International Journal of Knowledge and Language Processing
Volume 7, Number 2, **2016**

www.ijklp.org
KLP International ©2016 ISSN 2191-2734
pp.1–8

A Research on Automatically Proofreading Tones of Isolated Dialect Words

Wei Wei[1], Xiaohe Chen[1] and Weiguang Qu[2]
[1] School of Chinese Language and Literature
Nanjing Normal University
No.122, Ninghai Road, Gulou District, Nanjing, China
wwei906@163.com, chenxiaohe5209@126.com

[2] School of Computer Science and Technology
Nanjing Normal University
No.122, Ninghai Road, Gulou District, Nanjing, China
wgqu_nj@163.com

Received December 2016; revised December 2016

ABSTRACT. *It is necessary to automatically proofread the missing and improper tones of the isolated dialect words annotated manually in the Audio Database of Chinese Language Resources, using automatic speech recognition technology. In this paper, an automatic method has been investigated to solve the problem. After annotating the tones automatically, the errors can be found from the difference between the automatic annotations and the manual annotations. The result shows a satisfactory result in both accuracy and universal applicability.*
Keywords: Dialect; Isolated words, Tone, Automatic proofreading

1. **Introduction.** China's State Language Work Committee has launched the project of the Audio Database of Chinese Language Resources since October, 2008. It's an extensive and heavy task, especially annotating the tones of isolated words. Although the annotation has been proofread manually for several times, there might also be some errors left. Manual rechecking is a waste of time and effort, so this paper explores the method of automatic proofreading, in order to improve the work efficiency.

Automatic proofreading is an important research field of natural language processing, while the present research focuses more on the field of the text proofreading. It is a relatively new field to apply the technology of automatic proofreading to the construction of the audio database.

It has been pointed out that annotating is an important work in constructing the speech database in the work of [1], and the technology of Automatic Speech Recognition has been applied to speech annotating work, such as the Dutch massive spoken corpora annotation, which saves time and money using the technology of automatic transcription [2]. It's also been noticed that under the influence of accuracy, there is a gap between the present ASR technology in automatic transcription and the manual work, so there should be a post-processing to detect the errors [3].

As for the construction of Chinese dialect audio database, there has been a system which can make the manual work easier [4]. And it also uses ASR to verify some phones [5]. However, it didn't present the efficiency and some practical applications.

2. **Problem Statement and Preliminaries.** The acoustic feature and statistic model are significant for an ASR system. The time domain waveform of speech signal can't be used directly for speech recognition, and it should be transformed to the frequency domain which is a more stable acoustic parameter. The physical basis of tone is mainly the fundamental frequency (F0) of vocal cord vibration, so F0 is widely used in tone recognition, such as the research on mandarin tone recognition [6-7] and Cantonese tone recognition [8-9].

The Mel-Frequency Cepstral Coefficient (MFCC) is a parameter based on perception, which can present the frequency components of acoustic signal in less dimensions. In the study of ASR, MFCC is also useful and it is accepted widely due to its high accuracy, such as the research on mandarin tone recognition [10-11].

State-of-the-art tone recognition systems share similar techniques with most of the ASR research. Some statistical models are commonly used in tone recognition, such as the Gaussian Mixture Model (GMM)[12], the Hidden Markov Model (HMM)[13], the Support Vector Machine (SVM)[14], and the Artificial Neural Network (ANN)[15], etc.

As mentioned above, current ASR technology is mature and also applied to the construction of acoustic database. While in the construction of Chinese dialect acoustic database, the application of ASR is relatively less, and they focus more on only several Chinese dialects, such as mandarin and Cantonese. This paper will introduce an automatic proofreading method that can find out the improper and missing annotations using ASR, which will make the construction of acoustic database more efficient.

The data set used in this paper is the Jiangsu acoustic database, a sub corpus of the Audio Database of Chinese Language Resources. The Jiangsu corpus consists of 70 dialect collection stations, including Jianghuai mandarin, Wu dialect and Zhongyuan mandarin. There are 1000 single words types according to the single words list in the *Chinese Language Resources Investigation Manual*. Each token should be saved into a separate windows PCM format file, named by the ID and word, such as *0001Duo.wav*. The annotation file contains the IPA of the initial, final and the five-degree value of the tone, recorded manually.

