International Journal of Hybrid Information Technology
Vol.8, No. 5 (2015), pp. 303-316
http://dx.doi.org/10.14257/ijhit.2015.8.5.33

Chinese Accent Detection Research Based on Features Structured

[1,]*Zhao YunXue, [2,]*Zhang Long**, [3,]*Zheng ShiJie and [4,]**Zhang Wei

*College of Computer Science and Information Engineering Harbin Normal University Harbin, China
**School of Computer Science and Technology, Harbin Institute of Technology Harbin, China
E-mail: zhaoyunxue_1126@163.com
E-mail: zlwalkman@sina.com
**corresponding author
E-mail: zhengshijie@163.com
E-mail: zhangwei@sina.com

Abstract

Accent is a critically important component of spoken communication, and plays a very important role in spoken communication. In this paper, we conduct accent by using MFCC algorithm and RASTA - PLP algorithm to extract short-time spectrum features of each speech segment based on features structured information. We build short-time spectrum feature sets based on MFCC algorithm and RASTA - PLP algorithm. And we choose NaiveBayes classifier to model the two feature sets. NaiveBayes is to choose the class with maximum posteriori probability as the object's class. This classification method makes full use of the related phonetic features of speech segment. Based on short-time spectrum of MFCC feature set and short-time spectrum of RASTA - PLP feature set respectively achieve 82.1% and 80.8% accent detection accuracy on ASCCD. The experimental results indicate that based on sub-segment splicing feature structured method of MFCC and sub-segment splicing feature structured method of RASTA - PLP can be used in Chinese accent detection study.

Keywords: *accent, feature structured, short-time spectrum features, accent detection*

1. Introduction

The technology of Chinese accent detection is an important part in the field of prosodic features study. Prosodic features include accent, tone, and intonation, *etc.* Firstly, we study the accent in the prosodic features, because the accent plays a very important role in language communication.

Accent belongs to the concept of auditory perceptive category. From the perspective of the listener, the accent mainly indicates those words that sound more prominent than the word or words rounding [1]. A highlight is the prominent degree of sound or syllables in a context than other sound or syllables. Duration, pitch, accent and intensity are factors that affect the relative highlight [2-3]. Accent has a certain level in cognitive performance, such as word, phrase and sentence rhythm. The influence of the accent information loaded by the syllables should be different in diverse ranges. Accent information loaded by some syllables is just inside the prosodic word, and some syllables loading accent information is within the scope of the prosodic phrases or sentences. Respectively, those are called word accent and sentence accent [4-5]. Therefore, the study of accent must be limited within a certain level. This paper limits the accent within word accent.

ISSN: 1738-9968 IJHIT

Thanks to accent, people's language sounds cadenced rather than straightforward way. In addition, the accent avoids ambiguity and strengthens the role of semantic meaning. For example, "I was terrified," accent falls on the "I" and landed on the "terrified", which emphasize the content of the sentence is not the same. Again, for example, "He called you to go yesterday," if the accent falls on the "yesterday" and "you", although the same semantics, but the speaker emphasize content is different.

Chinese accent detection is the analysis of the speech signal processing. It extracts the speech features and sets up a corresponding speech model to determine each syllable. This paper mainly studies the MFCC and RASTA-PLP algorithms and constructs the MFCC and RASTA-PLP sub-segment splicing feature sets.

This paper introduces research situation in the second part. The third part introduces MFCC feature extraction process. The fourth part introduces RASTA - PLP feature extraction process. The fifth part describes sub-segment splicing features structured. It introduces the ASCDD reading text corpus in the sixth part. The seventh part describes experiment environment, and analyzes the experimental results. The eighth part describes the development trend of this research field.

2. Related Research

Review the domestic accent detection technology.Hu Weixiang *et al.*, [4] used duration and pitch acoustic related feature set and the adopt differentiation based on classification and regression tree to detect accent model. The Chinese accent detection accuracy can reach 80% on ASCCD. Shao Yanqiu *et al.*, [1] used neural network with acoustic related features, linguistic features and compounded features to detect Chinese accent. The Chinese accent detection accuracy can reach 78.4%, 83.2% and 84.3%. Chen Nan *et al.*, [6] proposed pitch synchronous features and pitch synchronous peak based on dynamic frame to detect English accent, which used a combination of new features and traditional features. It can decrease error rate of 6.65%. Chen Nan *et al.*, [7] used nonlinear weighted energy features and combined with the characteristics of traditional features to detect English accent. The nonlinear weighted energy features was more robust than the traditional features. The combination with new features and traditional features can decrease error rate of 3.58%. From the perspective of auditory model, Chen Nan *et al.*, [8] also used instantaneous frequency and intensity information of pitch synchronous peak amplitude features to detect accent at the same time. Li Kun *et al.*, [9] used the features of auditory perception with half interval and loudness features, and the normalization of the syllable peak instead of the average. The system accuracy reached 78.7%, and missing rate is 9.37%. On this basis, the model based on the masking was also put forward, and the system accuracy increased to 83.4%. Thus, missing rate dropped to 5.72%. NiChongJia and others [10-11] did the further research for the Chinese accent detecting. They made the use of acoustical relevant features and the grammar and dictionary relevant features to detect accent by boosting integration classification and regression tree. And they used the grammar and dictionary relevant features to detect accent by condition random field constructing a model. And finally, they integrated the Boosting classification and regression tree model and conditional random field model to gain higher recognition rate of hybrid model. It can achieve 76.3% on ASCCD corpus detection accuracy. Li Xin guang *et al.*, [12] studied the English sentence objective evaluation system based on accent and rhythm. It divided English accent sentences by extracting the energy features of speech.

