The Analysis of Influence Factors and Identification of Speaker-Dependent Primi Speech Recognition

Lin Guo, Yang Bai, Jie Su, Wen-lin Pan* and Tian-jun Zhang
Yunnan Minzu University, Kunming, 650500, Yunnan Province PR China
*Corresponding author

***Abstract*—The research object is an endangered minority language of Primi on the southern dialect. This paper takes isolated-word of Primi as the primitive and uses large vocabulary of speech corpora based on HTK. The effect of different quantity of vocabularies, the sample ratio of training and testing, and the data quality on recognition rate of speaker-dependent were investigated. In addition, the identification of speaker-dependent acoustic model were compared. The experimental results show that the high quality data can improve the recognition rate and has little effect on the recognition rate in the vocabulary size.**

Keywords-isolated-word; primi language; HMM; speaker-dependent speech recognition

I. INTRODUCTION

At present, there are a lot of researches on the speaker-dependent speech recognition. The majority of literatures are concerned with the dominated languages such as Chinese and English. This technology has been applied to smart products, such as the car controlled by voice, navigation map, and speech code of credit card, etc. But, as far as our knowledge, few researches have reported that the endangered language of Primi is applied on speech recognition.

Some characteristics studies of Chinese language about speaker-dependent speech recognition are as follows. Zhao solved problems of complex algorithm and the large storage space based on improved algorithm of dynamic time warping (DTW), that the rate of recognition arrived at more than 95% [1]. Wei analyzed the characteristic of speech parameters, and proposed the entropy sequence that not only reduce the amount of computation, but also represent the speech signal [2, 3]. Peng used the methods of global constraints based on DTW to recognize different backgrounds speeches, that the average recognition rate of three people arrived 96% in a relatively quiet condition [4]. Sun used hidden Markov model (HMM) to investigate the effect of speech training times and gender influence on recognition rate. And the result shows iteration three times superior to that of the previous two, that speech recognition accuracy rate can reach above 92% [5].

In the study of minority speech recognition of speaker-dependent, Li established acoustic model based on phonemes and semi-syllables and trained large-vocabulary continuous speech of Lhasa Tibetan on HTK platform, that the recognition rate is 92.2% [6]. Yao has improved the traditional method of endpoint detection, and the method increased the isolated-word speech recognition rate of Tibetan specific [7]. Nuo wrote a recognition tool of Uighur speech based on DTW, that the rate of speaker-dependent achieves 96.4% [8].

In order to satisfy people's demand for personalized products, the technology of speech recognition eagerly needs to solve problems, like multi-language, wide vocabulary and the rate of recognition [9]. The existing researches improve the ratio of speech recognition from many aspects, for instance, training model, background, feature parameters, training times, the recognition of primitive, and the endpoint of speech. However, little research focuses the recognition ratio on the effect of training method and speech quality.

Our research object is an endangered minority language of Primi on the southern dialect (Qinghua language). Three effects on recognition rate of speaker-dependent were investigated, include different vocabularies quantity, the sample ratio of training and testing, and data quality. In addition, using different speakers' speech corpora, the identification of two speaker-dependent acoustic models were studied.

II. EXPERIMENTAL METHOD

A. Acquisition and Pretreatment of Speech Corpus

About twenty thousand people of Primi lived in Lanping, Yongsheng, Ninglang, Weixi and Lijiang in Yunnan Province. Furthermore, Sichuan Province has more than 20,000 people. Primi dialect belongs to Qiangic of Tibeto-Burman language of Sino-Tibetan. In this paper, Qinghua language of the southern dialect was selected as the research object [10, 11].

Field investigations and recordings were conducted in Yu-Shi-Chang, Qinghua village, Lanping county, in July 2014. Nearly 20,000 speeches were recorded by a number of male and female speakers, which used 1,009 common words of local language as vocabulary. A tool of endpoint detection and semantic annotation was developed to dispose large original corpus. It can generate available speech corpus for HTK through automatic speech segmentation. On the basis of the above work, we have built a corpus of Primi language for the author affiliated team to share.

Articulate speech corpora from the speech corpus spoken by two men and two women were analyzed in this paper. A total of 1,009 words were recorded, and each word recorded eight times, which composed 4×8×1009 speeches.

B. Construction of Recognition Model

The speech sample was taken isolated-word of Primi as recognition unite, which was divided into training set and test set. The sample adopted Mel Frequency Cepstrum Coefficient (MFCC) of 39 dimensions as acoustic feature. Taking into account that DTW is only suitable for small vocabularies, we selected HMM to build a training set model, and used Viterbi algorithm to recognize the training set respectively. The process of speech recognition is shown in Figure I.

