Research on Tibetan Speech Recognition Speech Dictionary and Acoustic Model Algorithm

Chunwei Gong*1, Guanyu Li1 and Lisai Luo1

1 National Languages Information Technolog Department, Northwest Minzu University, Lanzhou, Gansu Province, 730000, China

*Corresponding author’s e-mail: hdd_gcw@163.com

Abstract. This paper mainly conducts related research on the algorithm of the Tibetan acoustic model. Firstly, a Tibetan Lhasa pronunciation dictionary is constructed based on words and monosyllables, and then compared. The experimental results show that the phonetic dictionary constructed by words is superior to the phonetic dictionary constructed by monosyllables. At the same time, taking phonemes and single syllables as the modeling unit of the acoustic model, the modeling unit of the acoustic model was studied and a comparative experiment was carried out. Experimental results show that phonemes are better than single syllables as acoustic model modeling units.

1. Introduction

Tibetan belongs to the Tibetan branch of the Tibetan-Burmese group of the Sino-Tibetan language family, with more than 6 million speakers. It is mainly distributed in China's Tibet Autonomous Region and 5 provinces including Sichuan, Yunnan, Qinghai, and Gansu. Tibetan is mainly divided into three major dialects: Wei Tibetan, Amdo and Kham. The three dialects have the same text, but there are large differences in pronunciation. This article selects the most widely used and most representative Lhasa dialect (which belongs to Wei Tibetan dialect) as the research object.

The challenges of Tibetan speech recognition include three aspects: (1) lack of resources; (2) the Tibetan word segmentation system has a great impact on Tibetan speech recognition; (3) the acoustic characteristics of Tibetan have not been well studied. The statistical characteristics of the automatic speech recognition method require a lot of language resources to ensure its good performance. For most minority languages, there is a problem of insufficient speech corpus, and there is also a lack of standardized automatic processing tools (standard character encoding, word processing software). At present, in Tibetan speech recognition research, dictionaries are built in units of single syllables [1]. When a dictionary is built with syllables as its unit, the language model can only learn the contextual relationship of the syllables, and cannot fully play the role of the language model. This article builds a dictionary based on words and trains to get a language model. In speech recognition, the acoustic model can be understood as the modeling of the acoustic characteristics of speech. The acoustic model can convert the input of speech into a probability representation that belongs to a certain acoustic symbol, which is the modeling unit of the acoustic model and the modeling of the acoustic model. The unit includes words, syllables, vowels, and phonemes, etc. This paper chooses to use single syllables and phonemes as recognition units for comparison experiments.

The rest of the paper is organized as follows: Section 2 mainly outlines Tibetan syllables, word concepts, the construction of a pronunciation dictionary, and the training of language models; Section
3 details the framework of the acoustic model and the choice of modeling units. Then analyze the experimental results in Section 4. Section 5 will summarize the experimental work.

2. The concept of Tibetan syllables and words
Tibetan is a type of pinyin, with 30 consonants and 4 vowels. These letters form syllables and then syllables form words. In Tibetan, syllables are separated by the syllable separator " - " (tsheg) [2]. However, there is no delimiter to distinguish between Tibetan words, which has also attracted much attention for Tibetan word segmentation. Tibetan syllables can be understood as Tibetan words. Computers can clearly understand the meaning of a word, but cannot easily understand the meaning of the word. However, for example, "I am a student", The computer can clearly understand the word student, but cannot clearly understand the two words of learning and students. At the same time, there is a strong correlation between words, but the correlation between words is weak. Choose to build a pronunciation dictionary in words.

Speech recognition is mainly divided into three parts: acoustic model, language model and pronunciation dictionary. The following figure can clearly show the relationship between the acoustic model, language model and pronunciation dictionary in speech recognition.

![Figure 1. Structure of speech recognition](image)

The pronunciation dictionary contains a set of words that can be processed by the speech recognition system, and the pronunciation of each word is marked. Through the pronunciation dictionary, the mapping relationship between the modeling unit of the acoustic model and the modeling unit of the language model can be obtained, so that the acoustic model It is connected with the language model to form a searched state space for the decoder to perform decoding [3]. The construction process of the pronunciation dictionary and the training of the language model will be described in detail below.