Current detection methods were generally based on the syllable to extract acoustic features and statistical features. This paper presents a structured improvement program for NaiveBayes input feature, which is the short-time

spectrum method based on the features structured. All speech frames are divided into many segments on average for a word. The method is equivalent to that the duration, pitch, initial consonant and vowel divided into many segments. It seems that short-time spectrum feature based on the features structured contains more information, which can reflect the details of speech and detect accent more powerfully.

3. MFCC Feature Extraction

3.1. Mel Frequency Cestrum Coefficient (MFCC)

The human ear has different perception on different frequency of speech. Research shows that there is a linear relationship between awareness and frequency below the frequency of 1000 Hz. And above the frequency of 1000 Hz, there is a logarithm relationship between auditory perception ability and frequency. So people put forward the concept of Mel frequency, its meaning is: 1Mel is 1/1000 of the 1000 Hz tone perception level. Frequency f and Mel frequency conversion formula is $Mel(f) = 2595\lg(1 + f / 700)$ (1) F is frequency, unit: Hz.

3.2. Standard MFCC Extraction Process

MFCC is based on the concept of the Mel frequency, and that its extraction and the calculation process are shown in Figure 1.

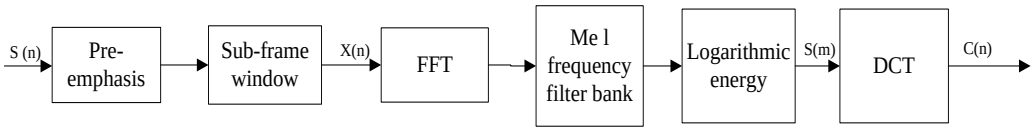

Figure 1. MFCC Extraction Process

Extraction and calculation process is:

(1) The wav voice signal s (n) after pre-emphasis, framing and adding window module processing get the time domain signals of each speech frame x (n). In the pre-emphasis module, this paper adopts digital filter H (z) = 1 - μ z – 1. The value of μ is 0.97.

(2) The time domain signal x (n) after a number of 0 to form long for the sequence of n (n=1024 in this paper), and then after Fourier transform (FFT) modules, it become linear spectrum x (k),

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi nk/N} \quad 0 \le n, k \le N-1 \tag{2}$$

Type: N - points of Fourier transform.

Definition of 20 band-pass filter Hm (k), and M is number of filter. The filter is the triangular filter. Its center frequency is f (M),

$$\begin{cases} H_m(k) = 0 (k < f(m-1)) \\ \dfrac{k - f(m-1)}{f(m) - f(m-1)} (f(m-1) \le k < f(m)) \\ \dfrac{f(m+1) - k}{f(m+1) - f(m)} (f(m) \le k \le f(m+1)) \\ 0 (k > f(m+1)) \end{cases}$$

Mel filter center frequency is defined as $f(m) = \frac{N}{F_s} B^{-1}(B(f_1) + m \frac{B(f_h) - B(f_1)}{M + 1})$

Where: f1, fm - the lowest frequency and the highest frequency of the filter frequency range of applications, Fs - sampling frequency, N - FFT transform points.

Where: $B^{-1}(b) = 700(e^{b/1125} - 1)$

(3) The above (2) linear spectrum X (k) obtained by Mel frequency spectral filter bank module and processed by the logarithmic energy spectrum obtained on the number of S (m).

$$S(m) = \ln[\sum_{k=1}^{N-1} |X(k)|^2 H_m(k)], 0 \leq m < M \quad (3)$$

The Mel frequency filter group set 20 band pass filter H (k) for voice frequency range, each filter has triangular filter property.

(4) The logarithm spectrum S (m) transform cepstrum domain after discretion cosine transforms (DCT). Mel frequency cepstrum coefficient can be obtained.

$$c(n) = \sum_{m=0}^{M-1} S(m) \cos\left\lfloor \frac{\pi n(m + \frac{1}{2})}{M} \right\rfloor, 0 \leq m < M \quad (4)$$

4. RASTA - PLP Feature extraction

4.1. Perceptual Linear Prediction (PLP)

PLP parameter is a kind of characteristic parameters based on auditory model. The feature parameters are a set of coefficients of polynomial. It is equivalent to a kind of LPC (Linear Prediction Coefficient) [13]. The difference is that PLP technology transforms the conclusions into engineering process by using the method of approximate calculation. And PLP technology applies it to the frequency spectrum analysis. The input speech signals treated by auditory model replace traditional LPC analysis using the time domain signal. After the processing, the speech of the spectrum considered auditory characteristic, thus it is advantageous to the speech feature extraction [14].

4.2. Standard PLP Extraction Process

PLP technology mainly imitates the perception mechanism of the auditory in the three levels [15]:
(1) the critical frequency band analysis;
(2) the loudness curve pre-emphasis;
(3) signal strength - auditory loudness transform.
PLP feature extraction step as shown in the figure below.

