Automatic Segmentation of Syllables in the Flow of Tibetan Lhasa Dialect

Haibo Shi[1, a], Yonghong Li[2, b, *]

[1]Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lan Zhou, China

[2]Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lan Zhou, China

[a]512772113@qq.com, [b]lyhweiwei@126.com

*Corresponding author

Keywords: Tibetanlhasadialect, HMM, HTK, Forcedalignment, Automatic segmentation

Abstract: This paper studies and realizes the problem of phoneme automatic segmentation of Tibetan Lhasa dialect. In this paper, the automatic segmentation of the related methods and details about based on hidden markov model (HMM) of the Tibetan capital Lhasa dialect phoneme principle and method for automatic segmentation of HTK as a set of specially set up and handle the HMM tools is the best tool to implement this method, the use of HTK toolkit experimental phonetics, combining with the characteristics of itself in the Tibetan language and language flow has been basically achieved the phoneme segmentation automatically, [1]based on the HMM model training, can undertake effective segmentation batch of corpora, greatly saves manpower and time cost.

1. Introduction

Corpus for human-computer interaction is very important in various fields, the establishment of the corpus need to deal with the corpus of the original data, one of the most important work is the phoneme segmentation, before this job need manual annotation segmentation, need to have the background knowledge of linguistics to repeatedly listen to the sound, tags, segmentation, when large corpus, spend a lot of manpower, still need to spend a lot of time cost, artificial segmentation error uniformity. Now, with the development of deep learning and the improvement of computer hardware, there is an urgent need for corpora with large amount of data, including tens of gigabytes or even hundreds of gigabytes. Such a large amount of data manual segmentation efficiency is too low, the workload is too large. To solve this problem, people began to think about how to use machines instead of humans to achieve more efficient high-speed segmentation of large scale speech signals, with the development of computational linguistics and relevant technical algorithms, many methods of automatic segmentation of phonemes have been developed recently, which greatly improves the efficiency. At present, there are many methods of phoneme segmentation, one of which is mainly based on hidden markov model, Mainly based on hidden markov model group, its advantage lies in the HMM containing statistical model framework, can in the limited training of voice and data collection training algorithm estimate the model parameters, the mass of the segmentation of speech signal has the possibility of implementation, the most important thing is experiment based on the HMM model kit has a lot of, can change the type, size, and even structure model, the flexibility of the system to distinguish all kinds of sounds, human language. while the other is based on bayesian network model or template. Now the main method is HMM model. The construction of Tibetan language corpus laid the foundation for the study of the ethnic minority languages, is necessary to work, to protect minority language culture for speech recognition, speech synthesis and the study of dialect culture provides the platform, the construction of the corpus pushed forward the development of speech research, provides information service for the Tibetan

people, greatly facilitate the Tibetan people's life, also let other people know more about Tibetan culture

2. Research status at home and abroad

Most of the manual segmentation methods adopt voice personnel with professional background to repeatedly listen to phonetic symbols with analysis tools. The advantages are high accuracy, while the disadvantages are poor uniformity. As people work longer hours, their attention decreases, and their accuracy decreases. To realize the automatic segmentation of phonemes can be divided into two categories, one is doing segmentation using a priori knowledge of linguistics, the second can use speech signal in the frequency domain or time domain feature segmentation, such as tsinghua university researchers syllable segmentation automata algorithm based on merging, use the include energy, the zero rate, pitch and cepstrum parameters, and other characteristics, to distinguish between the frame and the frame, and the Chinese academy of sciences, researchers at the hidden markov model (HMM) was proposed for the method of segmentation phonemes, this method is the most stable, and widely used method of tectonic units, and the corresponding unit voice library. In addition, northwest university for nationalities puts forward a method to explore commonness and transfer training by utilizing the relationship between languages in the state of language data shortage. There is also an automatic DTW-based segmentation algorithm that twisting the time axis of the speech signal, the acoustic feature sequence of the signal is forced to align with the acoustic feature of the reference template, so as to determine the boundary points of syllables., so as to determine the boundary points of syllables. At present, this algorithm works well in the segmentation of isolated words in corpus, but its disadvantages are also obvious. The figure below shows the manual segmentation method. *Fig.1* shows Tibetan students manually marking voice signals with praat software.