Theorem 2.1. Using ASR, the proofreading work is divided into two parts; the first part is the digital signal processing and the second part is the tone verification part.

In the first part, the endpoints of the speech signal are detected and the acoustic parameters are extracted. The commonly used endpoint detection (EPD) algorithm is based on short-time energy, short-time zero crossing rate (ZCR) and the spectral characteristics of acoustic parameters. The former two EPD algorithms are based on the time-domain waveform, of which the calculation is relatively easier and the speed is faster. However, the ZCR algorithm is sensitive to the noise, which leads to some necessary improvement, such as a low-pass filter as a preprocessor and a threshold for ZCR which is threshold-crossing rate (TCR) [16]. On the one hand, EPD can find the endpoints of a signal automatically, which is a precondition of an automatic system. On the other hand, the number of the syllables in the audio file can be determined, which is practical in finding out the missing annotation. As is mentioned above, F0 and MFCC are commonly used in tone recognition system. The following experiment will test whether they are suitable for Chinese dialect tone recognition.

In the second part, ASR technique is used to annotate the tones automatically. Then the different annotation compared with manual annotation will be displayed immediately, from which the possible problems can be found.

Limited by the corpus size and the tone system of a certain dialect, it is not feasible to train models using big data and to get tone models suitable for any tone system. Thus, when proofreading tones of a certain dialect, the models are trained just using the dialect data itself. Then the models can be used to annotate the tones in order to find such a token which is annotated wrongly by human but recognized correctly by our models. In addition, the result is a local optimum, that is to say, even if there were some errors left, when the models were retrained, they would be found eventually.

The typical indicator to evaluate the performance of ASR is the error rate [8-9,17] and another indicator commonly used is accuracy [10-11]. While in our specific automatic proofreading task, not only should the correct annotation be verified, but also the correct tones once annotated wrongly should be recognized.

Definition 2.1. With the specific task considered, the following accuracy P is defined accordingly.

$$P = (a+b)/s \tag{1}$$

It can evaluate the proofreading performance objectively. The factor a denotes the number of tones annotated correctly both by human and computer, the factor b denotes those annotated wrongly by human but recognized correctly by computer, and s is the total number of the tokens.

On the other hand, the proofreading task is also a classification problem, judging whether the transcription is true. Considering the process of manual verification in automatic proofreading, there should be a method to analyzing the efficiency of both automatic and manual effort.

Definition 2.2. As secondary indicators, the accuracy P_1 is defined to evaluate the efficiency of manual effort and the callback rate R is defined to evaluate the performance of automatic work, where E' denotes the number of annotations which are actually wrong when the computer judges that they are wrong, E'' is the number of annotations which computer thinks they are wrong, and E is the number of wrong annotations that are actually exist.

$$P_1 = E'/E'' \tag{2}$$

$$R = E'/E \tag{3}$$

When the proofreading process is repeated, it is necessary to evaluate the whole process.

Definition 2.3. *Considering the whole process, the following cumulative callback rate is used accordingly, where i denotes the times of the proofreading experiment, E_j denotes the number of wrong annotations newly found, while j denotes the times of experiment and $1 \leq j \leq i$, and E is the same as it is in Formula 3.*

$$R_i = \Sigma E_j/E \tag{4}$$

Before the experiment is conducted, the data set has been pre-processed. There are five tones in Nanjing dialect, a falling tone named yinping (41), a rising tone named yangping (24), a low-level tone named shangsheng (22), a mid-level tone named qusheng (44) and a high-level short tone named rusheng (5). Before applying F0 to tone recognition, abstracted by the improved auto-correlation algorithm [18], the tone nuclear should be determined, which is the most characteristic part of a tone, while the other parts are tone head and tone tail, accounting for 10%-20% of the whole tone [19]. To ensure the accuracy, the boundaries of the speech and non-speech and that of the initial and final were annotated manually. And to get the tone nuclear, 15% of the finals at the beginning and end were removed, because the pitch contour of each tone is straight. Moreover, a three-dimension feature vector is constructed to denote the tones, which consists of the mean of the F0, the first order difference of F0 and the duration. Then GMM is used as a classifier.