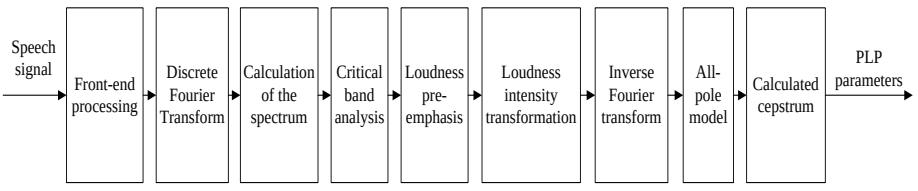

Figure 2. The Extraction Process of PLP

(1) Spectrum analysis

Speech signal after sampling, framing, adding window, and discretion Fourier transform, take a short speech spectrum of real part and imaginary part of the sum of squares, get short time power spectrum $p(f) = R_x[X(f)]^2 + I_m[X(f)]^2$

(2) Critical frequency band analysis

The division of critical band reflects the human ear auditory masking effect and is the embodiment of the human ear auditory model. Using the formula

$$Z(f) = 6\ln\left\{ f/600 + \left[(f/600)^2 + 1 \right]^{0.5} \right\} \quad (2)$$

The frequency of the shaft f of Spectrum f P (f) will be mapped to Bark frequency Z, a total of 17 bands.

This weighting coefficient 17 for each frequency band in the energy spectrum and the formula (3) are multiplied by the summation of the critical bandwidth of the auditory spectrum $\theta(k)$

$$\Psi(Z - Z_0(k)) = \begin{cases} 0, Z - Z_0(K) < -1.3; \\ 10\exp\left[(Z - Z_0(k) + 0.5) \right], -1.3 \le Z - Z_0(K) \le -0.5; \\ 1, -0.5 < Z - Z_0(K) < 0.5; \\ 10\exp\left[-2.5(Z - Z_0(k) - 0.5) \right], 0.5 \le Z - Z_0(K) < 2.5; \\ 0, 2.5 \le Z - Z_0(K). \end{cases} \quad (3)$$

$$\theta(k) = \sum_{Z - Z_0(k) = -1.3}^{2.5} p(f(Z))\Psi(Z - Z_0(k)), k = 1, 2, \ldots, 17. \quad (4)$$

This $Z_0(k)$ represents the k-th critical band center frequency of the auditory spectrum.

(3) Such as loudness pre-emphasis

Using simulated human ear is about 40 dB equal loudness curves E (f) for pre-emphasis. That $\Gamma(k) E\left[f_0(k) \right] \theta(k), (k = 1, 2, \ldots, 17).$ (5)

K denotes a center frequency of the auditory critical band corresponding to a frequency spectrum (in Hz). That is

$$E\left[f_0(k) \right] = \frac{\left(f_0(k)^2 + 1.44 \times 10^6 \right) f_0(k)^4}{\left(f_0(k)^2 + 1.6 \times 10^5 \right)^2 \times \left(f_0(k)^2 + 9.61 \times 10^9 \right)} \quad (6)$$

(4) Loudness intensity transformation

In order to simulate the strength of the nonlinear relationship between the voice and ear feeling of the loudness of strength - loudness is transformation $\Phi(k) = \Gamma(k)^{0.33}$ (7)

After the discretion flourier transformation, 12 orders all pole model is calculated by the Durbin algorithm. And the 16th order cestrum coefficient is also calculated. The end result is the PLP feature parameters.

4.3. RASTA Filter to Improve PLP Parameter

RASTA is a robust method that is not sensitive to the human auditory excitation source for grading by the designed. It can work in the logarithmic domain and cepstral domain. Speech spectrums are filtered through a sequence of speech and inhibit signal slowly varying spectral components from the speech spectrum in order to eliminate the effect of

channel distortion [16]. Perception experiments show, RASTA (Relative-Spectra) is a technology that is able to suppress PLP feature effective channel distortion. Perception also noted that the human auditory perception characteristic can suppress non-stationary linguistic background and voice information and can enhance the change [14]. Therefore, speech analysis method based on the characteristics of auditory perception is in favor of robust speech recognition.

Changes with respect to the transmission channel are constant and slowly vary relative change of the voice. RASTA-PLP technology use the relative stationarity of transmission channel on each PLP in the logarithmic spectral bands and employ a very low frequency of the low-side band to pass filter for filtering processing instead of the usual short-time spectrum. And its high-pass section is conducive to inhibition of convolution noise in the channel, and the low-pass section helps smooth short-time spectral analysis due to changes among frames. We solve RASTA-PLP parameters after the critical band analysis. The auditory critical band spectral value input a band-pass filter. And then this paper extracts PLP for the result. Lastly, we can obtain the final parameters RASTA-PLP. The transfer function for

$$H(z) = 0.1 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4}\left(1 - 0.98z^{-1}\right)} \quad (8)$$

Because PLP parameters have auditory perceptual features, RASTA is to eliminate noise interference technology. So RASTA-PLP speech feature is reliable and effective in speech recognition. And it is higher robust.