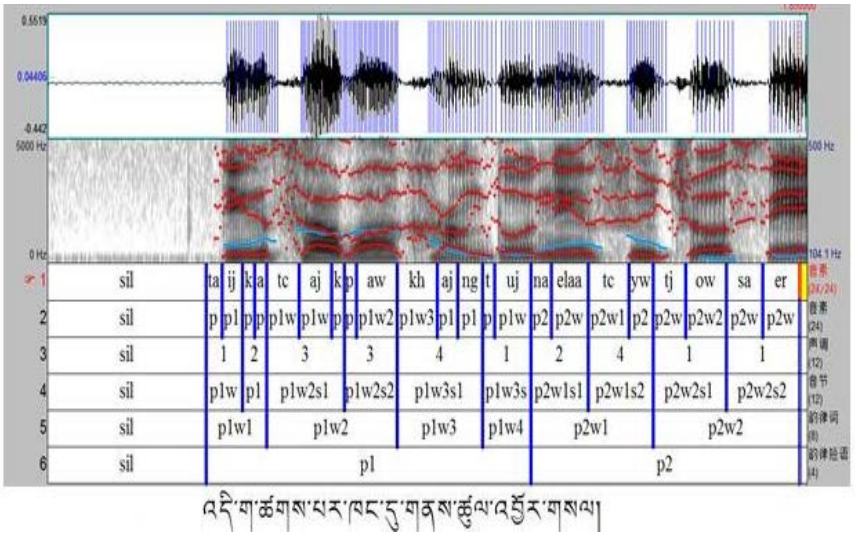

Figure 1. Manually annotated phonemes

Endpoint detection: endpoint detection refers to the determination of the starting and ending points of a speech signal.

2.1 Short-term energy algorithm

When the speech signal is at very low amplitude, it is obviously mute. We set a threshold, under which it can be determined as mute. However, the mute is not dependent on the friction of vocal cords, and the short-term energy is even smaller than the mute.

2.2 Short time zero crossing rate

By observing the waveform change of speech signal, it is found that the waveform change of silent segment is relatively slow, the waveform change of silent segment is relatively drastic in

amplitude, and the frequency of crossing the zero point is more than that of silent segment.Zero crossing is how many times the sample changes sign.

The short time energy is more suitable for detecting dullness and the short time zero crossing rate is more suitable for detecting voiceless.

3. Automatic sharing based on hmm

After the text HMM is actually divided into two parts. The first is markov chain, which describes how the frame state of a speech signal is transferred to another frame state, and the output state sequence of this process by using a set of statistical correspondence of state transition associated with probability distribution. The second process is a random process, which describes the relationship between state and observed value. The observed sequence is used to describe the implicit state, which is generated as a sequence of observed value. According to the different probability of observation values, HMM can be divided into discrete, semi-continuous and continuous HMM model structures. According to the different transition probability matrix, HMM can be divided into left-to-right model with parallel path from left-to-right model and left-to-right model without jump. In this paper, left-to-right model is adopted. Hidden markov model is widely used in various fields. It is a probabilistic model that describes the statistical characteristics of random process. Firstly, the continuous speech waveform is converted into a discrete vegetable vector sequence of equal length. The speech vector sequence and the hidden symbol sequence implement a mapping, while the viterbi algorithm can be regarded as a matrix to find the optimal path. It is a special dynamic programming algorithm, which uses the algorithm in dynamic programming to find the hidden information sequence that may produce the observation sequence. Automatic segmentation of phonemes is realized by using the automatic alignment principle in HTK.

4. The experiment tools

There are many tools to choose for the implementation of hidden markov model, including HTK, GHMM Library, UMDHMM, Jahmm Java Library, etc. This paper adopts the HTK Toolkit, Hidden Markov Model Toolkit: developed by the machine intelligence laboratory of Cambridge university, and the underlying language is C language. HTK provides a series of tools such as data recording, model training and speech recognition. *Fig.2* shows the software structure diagram is as follows.