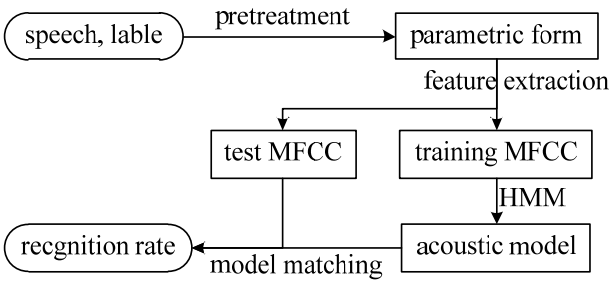

FIGURE I. SPEECH RECOGNITION PROCESS

C. Experimental Tools

The open source HTK toolkit, using the standard C programming language, is developed by the University of Cambridge to establish and process HMM. As far as our knowledge, the operation of HTK assembly is cumbersome because of composing by many standalone routine. In addition, it needs to consider the enormous quantity of speech in this paper. Hence, in order to simplify the speech recognition process based on HTK, the "HTK-Assistant" tool was developed to achieve the batch training and identification in Figure II. The manpower can be greatly reduced by using HTK-Assistant because relevant configuration files can be automatically created.

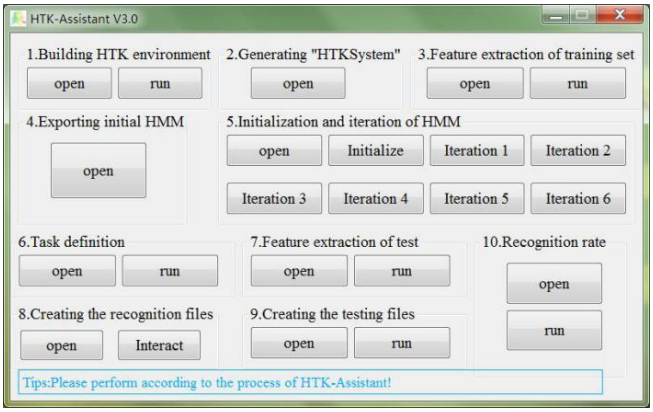

FIGURE II. THE INTERFACE OF HTK-ASSISTANT

III. The Experiment Analysis of Recognition Rate

In this paper, multi experimental groups were designed in order to investigate three effects, include different quantity of vocabularies, the ratio of training and testing sample, and data quality of speaker-dependent recognition rate. In this experiment, seven groups were carried out, including 2×4×1009 Primi speeches that the data pronounced by male A_1 and female B_1. The step length of each group is defined in (1). Taking into account the size of vocabularies, the first 500 words use 50 as a step, and the rest words' step is 100. Let n be the experiment number of each group and then N be the total number of words. The first six groups total number was N=1009.

$$I = \begin{cases} 50n & ,1 \le n \le 10 \\ 500 + 100(n-10), & 11 \le n \le 15 \\ N & ,n = 16 \end{cases} \tag{1}$$

The total sample size of each experimental group is defined as

$$S = 8I \tag{2}$$

The rate of speech recognition is defined as

$$R = \frac{c}{t} \tag{3}$$

where c is the correct number and t is a total number in each speech recognition group.

Six groups were designed according to factors of the incremental words, the ratio of training set and test set, different speakers. The data of experimental groups and the average recognition rate are shown in Table I. Experimental results are shown in Figure III.

TABLE I. SIX EXPERIMENTAL GROUPS AND AVERAGE RECOGNITION RATE

speakers	experimental groups	ratio of training set and test set	average recognition rate (%)
A_1	1	5:3	97.87
	2	6:2	97.31
	3	7:1	96.27
B_1	4	5:3	97.65
	5	6:2	97.30
	6	7:1	95.83

By analyzing the data of experimental groups of 1-5, the following conclusions are obtained for the recognition rate of speaker-dependent.

1) The recognition rate will be smaller with the increase of the ratio of training set and test set, when the number of speech corpora is the same. The recognition rate is the ratio of the correct speech number and the total speech number. And the acoustic model can be optimized through increasing the number of speech sample. But the decreased recognition rate

is resulted in the reducing number of test set and the basically invariable number of correct recognition.

2) The recognition rate decreases with the increase of the speech number, when the ratio of training set and test set is the same. The average rate of recognition is above 95.83% whenever scheme of ratio is used. It can be drawn from the data that the average recognition ratio of 5:3 is the best (97.87%).