2.1. Construction of pronunciation dictionary
Due to the lack of syllable separators in Tibetan, acronyms are formed. Acronyms appear frequently in Tibetan, and acronyms appear in about 30% of vocabulary. This is the focus and difficulty of Tibetan word segmentation. Therefore, when designing a pronunciation dictionary, considering the influence of acronyms, the pronunciation dictionary in this article does not split acronyms. The words in the pronunciation dictionary of this article are mainly compiled from the electronic version of the Tibetan Dictionary and a large number of Tibetan textbooks. A total of about 25,029 Tibetan words have been
arranged. In addition to the basic Tibetan words, place names, person names, Institution names, common spoken words, verbs and their variants. These 25,029 Tibetan vocabularies were compiled according to the Tibetan phonetic conversion rules and manually collated into a Lhasa pronunciation dictionary with words as the unit. Because Tibetan words are composed of one or more syllables, the syllables are separated by separators. According to different modeling units, a phoneme pronunciation dictionary and a syllable pronunciation dictionary are established. According to the syllable pronunciation rules, the 25029 words are syllable-matched and marked with pronunciation, and a phoneme pronunciation dictionary is established. The Tibetan words are divided according to syllable separators, and a pronunciation dictionary with syllables as a unit is established for experimental research.

Table 1 Syllable pronunciation rules

| k a | k a k | k a ng | k ecb | k ec | k a p |
|-----|------|-------|-------|------|-------|
| c a m | c a | c ec | c ec | c i k |
| ds a | ds a k | ds a ng | ds ec | ds ec | ds a p |

There are 6013 Tibetan syllables and corresponding pronunciations in the Tibetan syllable pronunciation rule table, as shown in the figure above, which is a part of the syllable pronunciation rule table.

2.2. Language model training

The purpose of the language model is to give a maximum probability text sequence based on the output of the acoustic model. Suppose there is a word sequence \( S = W_1W_2...W_K \) of \( K \) words, the probability of this word sequence appearing is:

\[
P(S) = P(W_1, W_2,..., W_K) = P(W_1)P(W_2 | W_1)...P(W_K | W_1, W_2,..., W_{K-1})
\]

This article uses SRILM for language model training. Due to the small size of the Tibetan text corpus, most words or collocations appear less or not at all in the corpus, which results in a sparse data situation. In order to reduce the performance of post-processing, this experiment tested a variety of smoothing algorithms during the training of the language model, thereby maximizing the performance of the language model.

Language model training requires a large amount of text corpus. The text corpus used for language model training in this article is a total of 300 trillion, a total of 1,957,039 sentences, of which 50 megabytes come from Tibetan-related news corpora on WeChat public account, 250 megabytes from Tibetan news Corpus. Use all the text data to train the language model. In addition, the size of the test text is 12 megabytes, for a total of 77120 sentences. When training a language model, the training corpus needs to be segmented. The syllables in Tibetan are separated by the syllable separator "་" (tsheg), but there is no obvious division between words. Due to the omission of the delimiters between syllables in Tibetan, acronyms are formed. Acronyms appear very frequently in Tibetan texts, and acronyms are present in about 30% of vocabulary, which makes Tibetan word segmentation encountered great difficulties. In this paper, the best-performing Tibetan word segmentation system TIP-LAS is selected to segment the text corpus. Based on this, the text after the segmentation is checked, the sentence of the segmentation error is corrected, and finally the text used for language model training is obtained.
When using the TIP-LAS system to segment the text corpus, the following three rules are summarized for correcting errors in the segmentation:

1: རྐྱོང་བ་ replaced with རྐྱོང་བས
2: རུབ་ེས་ི་ replaced with རུབ་ེས་ིེ
3: ད་ལྐྱོེ་ི་ replaced with ད་ལྐྱོེ་ིེ

Contrast experiments on different smoothing algorithms of language models when training language models. The language model smoothing algorithms used in this paper include: Good-Turing estimation method, Katz smoothing method, Witten-Bell smoothing method, The Kneser-Ney smoothing method compares the superiority of the perplexity of the language models trained by each method, and finally selects the language model trained by the Good-Turing estimation method. Will be explained in the experiment.

3. Research on acoustic model algorithm

With the continuous development of neural networks in speech recognition, compared with the traditional gmm-hmm built acoustic model has been greatly improved. The experiments in this paper are completed under kaldi. The neural network used in this paper is Time-Delay Neural Network.