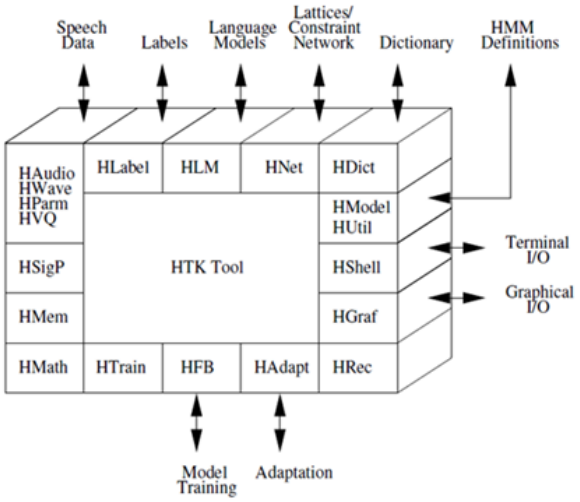

Figure 2. HTK tool structure diagram [2]

5. Experimental data preparation

The first step in any experimental project is data preparation. Training data and test data need to be prepared. For the automatic segmentation of phonemes, it is necessary to prepare the original

recording data. In the preparation of test data, the initial phoneme layer should be marked, which is helpful for data model training. Considering the syllabic characteristics of Tibetan Lhasa dialect, test training should be conducted according to the manually marked boundary. The experimental speech data are in WAV format, the speaker is a professional young announcer in Lhasa dialect, the sampling rate is 16k, and the precision is 16 bits, the single channel, and the normal speech speed.

According to the characteristics of Tibetan language, the phoneme list adopted in this paper is sampa-st designed by China institute of ethnic information technology, northwest university for nationalities. [3] *Fig 3* is the vowel list in Tibetan language.

Pronunciation characteristics			basic	long-tones	phary-ngeali-zation	Nasal-sounding	Next phoneme				
tongue position		lip shaped					k	m	p	u	ŋ
front	low		a	a:			ak	am	ap	au	aŋ
front	lower middle		ɛ	ɛ:	ɛʔ						
front	medium to high		e	e:	eʔ	ẽ		em	ep		eŋ
front	medium to high	lip rounding	ø	ø:	øʔ						
front	high		i	i:	iʔ	ĩ	ik	im	ip	iu	iŋ
front	high	lip rounding	y	y:	yʔ	ỹ					
rear	middle	lip rounding	o	o:			ok	om	op		oŋ
rear	high	lip rounding	u	u:			uk	um	up		uŋ

Figure 3. Tibetan vowel list

6. The experimental steps

Before the start of the experiment, the original voice signal needs to be preprocessed, the characteristics of the voice signal are extracted and input into the algorithm, the characteristic parameters extracted in this paper are MFCC. The WAV can be converted directly into mfcc format using HTK's Hcopy tool.

```
e:\htkdata>HCopy -A -D -C hcopy.conf -S hcopy.scp
HCopy -A -D -C hcopy.conf -S hcopy.scp

HTK Configuration Parameters[14]
  Module/Tool     Parameter                  Value
#                 ENORMALISE                 FALSE
#                 NUMCEPS                       12
#                 CEPLIFTER                     22
#                 NUMCHANS                      26
#                 PREEMCOEF               0.970000
#                 USEHAMMING                  TRUE
#                 WINDOWSIZE         250000.000000
#                 ZMEANSOURCE                 TRUE
#                 SAVEWITHCRC                 TRUE
#                 SAVECOMPRESSSED             TRUE
#                 SOURCEFORMAT                 WAV
#                 TARGETRATE         100000.000000
#                 TARGETKIND            MFCC_0_D_A
#                 SOURCEKIND              WAVEFORM


HTK Configuration Parameters[14]
  Module/Tool     Parameter                  Value
                  ENORMALISE                 FALSE
                  NUMCEPS                       12
                  CEPLIFTER                     22
                  NUMCHANS                      26
                  PREEMCOEF               0.970000
                  USEHAMMING                  TRUE
                  WINDOWSIZE         250000.000000
                  ZMEANSOURCE                 TRUE
                  SAVEWITHCRC                 TRUE
#                 SAVECOMPRESSSED             TRUE
                  SOURCEFORMAT                 WAV
                  TARGETRATE         100000.000000
                  TARGETKIND            MFCC_0_D_A
                  SOURCEKIND              WAVEFORM
```

Figure 4. HTK's Hcopy Tool interface

Label file: HTK defines its own label file format, supports multi-layer labeling, syllable layer, rhythm layer, phoneme layer, in HTK label file editing tool is Hled. Establish a dictionary for the voice unit corresponding to the HMM, in order to list the voice unit corresponding to the HMM.