5. Sub-Segment Splicing Features Structured

HMM can deal with different timing variables, but NaiveBayes only handle fixed-length input vector. Hence, if we use NaiveBayes to detect accents with length ranging speech segments (number of frames), and it must be converted into a unified dimension features. Common treatment methods of the dimension include re-sampling, based on the average frame, based on the state of stitching and so on. Figure 3 shows the literature [17] using a linear re-sampling feature normalization schematic manner (for example, with "I" word). In fact, this practice only use a part of the speech frame information. Figure 4 shows the frame-based average (Frame-averaged) feature normalization schematic way used in literature [18]. Each frame is directly as the input vector of NaiveBayes feature. And NaiveBayes uses the output of each frame on average. Figure 5 shows schematic diagram based on the state-concatenated warping feature in literature [18]. It uses HMM to determine the state of the further sequence of speech segment. It spliced mean feature vectors of each state corresponding to forming a fixed dimension composite feature vector. To take full advantage of speech segment feature information, this paper introduces features stitching treatment program based on NaiveBayes input requirements and features of the speech signal. The all voice features in the frames are divided into multiple segments based on the sub-segment that each syllable speech frames as a unit. The speech of this subdivision is called short-time spectral features shown in Figure 6. Speech segments of a syllable of the former sub-segment and sub-segment after stitching are shown in Figure 7, Figure 8.

