Selection of acoustic modeling unit for Tibetan speech recognition based on deep learning

Baojia Gong [1,3,4*], *Rangzhuoma* Cai [1,2,3,4], *Zhijie* Cai [1,3,4],*Yuntao* Ding [1,3,4] , and *Maozhaxi* Peng [1,3,4]

[1]College of Computer Science and Technology, Qinghai Normal University, Qinghai Xining 810016, China
[2]School of Computer Science and Technology, Southwest Minzu University, Sichuan Chengdu 610041, China
[3]Tibetan Information Processing and Machine Translation Key Laboratory of Qinghai Province,Qinghai Xining 810008, China
[4]Key Laboratory of Tibetan Information Processing, Ministry of Education, Qinghai Xining 810008, China

Abstract. The selection of the speech recognition modeling unit is the primary problem of acoustic modeling in speech recognition, and different acoustic modeling units will directly affect the overall performance of speech recognition. This paper designs the Tibetan character segmentation and labeling model and algorithm flow for the purpose of solving the problem of selecting the acoustic modeling unit in Tibetan speech recognition by studying and analyzing the deficiencies of the existing acoustic modeling units in Tibetan speech recognition. After experimental verification, the Tibetan character segmentation and labeling model and algorithm achieved good performance of character segmentation and labeling, and the accuracy of Tibetan character segmentation and labeling reached 99.98%, respectively.

1 Introduction

Automatic speech recognition technology is a key technology for human-computer interaction. In recent years, deep learning-based speech recognition technology has achieved leaps and bounds [1-2] and is widely used in such fields as voice search, personal digital assistants, and in-vehicle entertainment systems [3].

The selection of modeling units for Tibetan speech recognition is the primary problem facing acoustic modeling in Tibetan speech recognition, which provides important safeguards for the whole Tibetan speech recognition process. In Tibetan speech recognition system, researchers have considered modeling units with different granularity, including words and syllables [4], vowels [5-8] and phonemes [9-11], respectively. Tibetan not only has a large vocabulary, but also various variants exist. If words or syllables are used as modeling units, the requirements of the corpus are too high and can lead to data sparsity

* Corresponding author: 2352334454@qq.com

problems. As a result, it is difficult to include all modeling units in the training data, making the modeling units lack sufficient training samples to ensure the reliability of the acoustic model. If vowels or phonemes are used as modeling units, it is not possible to distinguish the differences in pronunciation of the same phoneme at different locations. In order to solve the above problems, the paper proposes a method of using Tibetan character as the modeling unit, and presents the flow of its segmentation and labeling algorithm.

2 Tibetan character selected as the acoustic modeling unit for speech recognition

The Tibetan script is a logical grammatical system of pinyin, which is not only horizontally spelled but also vertically spelled. Usually, if one knows the pronunciation pattern of Tibetan letters, one can spell the corresponding pronunciation of Tibetan characters. Tibetan characters are written vertically to form a stack of characters, known as character [12] or precomposed character stack [13]. Specifically, Tibetan character is defined as all single characters and Tibetan stacked combination symbols including the base character, head letter, subjoined letter, and vowel. For example, the Tibetan syllable བསྒྲིགས in ད, སྒྲ, ི and ས are each one character, and the syllable is composed of four characters.

2.1 Tibetan character segmentation and labeling model

The Tibetan character segmentation and labeling model in this article is mainly composed of preprocessing, segmentation and labeling modules. The specific model is shown in Figure1.

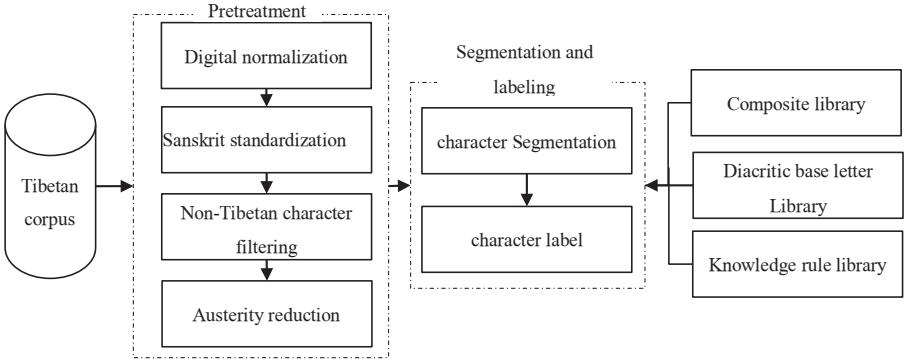

Fig. 1. Tibetan character segmentation and labeling model.