MFCC is adequate for syllable recognizer, that is to say, it is reasonable to apply MFCC to tone recognizer by syllable. Thus, the boundaries can be annotated automatically by the EPD algorithm. There are two significant parameters in an HMM, the number of the Gaussian models in a GMM, represented by *m*, and the number of states of an HMM, represented by *n*. So the short form of a model is like HMM(*m*,*n*).

In the Nanjing dialect corpus, there have been 1095 isolated words annotated before. But the total number is actually 1096, when the missing annotation is found by both manual boundaries annotation and automatic EDP.

3. **Main Results.** Here are the main experiments and results in this paper.

Example 3.1. Let us consider the following example. After the tones recognized respectively by GMM with F0 and HMM with MFCC, two confusion matrices are got as shown in Table 1 and Table 2.

TABLE 1. CONFUSION MATRIX USING GMM WITH F0

	41	*24*	*22*	*44*	*5*
41	*204*	*0*	*0*	*5*	*1*
24	*0*	*193*	*3*	*0*	*0*
22	*1*	*1*	*155*	*8*	*0*
44	*7*	*0*	*21*	*246*	*14*
5	*3*	*0*	*1*	*30*	*202*

In Table 1, there are 95 tokens confused, from which the wrong annotation of the second pronunciation of *0041Jin* is found, annotated as 41 manually, while in Table 2, there are only 15 tokens confused and the existing wrong annotation is also found. More precisely, the actual tone *44* is recognized correctly by both the two experiments.

TABLE 2. CONFUSION MATRIX USING HMM WITH MFCC

	41	*24*	*22*	*44*	*5*
41	*207*	*0*	*2*	*1*	*0*
24	*0*	*188*	*3*	*4*	*1*
22	*0*	*1*	*164*	*0*	*0*
44	*0*	*1*	*1*	*286*	*0*
5	*0*	*0*	*1*	*0*	*235*

The confusion matrices do not reflect the performance of proofreading missing annotation. Additionally, the missing annotation, the second pronunciation of *0835Sheng*, is recognized properly as *22* in both the experiments. In conclusion, the evaluating result is shown in Table 3.

TABLE 3. THE EVALUATING RESULT

Number	*Parameter*	*Model*	*P*	P_1	*R*
1	*F0*	*GMM*	*91.33%*	*2.08%*	*100%*
2	*MFCC*	*HMM(5,8)*	*98.72%*	*12.50%*	*100%*

Comparing the two experiments, it is obviously shown that the HMM with MFCC is more efficient than GMM with F0. Even the tone nuclear is not tagged explicitly in Experiment 2, recognizing tones by syllable is more precise owing to the HMM and MFCC, which not only can make the process more automatic, but also make the proofreading method more universal. Therefore, MFCC and HMM are the parameter and statistical model finally chosen.

In addition, though the value of P_1 is extremely low, the actual workload is small. Taking Experiment 2 as example, there are only 16 tokens to be rechecked. While evaluated by P, the total number of different dialect is almost the same, which is convenient to compare the results of different dialect. Thus, P is main indicator in the following experiment.

Example 3.2. In the two real experiments above, there are only two representative mistakes, a missing annotation and a wrong annotation. In fact, the mistakes rarely exist, which makes the proofreading task something like looking for a needle in the ocean.

However, it is still reasonable to predict that there might be some more mistakes, which challenges the result of the above two experiment. Therefore, the following artificial mistakes are made in Table 4, simulating a possible situation.

First of all, 20 artificial mistakes are made, chosen randomly from the annotation file by computer and numbered 1 to 20. Then, the tones are modified randomly, too. The artificial mistakes are shown in Table 4. It can be analyzed that the artificial mistake is typical, since the wrong tones contain both the same and different F0 contour with the right tones.

Then, with the parameters adjusted properly, the automatic proofreading is repeated, by which all the mistakes are found finally. The process and result is shown in Table 5. In the process of auto-proofreading, HMM(5,8) is used for the first four times. When the rate of calling back is lower, the parameters should be adjusted relatively lower. Thus, HMM(5,5) is available, even if the accuracy is a little bit lower accordingly.