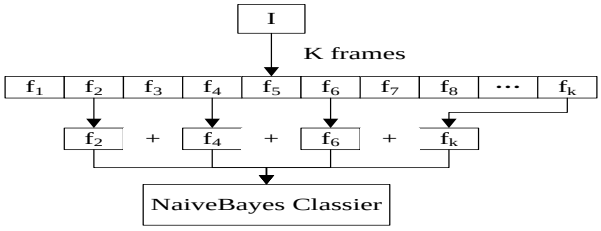

Figure 3. Re-Sampling

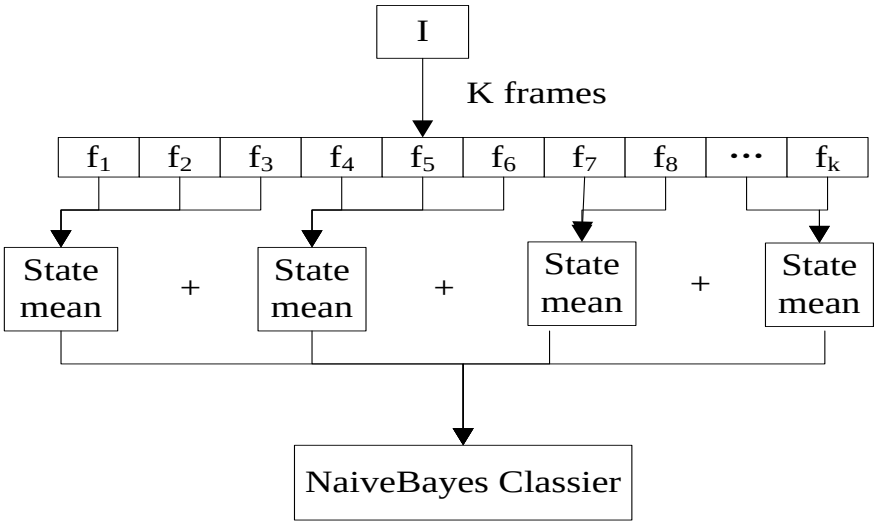

Figure 4. Status Splicing

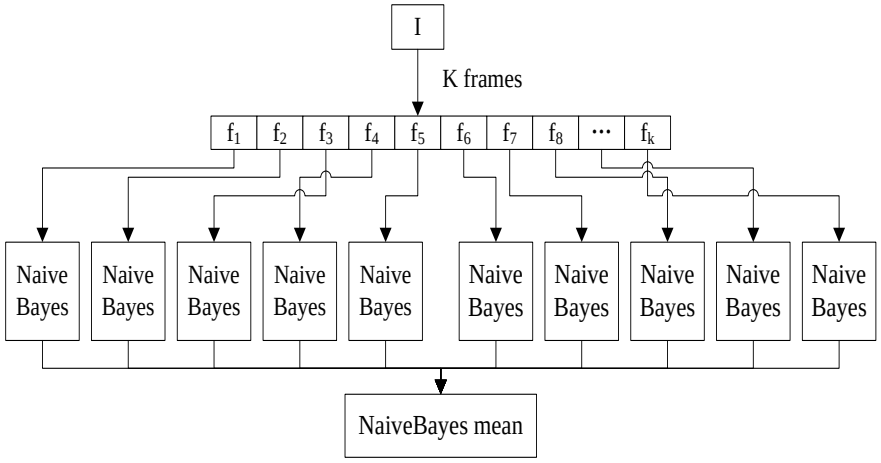

Figure 5. Frame Averaging

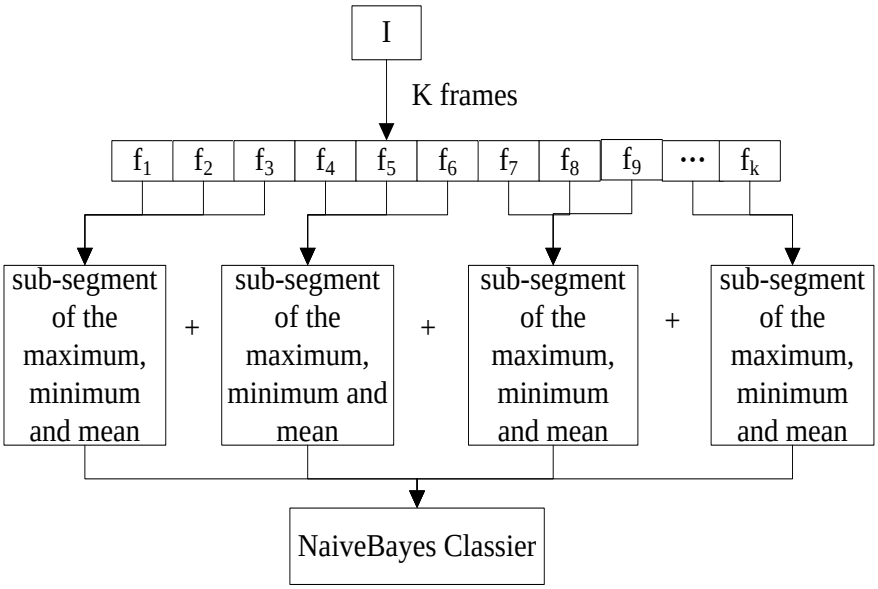

Figure 6. Sub-Segment Stitching

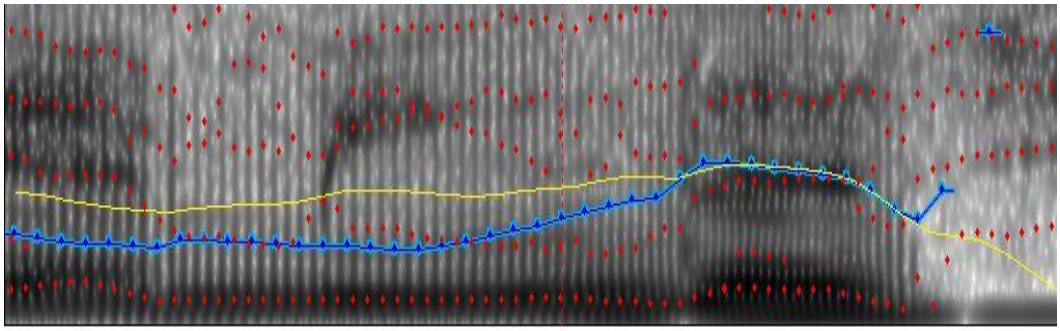

Figure 7. A Complete Speech Syllable Frames

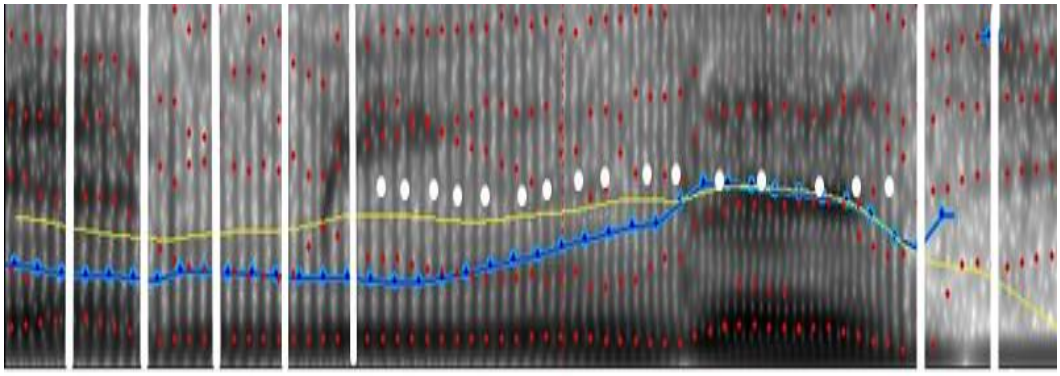

Figure 8. Full Speech Frames a Syllable is Divided into 18 Sections

The timing of the speech determines that we must consider the time change of frequency domain when we conduct the acoustic model for speech. Linear re-sampling scheme described the timing feature to some extent. However, variation characteristic of each phoneme segment frame is not deeply dig. Re-sampling frames can not adequately describe the characteristic properties of the phoneme segment (the number of sampling frames is too few to loss information, and the number of sampling frames is too more to increase the characteristic dimension). The average of the frames does not reflect the changes in the timing of the features due to the direct manipulation program for each speech frame isolated. It is essentially segmental modeling for phoneme segments state stitching method and sub-segments stitching. It strengthened the dynamics of timing described. However, the splicing process is based on the average characteristics of each state of the stitching, which requires prior to work out by the corresponding HMM state sequence, so it increases the amount of computation. Compared with based on the state of stitching, the method based on sub-segment stitching has small amount of calculation, and yet. Fully reflect the syllable internal speech change process.

6. ASCCD

ASCCD applies to the study of language speech, speech engineering development and mandarin teaching. Corpus contains eighteen narrative texts and discussing discourses. Each text contains three to five paragraphs, and each paragraph has five hundred to six hundred syllables. A total of this corpus is nine thousand syllables. There have ten people, five women and five men. They are recorded as M001, M002, M003, M004, M005, F001, F002, F003, F004, and F005. All syllables are annotated. Speech segment uses SAMPA - C standard labeling [19]. Rhythm uses the C - ToBI tagging system. It annotates pinyin, initial and final, tone and marking the syllable prosodic information level and sentence accent boundary [20]. Accent of each unit of rhythmic is divided into level zero, one, two

and three. Select one of the voice file f00101_01.wav and an annotation file f00101_01. TextGrid, with PRAAT software view as shown in Figure 9.