To use this model for Tibetan character segmentation and labeling, first, the system reads the text and preprocesses the read text. After that, the task of Tibetan character segmentation and labeling is carried out with the help of combinatorial library, Diacritic base letter library and knowledge rule library.

The pre-processing section consists of modules for digital normalization, Sanskrit normalization, non-Tibetan character filtering, and contraction reduction.

Segmentation and labeling are mainly composed of character segmentation and character labeling modules. Character segmentation refers to the longitudinal segmentation of Tibetan syllables, for example: ལེགས is divided into ལ, ེ and ས .Character labeling refers to labeling the segmented character according to the pronunciation of the same character in different positions. For example: ལེགས is marked as ལ, sེ and fས.

The composite component library is 476composite components counted in the literature [14] .The pronunciation of the same character in different positions in Tibetan language, for

example: single-component base letter and multi-component base letter, when is a single-component base letter is (ཨ་ག) pronounced 'ka' and appears in the multi-component base When the letter position (དགཨ) pronounced 'ga' sound, and so on, ང, ད, ཟ, ཨ and other 40 characters, their pronunciation will change. The modeling unit is the same , and there are obvious differences in voice features. Therefore, a base character library with diacritics was constructed. In addition to the difference between base letter and base letter, there are also pronunciation differences between base letter and post-added characters, base letter and post-added characters, and pre-added characters and base letter. In order to distinguish the same character in different positions, the Tibetan character is segmented, and then the prefixed letter, Diacritic base letter, suffixed letter and second suffixed letter are marked.