3.1. Construction of acoustic model based on TDNN

TDNN is a multi-layer feedforward neural network [6]. Unlike the traditional feedforward neural network, the input of each hidden layer is the output of the previous layer. TDNN expands the output of each hidden layer and hides it. The current output of the layer is spliced with the output of several moments before and after it as the input of the next hidden layer, so the TDNN can better learn the context information of the speech. The following figure is the standard TDNN structure:

![Time-delay neural networks structure diagram](image)

TDNN is used for phoneme recognition. The leftmost 5 matrices of the first layer are shared in the time dimension. The delay of the delayed neural network is 4, and the next 4 frames of the current frame will be connected to the hidden layer of the second layer. These five weight matrices will be shared in subsequent time steps, so as to ensure that the calculation amount is within a reasonable range, and the second layer hidden layer output at different times can be obtained with the first layer calculation to the end. The delay of the second layer is set to 3, that is, the current hidden layer output of the second layer and the hidden layer output at the next 3 times are connected to the third layer hidden layer. The last layer is set to a delay of 2, so that the output of the third hidden layer and the output of the hidden layer at the next two moments are both connected to the output layer.
Due to the traditional TDNN, the activation function of the hidden layer is calculated once for each frame. As the number of network layers increases, the input of the hidden layer will also have a lot of repeated context information, and the amount of calculation will be large during training. To solve these problems, a sub-sampling method is usually used. In this way, TDNN can use non-adjacent frames, which reduces the overlap of context information and reduces the amount of calculation during training. The acoustic model framework in this paper uses the TDNN of the subsampling method is shown in the figure below.

![Figure 3. Time-delay neural networks using subsampling method](image)

The above figure is the TDNN structure using the subsampling method. The input of the first layer is two frames before and after the current frame, and the input of the second layer is only using the frame from the previous time and the second time from the current frame. Frame, the input of the third layer is to use only the frame from the current frame to the third moment before and the frame from the third to the next moment, and the input of the fourth layer is to use the frame from the current frame to the seventh moment and for the frames at the second moment in the future, it can be seen from the TDNN structure of the subsampling method that compared to the traditional TDNN structure, the TDNN of the subsampling method is more streamlined and the redundant information of the context can also be eliminated.

4. Experimental results and analysis

4.1. Tibetan Phonetic Corpus

The corpus text used in this article comes from the Tibetan news text corpus, and the speech corpus comes from the self-built Tibetan Lhasa reading corpus. It has a total of 160 Tibetan speakers and a total of 85 hours of corpus, including 50 hours of mobile phone recording corpus and computer recording of corpus. 35 hours. The specific situation is shown in the table:

| Number of Utterances | Speakers |
|----------------------|----------|
| Dev-set              | 1640     | 4       |
| Test-set             | 1631     | 4       |
| Train-set            | 73845    | 152     |

4.2. Analysis of experimental results of language models

The number of syllables in the syllable dictionary is 6013. There are 25029 words in the pronunciation dictionary in terms of words. The text size of the trained language model is 300 trillion. There are
5935 types, 2-element statistical combinations are 413602, and 3-element statistical combinations are 1,948,889. One-element statistical combinations of the language model trained by the pronunciation dictionary constructed by word units are 25,913, and 2-element statistical combinations are 986,994. The number of ternary statistical combinations is 2908127, and the following are the experimental records when using different methods to train language models.

The vocabulary-based pronunciation dictionary uses various smooth-trained language models as shown below:

Table 3: Smoothing algorithms for syllable pronunciation dictionary training language models

| Smoothing algorithm       | PPL  |
|--------------------------|------|
| goodTuring algorithm     | 14.95244 |
| Katz algorithm           | 15.54064 |
| Witten-Bell algorithm    | 15.89334 |
| Kneser-Ney algorithm     | 17.48115 |

The pronunciation model in terms of words uses various smooth trained language models as shown below:

Table 4: Word pronunciation dictionary training language model smoothing algorithms

| Smoothing algorithm       | PPL  |
|--------------------------|------|
| goodTuring algorithm     | 14.88157 |
| Katz algorithm           | 15.56837 |
| Witten-Bell algorithm    | 16.01909 |
| Kneser-Ney algorithm     | 19.37291 |

By comparing ppl, this paper chooses a language model trained using Goode-Turing smoothing algorithm.