TABLE 4. ARTIFICIAL TONE MISTAKES

ID	*Tokens*	*No.*	*Real Tone*	*Artificial Mistakes*
(1)	*0061Bu*	*1*	*22*	*44*
(2)	*0094Qu*	*2*	*24*	*44*
(3)	*0223Shi*	*1*	*24*	*41*
(4)	*0237Shi*	*1*	*24*	*22*
(5)	*0267Fei*	*1*	*44*	*41*
(6)	*0270Wei*	*1*	*44*	*22*
(7)	*0304Zhao*	*1*	*44*	*22*
(8)	*0335Dou*	*1*	*41*	*44*
(9)	*0353Xiu*	*2*	*41*	*44*
(10)	*0365Jiu*	*1*	*41*	*44*
(11)	*0392Gan*	*1*	*44*	*41*
(12)	*0437Fan*	*1*	*41*	*44*
(13)	*0458Ji*	*1*	*44*	*5*
(14)	*0478Ge*	*1*	*44*	*5*
(15)	*0608Wan*	*1*	*41*	*44*
(16)	*0634Chen*	*1*	*44*	*24*
(17)	*0783Jiang*	*1*	*44*	*22*
(18)	*0795Deng*	*1*	*44*	*22*
(19)	*0822Ce*	*1*	*44*	*5*
(20)	*0910Heng*	*1*	*44*	*24*

There is no doubt that some tokens would be verified repeatedly if the repetitive tokens were not removed. So there is a filter in the proofreading system which can ignore the tokens presented before. The only tokens should be rechecked is those newly found. For example, in the second proofreading in Table 5, the fraction *13/22* means that there are 22 tokens to be rechecked, in which 9 tokens have been verified in the first time, so only 13 tokens should be rechecked this time. After the sixth proofreading, there are only 139 tokens rechecked, in which the total 22 mistakes are found eventually. That is to say, only 13% of the total tokens are rechecked, which makes the proofreading obviously more efficient.

TABLE 5. THE RESULT OF PROOFREADING ARTIFICIAL MISTAKES

ith	*Model*	*Numbers Verified*	*P*	*R*	*Mistakes Newly Found*	ΣE_j	R_i
1	*HMM(5,8)*	*21/21*	*97.35%*	*31.82%*	*(12)(16)(17) (18)(19)(20) 0835Sheng(No.2)*	*7*	*31.82%*
2	*HMM(5,8)*	*13/22*	*97.26%*	*36.36%*	*(5)(6)(11) (13)*	*11*	*50.00%*
3	*HMM(5,8)*	*13/21*	*97.17%*	*31.82%*	*(1)(3)(14)*	*14*	*63.64%*
4	*HMM(5,8)*	*23/35*	*96.35%*	*45.45%*	*0041Jin(No.2)*	*15*	*68.18%*
5	*HMM(5,5)*	*46/73*	*93.52%*	*63.64%*	*(4)(7)(9) (10)(15)*	*20*	*90.91%*
6	*HMM(5,5)*	*23/53*	*95.35%*	*54.55%*	*(2)(8)*	*22*	*100.00%*

Example 3.3. After the exploring experiments, a practical test is conducted as follows. With the data set of 70 dialects integrated, 72059 tokens are detected automatically by the algorithm of EPD, while there are 71954 tokens in fact after the manual verification. So the EPD algorithm reaches an accuracy of 99.85% in the experiment. The main causes leading to the errors are the strong noises in the signals, which are a little bit higher than the threshold, and the creaky voice in the syllables, which is quite low just like a silence period.

After all the proofreading tests on the other 69 dialects, the average accuracy of HMM(5,8) on 70 dialects is 97.79%, of which the minimum is 92.47% and the maximum is 99.90%. The results show that our proofreading system is of a high accuracy and a suitable universal applicability.

4. **Conclusion.** An automatic method to proofread the tones has been investigated in this paper, aiming at the missing and wrong annotation and finally an automatic tone proofreading system has been developed for future applications. First, with an accuracy of 99.85, the EPD algorithm of TCR is available for the proofreading task. Next, the model trained by the data set itself is effective, with an average accuracy of 97.79%, to find the wrong annotation after annotating the tones automatically. Moreover, HMM with MFCC is universal in ASR system, which ensures the universal applicability of our system.