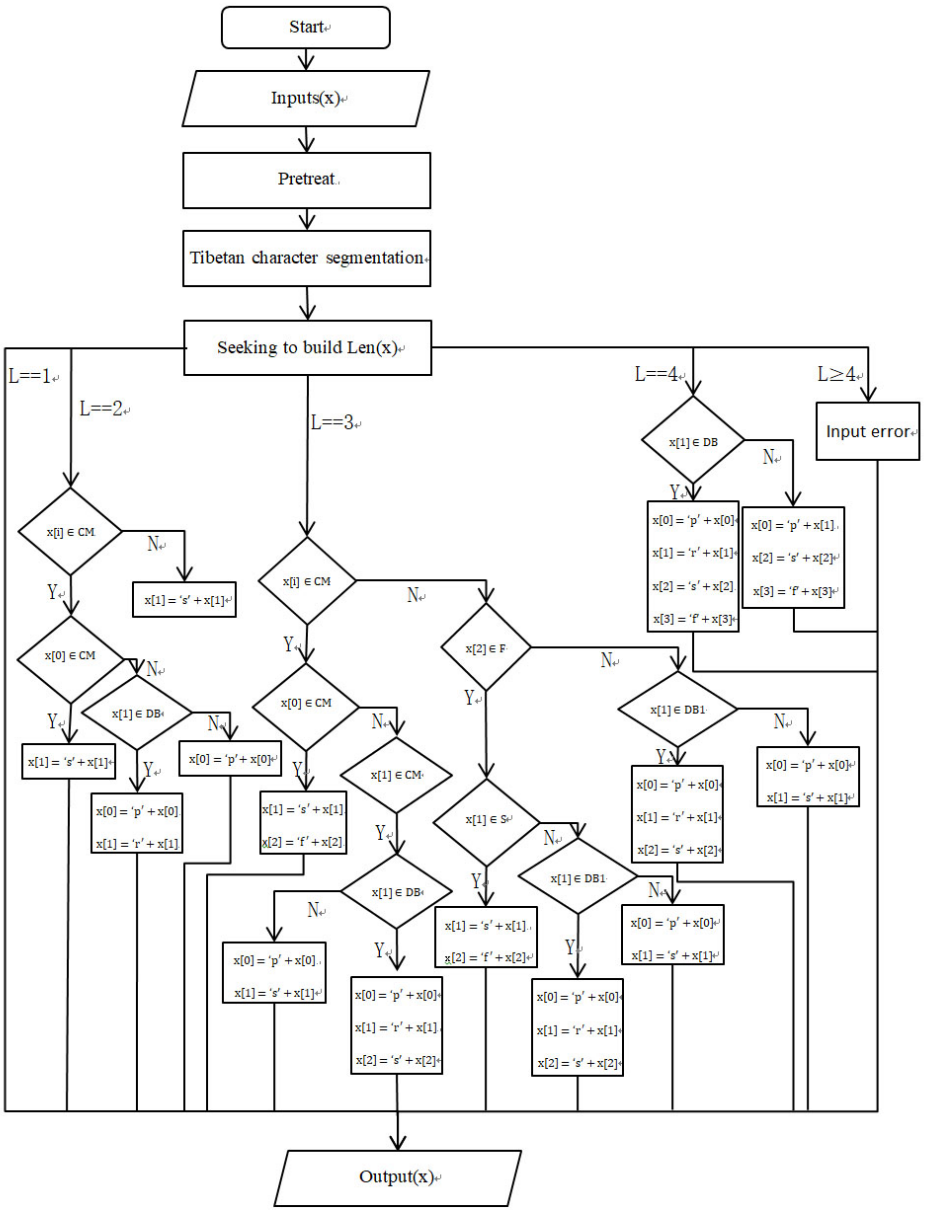

Fig. 2. Flow chart of Tibetan character segmentation and labeling algorithm.

2.2 Tibetan character segmentation and labeling algorithm flow

The text structure in Tibet is (prefixed letter+) base letter/ Combination construction (+suffixed letter)(+ second suffixed letter)[15]. Tibetan characters can be divided into four categories: single-component Tibetan characters, double-component Tibetan characters, three-component Tibetan characters and four-component Tibetan characters. Through the characteristics of Tibetan phonetics, Tibetan characters can be further divided into thirteen subcategories. For thirteen sub-categories segmentation and labeling, the algorithm flow chart of Tibetan character segmentation and labeling shown in Figure2 is designed.

As shown in Figure 2, in the algorithm flow, after inputting a Tibetan text, the output is the Tibetan text segmented and labeled by character. Among them, CM refers to the combined construction, DB refers to the diacritic base character library, S refers to ཀ, ང, ད and ས, and DB1 refers to the ང, ཞ and ཟ in the diacritic base font. F refers to ས.

3 Experiments

3.1Experimental data description

Since there is currently no shared text corpus in Tibetan, this article character the existing Tibetan corpus in the research group to establish three types of sample corpora that can be used for character segmentation and labeling of Tibetan, in order to facilitate the description below , These three types of sample corpus are represented by C1, C2 and C3 respectively. The C1 corpus contains 220,000 sentences, 4.35 million words and 553 character; C2 corpus contains 240,000 sentences, 5.34 million words and 561character; C3 corpus contains 260,000 sentences, 4.1 million words and 563 character. The sources of corpus are Tibetan textbooks for elementary and middle schools, Weirenwang.com, China Tibetans.com and Qiong mai Tibetan Literature.

3.2 Experiment and result analysis

Experiment 1 In order to verify the segmentation and labeling effects of character, we designed the first set of experiments to investigate the segmentation effects on different data sets and count the size of the characters.

Table 1. Splitting and labeling experiment results of Tibetan character.