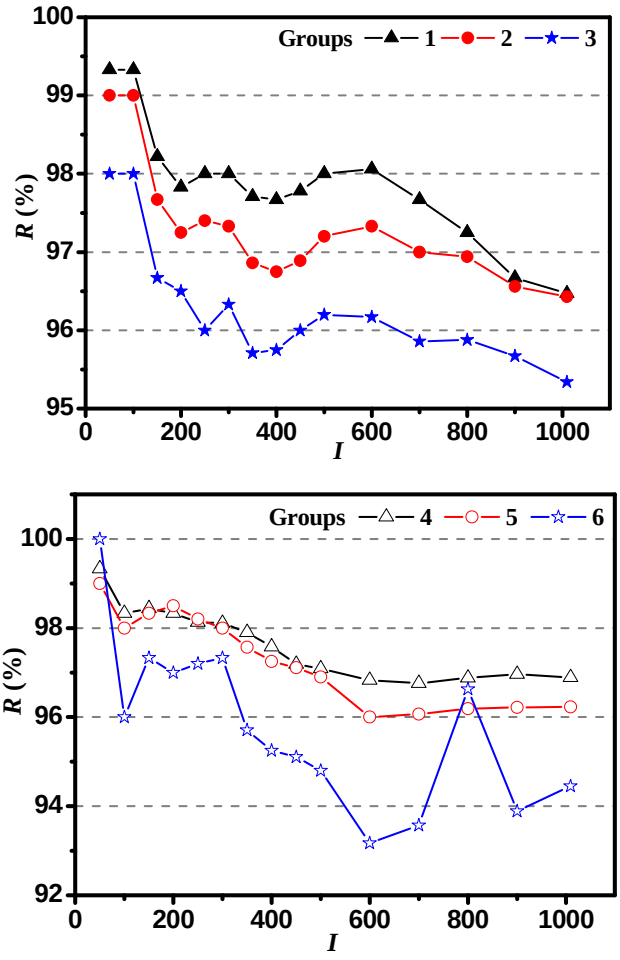

FIGURE III. THE RECOGNITION RATE OF THE SIX EXPERIMENTAL GROUPS WITH DIFFERENT VARIABLES

The following results are obtained by analyzing the sixth groups. Controllable factors affected recognition rate firstly is that two words with the same spelling and the same meaning should be labeled the same. Secondly, in the end of the original sound of the word, there is an abnormal noise such as the sound of recording pen. Uncontrollable factors include strong airflow of B_1's speeches, background noises, noise current, unvoiced consonant as the first syllable (such as t and k). Uncontrollable factors, like influencing the speech model, will affect the recognition rate. Nevertheless, the rate of recognition can be enhanced by means of wiping out controllable factors.

On the basis of the above conclusions, the speech data of the sixth experimental group with the lowest average recognition rate was chosen to improve the rate in the seventh experimental group. It was deleted the same words with a multi label and renewed cut the words with noise. Let 6:1 be the rate of training set and test set and total words is 1006. Figure IV shows comparison of the experimental group between sixth and seventh.

In the seventh experimental group, only five speeches can't be recognized in 1006 speeches which were supposed to be cut correctly. The overall average recognition rate was above 99.66%, while recognition rate of less 200 words was 100%.

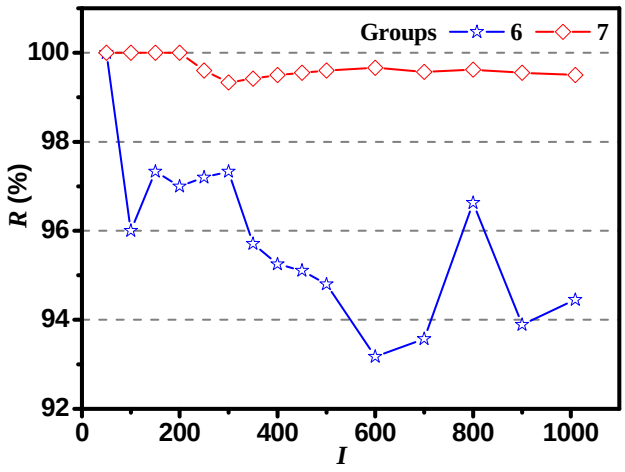

FIGURE IV. COMPARISON OF THE SPEECH QUALITY

IV. EXPERIMENTS OF IDENTIFICATION

In this paper, three groups of contrast experiment were proceeded in order to study whether the speech model of speaker-dependent has a good identification rate. In previous experiment, the rate of the first group is the highest. Hence, using the speech data of A_1 and the rate of 5:3, the acoustic model was established in this experimental group.