Acknowledgment. This work is funded by the Project of the Research and Development of Acoustic Database Technical Tools (Project Number: 2014BAK04B02). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] M. A. Pitt, K. Johnson, E. Hume, S. Kiesling and W. Raymond. The Buckeye Corpus of Conversational Speech: Labeling Conventions and a Test of Transcriber Reliability. *Speech Communication*, vol. 45, no. 1, pp. 89-95, 2005.

[2] C. V. Bael and L. Boves. Automatic Phonetic Transcription of Large Speech Corpora. *Computer Speech & Language*, vol. 21, no. 4, pp. 652-668, 2007.

[3] K. Voll, S. Atkins, B. Forster. Improving the Utility of Speech Recognition through Error Detection. *Journal of Digital Imaging*, vol. 21, no. 4, pp. 371-377, 2007.

[4] X. Han, L. Li, W. Pan. Computer Based Field Investigation and Processing System for Languages. *Journal of Tsinghua University*, vol. 53, no. 6, pp. 888-892, 2013.

[5] M. Yan. The Integration and Interactive Platform of Network Information Resources of Linguistics in Big Data Era. Shanghai: Shanghai Normal University dissertation, 2014.

[6] Q. Liu, J. Wang, M. Wang, P. Jiang, X. Yang. A Pitch Smoothing Method for Mandarin Tone Recognition. *International Journal of Signal Processing Image Processing & Pattern Recognition*, vol. 6, no. 4, pp. 245-253, 2013.

[7] H. Huang, J. Zhu. Tone Modeling Based on Discriminative Training for Mandarin Speech Recognition, *Computer Engineering and Applications*, vol. 45, no. 11, pp. 178-182, 2009.

[8] M. Emonts, D. Lonsdale. A Memory-based Approach to Cantonese Tone Recognition, *Paper presented in 8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003.

[9] Y. Qian, T. Lee, F. K. Soong. Tone Recognition in Continuous Cantonese Speech using Supratone Models. *Journal of the Acoustical Society of America*, vol. 121, no. 5, pp. 2936-2945, 2007.

[10] Y. Tian, J. L. Zhou, M. Chu, E. Chang. Tone Recognition with Fractionized Models and Outlined Features. *Paper presented in 29th IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Quebec, Canada, 2004.

[11] H. Xiao, C. Cai. Study of Speaker-Independent Tone Recognition Based on Support Vector Machine. *Computer Engineering and Applications*, vol. 45, no. 9, pp. 174-176, 2009.

[12] Z. Xu, F. Yu. Chinese Tone Recognition Based on SPWD Time-Frequency Ridge Feature Extraction, *Computer Applications and Software*, vol. 31, no. 3, pp. 142-145, 2014.

[13] Z. Li. HMM Based Recognition of Chinese Tones in Continuous Speech. *Journal of Electronics*, vol. 17, no. 1, 2000.

[14] D. Fu, S. Li, S. Wang. Tone Recognition Based on Support Vector Machine in Continuous Mandarin Chinese. *Computer Science*, vol. 37, no. 5, pp. 228-230, 2010.

[15] F. Sun, G. Hu. Chinese Tones Recognition Based on a New Neural Network. *Journal of Shanghai Jiaotong University*, vol. 31, no. 5, pp. 36-38, 1997.

[16] D. Xi, R. Li, H. Chen. A Speech Endpoint Detection Algorithm Based on Maximum of Auto-correlation Function and Amended Threshold-crossing Rate. *Audio Engineering*, vol. 34, no. 4, pp. 53-57, 2010.

[17] J. Zang. *Fundamental of Chinese Man-Machine Voice Communication*. Shanghai Science and Technology press, Shanghai, 2010.

[18] P. Boersma. Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-To-Noise Ratio of a Sampled Sound. *Ifa Proceedings*, vol 17, pp. 97-110, 2000.

[19] X. Zhu. *Phonetics*. Commercial Press, Beijing, 2010.