Corpus	Word count (million)	Number of characters	Accuracy/%	Instance	
C1	4.35	562	100	original	སངས་རྒྱས་ཐམས་ཅད་ལ་ནི་སྐུ་ལས་ཆིང་གསོལ་བ་འདེབས་སོ
				result	ས sང fས རྒྱ sས ཐ sམ fས ཅ sད ལ ནི སྐུ sལ ཆི sང pག སོ sལ བ pའ དེ sབ fས སོ
C2	5.34	583	99.98	original	བང་ཀ་རམ་བརྟེད་ལྡན་པའི་རི་བོ་ཐའེ་ཏྲན་གྱི་ཡང་རྩེར་འགོས་རྒྱོང
				result	pབ rད sག རམ pབ རྟེ sད ལྡ sན་ བ sའི རི་ བོ ཐ sའེ ཏྲ sན་ གྱི ཡ sང རྩེ sར pའ གོ sས རྒྱ sང *fའ*
C3	4.10	567	100	original	ཀྱང་བབྱང་ངམག་ཚུད་ཀྱི་ཀུའུ་ཞེ་ཞེ་ཅེན་ཐོང་གེས་རྒྱལ་ཁས་ཀྱི་བྱས་རྟེའ་ཀྱགས་གཅད་བ་རེད
				result	ཀྱ sང pད ཐ sང pད ཨ sག ཚུ sད ཀྱི ཀྱ sའ ཞེ ཞེ ཅེ sན་ ཐེ sང ཐི sས རྒྱ sལ ཁ sའ ཀྱི ཐ sས རྟེ sས pག rཟེ sང fས རྐ sག fས pག ན sང བ རེ sད

From the experimental results in Table 1, it can be seen that the accuracy of the Tibetan character segmentation and labeling model reached 99.98%, and the error rate of 0.02% is that there are two syllables in the text that lack syllables and label errors. It shows that the Tibetan character segmentation and annotation model proposed in this paper has achieved good segmentation and annotation effects.

Experiment 2 According to the frequency of use of character in all modern Tibetans, it is divided into four levels: prefixed character, base character, Diacritic base character, suffixed character, and second suffixed character. Figure 3 lists the relationship between the number of characters and frequency of use.

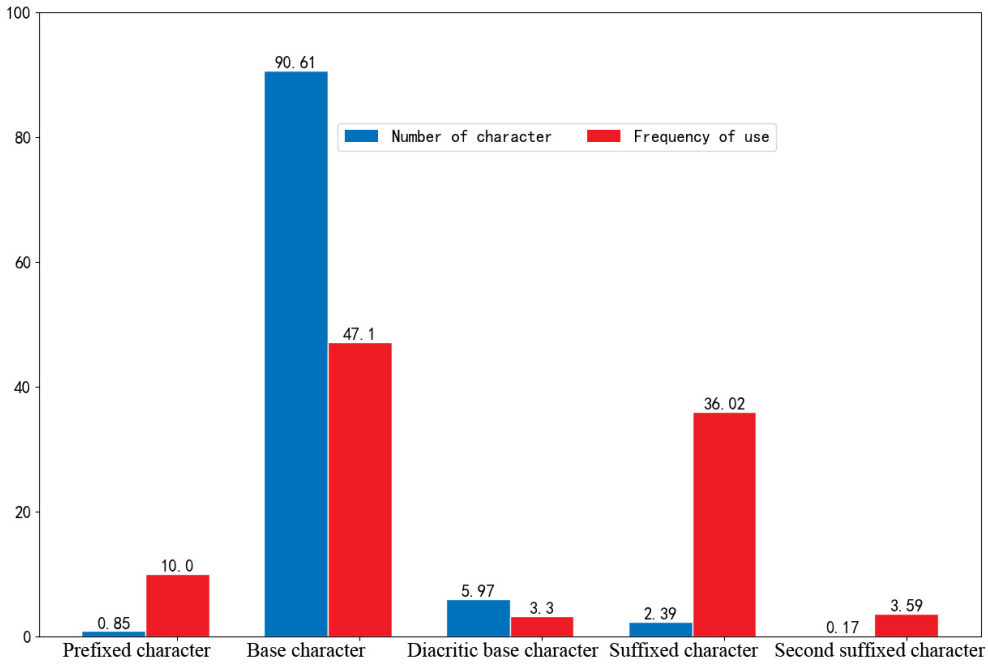

Fig. 3. Frequency distribution table of Tibetan character

It can be seen intuitively from Figure 2 that only less than 4% of characters are used when composing Tibetan scripts, but their use frequency is as high as 49.61%, and 4% of characters are all diacritical characters. In the latter 90.61% of characters, only 47.10% appeared when composing Tibetan, and 90.61% belonged to invariant characters. This shows that in Tibetan, it is impossible to distinguish the difference in pronunciation of the same phoneme in different positions. Therefore, the segmentation and labeling of Tibetan character has resolved the differences in phoneme pronunciation.