The tests set were selected three speakers (male A_2, female B_1 and B_2) as the research object. The speech quantity was 4×400. Let ten be the increment. Table II shows the results of average recognition rate, and Figure V is the line chart.

TABLE II. THE AVERAGE RECOGNITION RATE OF CONTRAST EXPERIMENTS

experimental groups	**i**	**ii**	**iii**
speakers	A_1-B_1	A_1-B_2	A_1-A_2
average recognition rate (%)	28.11	18.04	60.54

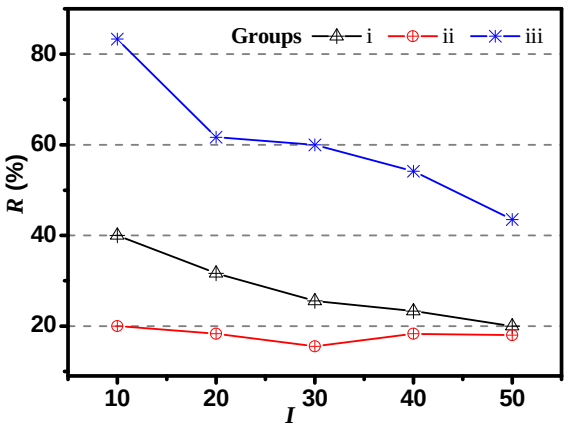

FIGURE V. THE RECOGNITION RATE OF THREE GROUPS OF CONTRAST EXPERIMENTS

The recognition rate of the group iii is overtopping than i and iii, because the parameters of MFCC are close to the vocal feature of human. While the overall rate of recognition is very low, it identifies that the acoustic model of speaker-dependent owns a well identification.

V. CONCLUSION

The research object was the southern dialect of Primi in this paper. The effects on speech recognition rate, include quantity of vocabularies, training and testing sample ratio, and data quality, are studied. It also investigated that whether speaker-dependent acoustic model is good to identify other speakers' speech. From experimental results, when the ratio of training set and test set is 5:3, the ratio of a large speech number can reach to 97.87% after rough segmentation. With the increase of cutting words, the recognition rate was basically invariable. The average recognition rate reached 99.66% and the recognition rate of small vocabulary can reach 100%.It shown that the established acoustic model of speaker-dependent in this paper has a good performance of identification.

ACKNOWLEDGMENT

This work was financially supported by Science Research Fund Project of Yunnan Provincial Education Department (2014Z091), Innovation Program of Yunnan Minzu University (2015YJCXY285) and Key Laboratory of IOT Application Technology of Universities in Yunnan Province.

REFERENCES

[1] Z. Q. Zhao, and J. D. Fang, "An improved speaker-dependent speech recognition systems and algorithms," J. Electronic Design Engineering, vol. 22, pp. 31-34, August 2014.

[2] L. X. Wei, M. Zhang, and Y. C. Zhong, "Noise speaker dependent recognition system based on PCNN," J. Computer Engineering and Applications, pp. 133-136, March 2012.

[3] L. X. Wei, "The Noise Speaker Dependent Speech Recognition System Based on PCNN," D. Guangdong University of Technology, 2011.

[4] H. Peng, W. Wei, and J. H. Lu, "Research on Speaker-Depended Isolated-Word Speech Recognition System," J. Control Engineering of China, pp. 397-400, March 2011.

[5] D. D. Sun, "Design of specific Person Speech Recognition Intelligence Heelchair System Based on Embedded," D. Changchun University of Science and Technology, 2011.

[6] G. Y. Li, and M. Meng, "Research on Acoustic Model of Large-vocabulary Continuous Speech Recognition for Lhasa Tibetan" J. Computer Engineering, pp. 189-191, May 2012.

[7] X. Yao, G. R. Shan, and Y. H. Li, "Research on Tibetan Isolated-word Speech Recognition System," J. Journal of Northwest University for Nationalities(Natural Science), pp. 29-36, January 2009.

[8] M. H. Nuo, "Dsign and Realization of Uyghur Isolated-word and Connected Digit Speech Recognition system," D. Xinjiang University, 2006.

[9] J. T. Yu, W. Y. Liu, and M. Z. Yi, "The Research Summary of Domestic Speech Recognition," J. Computer CD Software and Applications, pp. 76-78, October 2014.

[10] S. Z. Lu, "The Survey of Primi Language," J. pp. 58-73, April 1980.

[11] L. Y. Xie, "A Summary of Pumi Studies in China," J. Journal of Yunnan Minzu University of the Nationalities, pp. 75-78, Janury 2003.