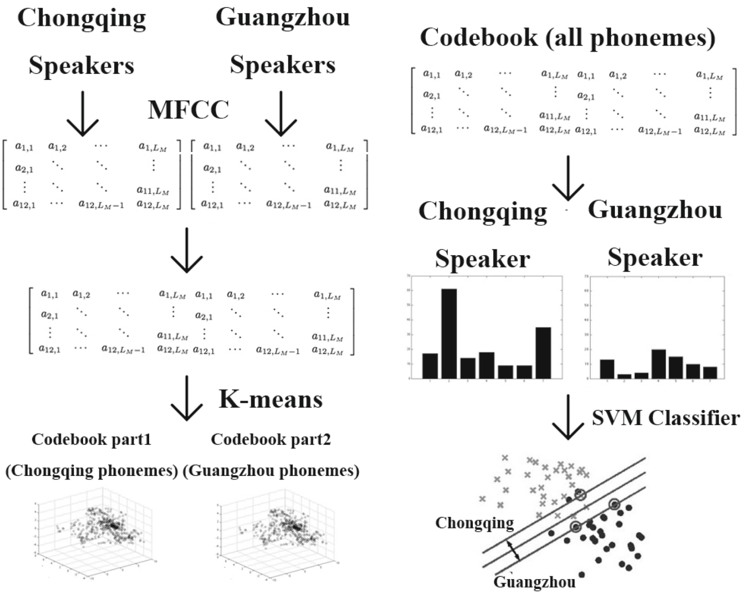

Fig. 2. A illustrative process of the codebook generation (left-hand side) and classification of Mandarin speakers from Chongqing and Guangzhou (right-hand side) by using the BOP model.

Shanghai, Xiamen and Guangzhou, each city has 100 speakers (50 male and 50 female). All the speech clips are mono channel. During the pre-emphasis process, the parameter $\alpha = 0.95$. The window length $\beta = 1024$ points, and the framing shift is 256 points. Each sound clip is transformed into a 12-dimensional matrix.

4.1 Influence of the Size of Codebook

In this paper, the BASC863 database including 800 speech clips in 8 types is used for testing. A 50 % cross-validation is used to validate this model, precisely, we use 400 speech clips as the training and other 400 clips for testing, then we exchange the testing and training sets and run the test again. As a result, the average accuracy with 2 times calculation is regarded as the final result.

In order to study the influence of the length of Codebook, the BOP model tested with the size of Codebook of 800, 720, 696, 680, 640, 600, 560, 504, for a 8-types codebook, its size should be a multiple of 8. And we have also calculated the accuracy with SVM classifier in different kernel functions (linear and polynomial), where k is the number of clusters. When $k = 696$, the SVM classifier with linear function reaches a hit rate of homeplace classification of 63.67 %. The accuracy in different sizes of codebook as shown in Table 1:

Table 1. Accuracy of BOP model with SVM classifier.

k	504	560	600	640	680	696	720	800
Polynomial	57.13 %	56.88 %	56.25 %	56.38 %	56.38 %	55.67 %	55.38 %	55.00 %
Linear	60.50 %	60.75 %	61.50 %	62.63 %	62.13 %	**63.67 %**	61.88 %	61.25 %

4.2 Comparisons to Pervious Works

Given the RASC863 database, the Bag-of-phonemes model is compared to Gu *et al.* [1], which uses the GMM symbolization method for dimensionality reduction and SVM classifier for classification. In this comparison, we used the same testing and training sets and the same calculation method. With 384 Gaussian models, the SVM classifier with polynomial function reaches a hit rate of homeplace classification in 61.75 % (see Table 2). It is obviously lower that the best results we obtained by using the BOP model in Table 1.

Table 2. Accuracy of GMM symbolization model with SVM classifier.

GMM	32	64	96	128	256	384	512
Polynomial	58.88 %	59.88 %	59.75 %	59.88 %	61.00 %	**61.75 %**	61.25 %
Linear	58.75 %	59.88 %	59.25 %	59.63 %	60.00 %	61.50 %	61.00 %

5 Conclusion and Future Work

In this paper, we have presented a simple but novel model named the Bag-of-phonemes for homeplace classification of mandarin speakers. The codebook is built up by clustering the MFCC features of training data. Following the similar idea of Bag-of-words model in natural language processing, each input of voice information is classified according to these basic phonemes (codewords) representations. Classical classifier SVM was used to testify this model. Empirical evidence showed that the new model outperforms GMM on the BASC863 database. In order to testify the effectiveness of the new proposed model, more comprehensive experimental studies on different datasets and more comparisons to other speech recognition models should be considered as the future work.

Acknowledgments. This work is supported by the National Science Foundation of China No. 61305047 and the Chinese National Undergraduate Training Programs for Innovation and Entrepreneurship No. 201410006113.

References

1. Gu, M., Xia, Y., Zhang, C.: Chinese dialect identification using clustered support vector machine. In: Signal and Image Processing, pp. 396–399 (2008)

2. Hou, J., Liu, Y., Zheng, T.F., Olsen, J., Tian, J.: Using cepstral and prosodic features for chinese accent identification. In: Chinese Spoken Language Processing, pp. 177–181 (2010)
3. Malhotra, K., Khosla, A.: Automatic identification of gender and accent in spoken hindi utterances with regional indian accents. In: IEEE Workshop on Spoken Language Technology, pp. 309–312 (2008)
4. Teixeira, C., Trancoso, I., Serralheiro, A.J.: Accent identification. In: International Conference on Spoken Language Proceedings, vol. 3, pp. 1784–1787 (1996)
5. Gholipour, A., Sedaaghi, M.H., Shamsi, M.: The contribution of prosody to the identification of persian regional accents. In: IEEE Symposium on Industrial Electronics and Applications, pp. 346–350 (2012)
6. De Santo, M., Napoletano, P., Pietrosanto, A., Liguori, C., Paciello, V., Polese, F.: Mixed graph of terms: beyond the bags of words representation of a text. In: Hawaii International Conference on System Sciences, pp. 1070–1079 (2012)
7. Zhao, Q., Qin, Z., Wan, T.: What is the basic semantic unit of chinese language? a computational approach based on topic models. In: Kanazawa, M., Kornai, A., Kracht, M., Seki, H. (eds.) MOL 12. LNCS, vol. 6878, pp. 143–157. Springer, Heidelberg (2011)
8. Zhao, Q., Qin, Z., Wan, T.: Topic modeling of chinese language using character-word relations. In: Lu, B.-L., Zhang, L., Kwok, J. (eds.) ICONIP 2011, Part III. LNCS, vol. 7064, pp. 139–147. Springer, Heidelberg (2011)
9. Yuan, X., Yu, J., Qin, Z., Wan, T.: A SIFT-LBP image retrieval model based on bag of features. In: IEEE International Conference on Image Processing (2011)
10. Yu, J., Qin, Z., Wan, T., Zhang, X.: Feature integration analysis of bag-of-features model for image retrieval. Neurocomput. **120**, 355–364 (2013)
11. Qin, Z., Liu, W., Wan, T.: A bag-of-tones model with MFCC features for musical genre classification. In: Motoda, H., Wu, Z., Cao, L., Zaiane, O., Yao, M., Wang, W. (eds.) ADMA 2013, Part I. LNCS, vol. 8346, pp. 564–575. Springer, Heidelberg (2013)