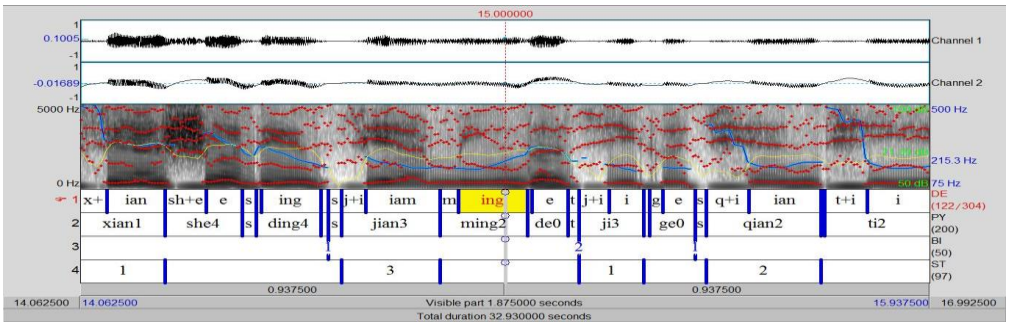

Figure 1. The Corresponding Voice Files and Annotation Instances

Accent has rhythm structure corresponding Chinese hierarchy. The heaviest prosodic syllable weighs the phonetic symbol of one, the secondary prosodic phrase weighs the heaviest syllable the phonetic symbol of two, mainly the heaviest prosodic phrase syllable the phonetic symbol is three, zero indicates unstressed, namely normal pronunciation. In this study, we divide syllables into normal pronunciation and accent. We don't distinguish the differences among them. We consider prosodic word accent and the secondary prosody phrase (MIP) accent as a normal pronunciation, only the main rhythm of phrases (MAP) accent as stressed. ASCCD of accent distribution in the corpus as shown in Table 1.

Table 1. Distribution of Accent in ASCCD Corpus

pronunciation	the number of syllables	the percentage
all	87586	100.00
normal pronunciation	76860	87.75
accent	10726	12.25

7. Test and Analysis of Experimental Results

7.1. Experiment Environments

On Chinese language corpus ASCCD, we choose F001, F002, F003 and F005 four as training set, and select F004 one as testing set. In a sentence level, the size of the training set and testing set is 4:1. On the syllable level training set contains 35060 syllables. On the testing set contains 8761 syllables, including about 964 in the accent syllables. This paper uses machine learning classification methods. This method extracts relevant features from the existing training set, such as acoustic features, based on short-time spectrum of MFCC features, based short-time spectrum of RASTA - PLP features, and optimization fusion features. This paper input these features and trains the model, using the trained model to generate the rhythm in the end. For machine learning method, we firstly choice WEKA NaiveBayes classifier. The classification principle of NaiveBayes classifier is the prior probability through an object using the Bayesian formula to calculate the probabilistic and following which object belongs to a kind of probability. And we choose the class with maximum a posteriori probability to be as [21] the objects belonging to the class and use the WEKA training by default setting.

Short-time spectrum features of speech frames of each syllable for a unit, these frames are divided into many segments every feature on average. The number of speech frames is different for every syllable. But detection on the basis of classifiers needs the same length of input variables. Therefore, we must transform short-time spectrum features of the

different number of speech frames into the uniform features. The dimensions of the treatment methods mainly are re-sampling, based on the frame average, based on state splicing, and based on features structured, *etc.* This paper chose the method based on features structured. For this method reflects the syllables in speech change process. The calculation is less and the speed is much more.

7.2. The Method Based on MFCC Feature

In the short-time spectral features based on MFCC, the recognition rate of accent detection is shown below for different numbers of sub-segment.

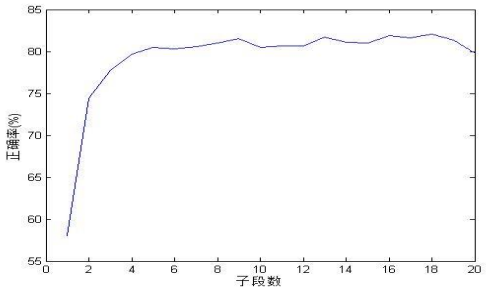

Figure 1. Under Different Segment Number Based on MFCC Feature

From the above chart we can see the short-time spectrum features from five sub-segments to 20 sub-segments that the short-time spectral features of different sub-segments of the recognition rate change is not large. They are maintained at about 80%. When a speech segment is divided into 18 sub-segments, the short-time spectral features the highest recognition rate is 82.1%. In this paper, short-time spectral features of 18 sub-segments extracts MFCC parameters (dimension of 13) in each sub-segment of the maximum, minimum and mean. It builds short-time spectral feature sets based MFCC and makes a total of 13 * 18 * 3 = 702 features.

7.3. The Method Based on RASTA - PLP Feature

In the short-time spectral features based on RASTA - PLP, the recognition rate of accent detection is shown below for different numbers of sub-segment.