4 Summary

Tibetan character segmentation and labeling are the basic work of selecting acoustic modeling units for Tibetan speech recognition. This paper proposes the algorithm flow of Tibetan character segmentation and labeling by designing the Tibetan character segmentation and labeling model. Experiments show that the Tibetan character segmentation and labeling model and algorithm flow have achieved good character segmentation and labeling performance. The accuracy of Tibetan character segmentation and labeling has reached 99.98%, and the segmentation and labeling effects can be achieved. Satisfying

practical needs has laid the foundation for the subsequent establishment of the acoustic model of Tibetan speech recognition based on the word D and speech recognition. We plan to study the acoustic model of Tibetan speech recognition based on the neural network of Tibetan characters on the basis of the work of this paper in the future to improve the performance of Tibetan speech recognition.

This research was financially supported by the National Natural Science Foundation of China (61966031,61866032); Ministry of Education "Chunhui Program" (Z2016077, Z2012093); Qinghai Province Science and Technology Project (2019-SF-129); Qinghai Province Key Laboratory Project (2014-Z-Y32, 2015-Z-Y03); Key Laboratory of Tibetan Information Processing and Machine Translation (2013-Y-17) ; Qinghai Normal University 2020-2021 Innovative Training Project.

References

1. Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, **29(6)**: 82–97.
2. Graves A, Mohamed A R, Hinton G. Speech recognition with deep recurrent neural networks[C]// InterSpeech.Canada, 2013: 6645–6649.
3. Seltzer M L, Ju Y C, Tashev I, et al. In-car media search[J]. IEEE Signal Processing Magazine, 2011, **28(4)**:50–60.
4. Zhao Erping, Wang Conghua, Dang Hongen, Luo Weiqun. Research on the Speech Recognition Technology of Tibetan Isolated Words[J]. Journal of Northwest Normal University (Natural Science Edition), 2015**(51)**: 54.
5. Li Guanyu, Meng Meng. Research on Acoustic Model of Continuous Speech Recognition in Tibetan Lhasa Vocabulary[J]. Computer Engineering, 2012, **38(5)**: 189-191.
6. Deqing Zhuoma. Research on the extraction of Tibetan phonetic features based on a small vocabulary of a specific person[D]. Tibet University, 2010.
7. La Long Dongzhi. Research on Tibetan Speech Recognition Technology [D]. Tibet University, 2015
8. Deji. Tibetan phonetic labeling and recognition research [D]. Qinghai University for Nationalities, 2014
9. Huang Xiaohui, Li Jing. Tibetan speech recognition acoustic model based on recurrent neural network [J]. Journal of Chinese Information Processing, 2018, **32(05)**: 49-55.
10. Nan Cuoji, Cairang Zhuoma, Du Gecao.Tibetan speech recognition based on BLSTM and CTC[J].Journal of Qinghai Normal University (Natural Science Edition),2019(4).
11. Wang Song. Tibetan Lhasa speech recognition system based on LSTM-CTC[D]. Lanzhou. Northwest University for Nationalities. 2019
12. WANG weilan, CHEN Wanjun Tibetan script, syllable frequency and information entropy [J] product safety and recalls, 2004,**000 (002)**: 27-31.
13. Jiang Di. *Research on Tibetan Characters*[M]. Social Sciences Archive Press, 2010.
14. Cai Rang Zhuoma, Cai Zhijie.An algorithm for word component decomposition in Tibetan character frequency statistics system[J].Computer Engineering and Science,2011,**33(3)**:159-162.
15. Cai Zhijie. Research on Key Techniques of Tibetan Word Vector Representation[D]. Qinghai Normal University, 2018