4.3. Experimental results and analysis of acoustic models

This article uses the TDNN structure of the sub-sampling method. The input features use 40-dimensional MFCC features to stitch 3-dimensional Pitch features, using 6 hidden layers, 900 hidden layer nodes, 2944 output layer nodes, and hidden layers according to \{0\} \{-1,0,2\} \{-3,0,3\} \{-7,0,2\} \{-3,0,3\} \{0\} for framing, and the input layer is stitched in addition to the current frame A total of 5 frames before and after. A total of 4 Epochs were trained, a total of 224 iterations were performed, the Minibatch size was 512, and the initial learning rate was 0.0015 and the termination learning rate was 0.00015.

The error rate calculation method used in this article is the syllable error rate, which is the Syllable Error Rate. The following will use the SER. The specific calculation formula is as follows:

$$\text{Syllable Error Rate} = 100 \times \frac{I+S+D}{T}$$

Among them, I represents an insertion error, S represents a replacement error, D represents a deletion error, and T represents a total number of words.

The pronunciation dictionary constructed by word consists of 20921 Tibetan words in total, and the pronunciation dictionary consists of 5,933 syllables in total. To compare the effects of the construction unit of the pronunciation dictionary, the acoustic model modeling unit (here The selected acoustic model modeling unit is phone.) The framework of the acoustic model is the same as the acoustic model.

Table 5: Comparison of experimental results of pronunciation dictionary building units

|               | Dev  | Test |
|---------------|------|------|
| Pronunciation dictionary by word | 11.93 | 16.48 |
| Pronunciation dictionary in syllables | 16.27 | 20.32 |
The error rate in the figure above is the syllable error rate. It can be seen from the experimental results that the pronunciation dictionary constructed in terms of words has a higher recognition accuracy than the pronunciation dictionary in terms of units.

In the course of the experiment, it was found that the modeling unit of the acoustic model may affect the experimental results. Therefore, this article uses phone and char as the modeling unit of the acoustic model to conduct experiments. The selected acoustic model framework and language model are the same. Constructed in terms of words, the experimental results are shown below:

|         | Dev  | Test |
|---------|------|------|
| Phone   | 11.93| 16.48|
| Char    | 13.27| 19.67|

When comparing the modeling granularity of the acoustic model, it can be found from the experimental results that the fine-grained modeling unit will perform better on a small data set than the coarse-grained modeling unit. The purpose of doing this set of comparative experiments is to In the future, when the amount of data increases, make preparations, and also make some relevant preparations for the multi-task learning of the acoustic model.

5. Summary
Speech recognition plays an important role in the field of artificial intelligence. This paper mainly studies the algorithm of acoustic model in Tibetan speech recognition, and has achieved good results on the experimental results. In the future we will also focus on research in the field of artificial intelligence.

Acknowledgments
The author thanks Li Guanyu and Luo Lisai for their valuable suggestions. We thank the Northwest University for Nationalities for its research and innovation project "Research and Implementation of Tibetan Speech Keyword Detection Technology Based on LSTM" (Yxm2018122) for its support.

References
[1] Li Yachao, Jia Yangji, Yu Hongzhi. TIP-LAS: An open source Tibetan part-of-speech tagging system [2] [C]. The 14th National Conference on Computational Linguistics (CCL 2015).2015.11. 13-14, Guangzhou.
[2] Li Yachao, Jia Yangji, Zong Chengqing, Yu Hongzhi. Research and Implementation of Tibetan Automatic Segmentation Method Based on Conditional Random Field [J].Journal of Chinese Information Processing. No.4, Volume 27, No 52-58, ISSN: 1003-0077.2013,7. (CSCD core)
[3] Deqing Zhuoma. A Review of Tibetan Speech Recognition [J]. Journal of Tibet University, 2010 (s1): 192-195. [4] Li Guanyu, Meng Meng. Acoustics of Tibetan Lhasa Large Vocabulary Continuous Speech Recognition Model Research [J]. Computer Engineering, 2012, 38 (5): 189-191.
[4] Li Guanyu. Research and Implementation of Continuous Speech Recognition in Tibetan Lhasa Large Vocabulary Based on HTK [J]. Journal of Northwest University for Nationalities (Natural Science Edition), 2011, 32 (3): 19-23.
[5] Wang Hui, Zhao Yue, Liu Xiaofeng, et al. Tibetan Speech Recognition Based on Deep Feature Learning [J]. Journal of Northeast Normal University (Natural Science), 2015, 47 (4): 69-73.