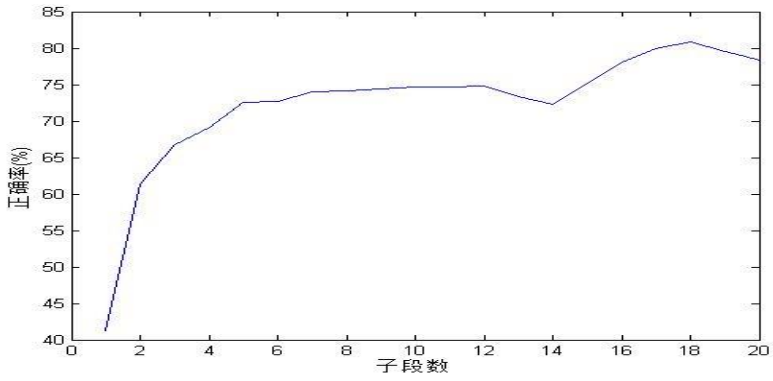

Figure 1. Under Different Segment Number Based on RASTA - PLP Feature

From the above chart we can see the short-time spectrum features from five sub-segments to 20 sub-segments that the short-time spectral features of different sub-segments of the recognition rate change is not large. They are maintained at about

80%. There is a significant decline in 14 sub-segments. It is gradually increased from 15 sub-segments spectrum features to 18 sub-segments short-time spectrum features. The recognition rate in 18 sub-segments reach to the highest of 80.8%. In this paper, short-time spectral features of 18 sub-segments extracts RASTA - PLP parameters (dimension of 9) in each sub-segment of the maximum, minimum and mean. It builds short-time spectral feature sets based RASTA - PLP and makes a total of 9*18*3=486 features.

7.4. MFCC and RASTA - PLP Contrast

MFCC parameters describe the spectral envelope of showing the channel, which it ignores the impact of the pitch frequency. The perceptual linear prediction coefficient (PLP) takes advantage of the psychological awareness. PLP coefficient has more robust than the linear prediction coefficients to noise.

Based on MFCC short-time spectrum features and RASTA - PLP short-time spectrum features divided into 18 are the highest recognition rate. The recognition rate is stationary based on short-time spectral features of MFCC from 5 sub-segments to 20 sub-segments. The recognition rate is some ups and downs based on the RASTA – PLP short-time spectrum features from sub-segments 5 to 20 sub-segments. Known from the analysis of the above, this is a little difference caused by the different of the MFCC and RASTA - PLP.

7.5. The Result of the Experiment and Analysis

Chinese accent detection accuracy

Feature name		Accuracy
MFCC(18 sub-segments)		82.1%
RASTA-PLP(18 sub-segments)		80.8%
literature[4] duration and pitch acoustic related feature sets		80%
literature [1]	acoustic related features	78.4%
	linguistic related features	83.2%
	compounded features	84.3%
literature[10] acoustical relevant features and the grammar and dictionary relevant features		76.3%

In linguistics, each speech contains two aspects: segment information (information constitutes a single phoneme syllables, such as consonants, vowels, *etc.*) and suprasegmental information (from phonemes or phoneme group loads, including audio high intensity, duration and other sound cues, such as tone, accent, *etc.*). Suprasegmental information also called prosodic information. In Chinese, accented syllables and unaccented syllables in pitch, intensity, and speech structure are different. Each pronunciation of Chinese characters is made of consonants and vowels. Consonant

pronunciation part of the time is short, intense signal changes; while the time of the part of vowel pronunciation is long, the signal is relative stable [22-23].

Short-time spectrum features within a syllable concatenation are able to descript duration, intensity, initial and final more detailed of the length of each speech segment. Intensity, consonants and vowels make the distinction between accented and unaccented further characterization, thus it makes the accent more prominent. Additionally, Mel Frequency Cepstral Coefficients MFCC and RASTA-PLP take advantage of the characteristics of the human ear's auditory perception. The combination of features and sound generation mechanism, which better simulate the human ear to the voice signal processing that extracts the features to reflect the unique nature of the auditory system. Therefore, based on the short-time spectrum MFCC feature set and RASTA-PLP feature set can obtain better recognition rate in Chinese accent detection research. The experimental results also confirmed this speculation.

8. Summary and Outlook

This paper describes the short-time spectral MFCC feature set based on features structured and RASTA-PLP feature set based on features structured. Then, the paper uses the classification algorithm NaiveBayes to model on ASCCD with the current short-time spectral features. NaiveBayes select the class with the largest posterior probability so that the object belongs to the class; this classification method makes full use the voice-related features of current voice. Experimental results show that the short-time spectral MFCC feature set based on features structured and RASTA-PLP feature set based on features structured have a high recognition rate. In the future, we will want to simplify the use of features or improve MFCC and RASTA-PLP algorithm. Other we will design to extract a minimum features to get higher recognition rate.

Acknowledgment

Our thanks to supports from the Natural Science Foundation of Heilongjiang Province of China (F201321), the Research and Development Program of Application Technology of Heilongjiang Province (GZ13A003), the Scientific Research Fund of Heilongjiang Provincial Education Department(12541z007), the Scientific Innovation Project for Harbin Normal University (SIP2014001), and the Heilongjiang Provincial Key Laboratory of Intelligence Education and Information Engineering. The authors are grateful for the anonymous reviewers who made constructive comments.

References

[1] S. Yanqiu, H. Jiqing, L. Ting and Z. Yongzhen, "Study on automatic prediction of sentential stress with natural style in Chinese", Acta Acoustical, **(2006)**, vol. 31, no. 3, pp. 203-210.
[2] Y. Jun, "The Phonetic Study of Chinese Rhythmical Words", Journal of Jilin Normal University (Humanities & Social Science Edition), **(2014)**, vol. 2, pp. 36-43.
[3] L. A. Hua, "Comparison of Word Stress between Chinese and English", Hangzhou Normal University, **(2013)**.
[4] H. Weixiang, D. Honghui, T. H. Taiyi, "Study on stress perception in Chinese speech", Journal of Chinese Information Processing, **(2005)**, vol. 19, no. 6, pp. 78-83.
[5] L. C. Ping, "A Prosodic-Syntactic Approach to Resultative VC Constructions", Shanghai International Studies University, **(2014)**.
[6] C. Nan, H. Q. Hua, W. W. Ning and C. R. Yan, "Application of pitch synchronization dynamic frame-length features in English lexical stress detection", Computer Applications, **(2008)**, vol. 6, pp. 1533-1536.
[7] C. Nan and H. Q. Hua, "Application of nonlinear weighting energy features in English lexical stress detection, Acta Acustica, **(2008)**, vol. 6, pp. 520-525.
[8] C. Nan, H. Q. Hua and L. Tao, "Application of Auditory Model-based Feature in English Lexical Stress Detection", Computer Engineering, **(2009)**, vol. 8, pp. 26-30.
[9] L. Kun and L. Jia, "English sentence accent detection based on auditory features", Tsinghua Univ (Sci & Tech), **(2010)**, vol. 4, pp. 613-617.

[10] N. Chongjia, Z. Aiying and L. Wenju, "Mandarin Stress Detection Using Acoustic", Lexical and Syntactic Features, Chinese Journal of Computers, vol. 34, no. 9, **(2011)**, pp.1638-1647.
[11] N. Chongjia, L. Wenju and X. Bo, "Mandarin Stress Detection based complementary model", Computer Enginnering, **(2011)**, vol. 37, no. 23, pp. 20-23.
[12] L. Xinguang, W. Guizhen and Y. Sizhe, "Research on objective evaluation system of English sentences based on stressed syllables and prosody", Computer Engineering and Applications, vol. 49, no. 8, **(2013)**, pp. 105-109.
[13] H. Cheng-wei, Z. Yan, J. Yun, Y. Yin-hua, Z. Li, "A Study on Feature Analysis and Recognition of Practical Speech Emotion", Journal of Electronics & Information Technology, **(2011)**, vol. 1, pp. 112-116.
[14] W.Yan, Z. Xue-ying, "A PLP Speech Feature Extraction Method in Noisy Environment", Jouranal of Taiyuan University of Technology, **(2009)**, vol. 3, pp. 222-224.
[15] W. Yan, "Study of Extraction Algorithm of Improved RASTA-PLP Speech Characteristic Parameters," Taiyuan University of Technology, **(2009)**.
[16] Y. Datao, "Auditory Mechanism Based Robust Feature Extraction and its Application in Speaker Recognition", Harbin Institute of Technogy, **(2014)**.
[17] D. Bin, Z. Qingwei and Y. Yonghong, "Objective evaluation of vowels of standard Chinese Pronuneiation based on formant Pattern," ACTA ACUSTICA, vol. 2, **(2007)**, pp. 122-128.
[18] L. Hong-Yan, H. Shen, W. Shi-Jin, L. Jia-En and X. Bo, "Automatic Mispronunciation Detection for English Learners by GMM-UBM and GLDS-SVM Methods," ACTA AUTOMATICA SINICA, vol. 2, **(2010)**, pp.332-336.
[19] C. Xiao-Xia, L. Ai-Jun, S. Guo–Hua, W. Hua and Y. Zhi-Gang, "An application of SAMPA-C for standard Chinese", Proceedings of the International Conference on Spoken Language Processing, Beijing, China, **(2000)**, pp. 652-655.
[20] L. A. Jun, "Chinese prosody and prosodic labeling of spontaneous speech", Proceedings of the Speech Prosody 200.Aix-en-Provence, France, **(2002)**, pp. 39-46.
[21] L. LinLin, "Automatic Detection of English Speech", Ocean University of China, **(2013)**.
[22] Y. Jue, "Rhythm Patterns in the Speech of Chinese EFL Learner", Zhejiang University, **(2013)**.
[23] S. S. Huang, "A study on prosodic features of Chinese simple introduction sentences spoken by Korean students", Peking University, **(2013)**.

Authors

Zhao Yun Xue, was born in 1986. She is a M.S. at Harbin Normal University. Her research interests include Computer Assisted Language Learning.

Zhang Long, was born in 1978. He is a Ph.D. at Harbin Institute of Technology. He is an associate professor at Harbin Normal University. His research interests include Computer Assisted Language Learning.

Zheng Shi Jie, was born in 1986. She is a M.S. at Harbin Normal University. Her research interests include Computer Assisted Language Learning.