6.1 Define the HMM

In HMM, quintuple is used as the basic parameters, including observation sequence, hidden state, initial probability, transition probability and launch probability.

In HTK, HMM prototype of basic unit is defined, and basic parameters include HMM topology. State number, mixed gaussian number, universal parameter dimension, etc. The model is a horizontal HMM structure, and the sequence of each speech signal corresponds to the HMM model. The following figure shows the training mode of HMM model.[4]

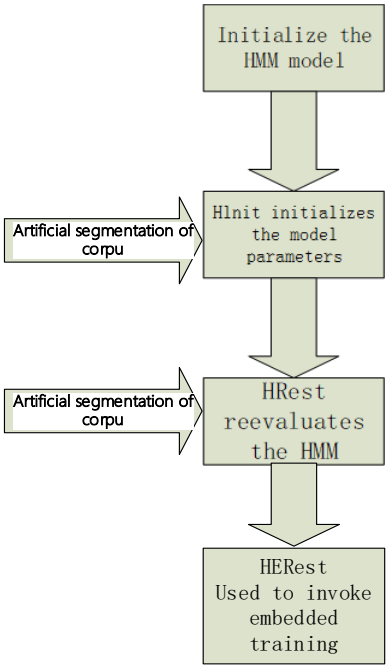

Figure 5. HMM's training flow chart

The parameters are estimated in the case of given observation sequence and corresponding state sequence.

In the first step, HTK is used to train HMM models in corpus, Hlint is used to initialize HMM models, viterbi algorithm is used to optimize parameters, Hrest is used to reevaluate forward and backward algorithms of each HMM, and forward and backward algorithms are used to adjust model parameters. Each training file must correspond to an annotation file.

In the second step, the prepared artificial segmentation boundary (200 sentences) is used to train the boundary model, and the parameters are constantly adjusted during the test to optimize the segmentation performance. After training the HMM model, the viterbi algorithm is used to process and segment the data. The segmented signal data is then sent into the HMM model for repeated training to improve the accuracy of the model.

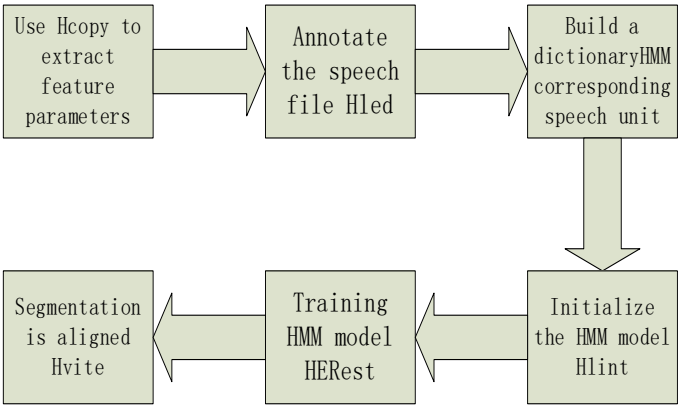

Figure 6. The experimental process

The third step, after training, prepare the speech data to be marked, generate the file list, and extract the parameters.

The fourth step is to use HTK's Hvite tool to align and shard the speech signals after parameter extraction.

The fifth step is to process the segmented data and check the annotation file of Textgard corresponding to the segmented speech signal.

7. Experimental results and analysis

This article selected 100 sounds pure and experiment time similar sentence, there are 200 sentences after artificial segmentation, a background in linguistics staff in order to maintain the consistency of the segmentation scale artificial segmentation of only one person (with the Tibetan linguistics background), the experiment of automatic segmentation of the point in time compared with the original artificial segmentation point in time. [5]Find the error between the two, and set the start time of automatic sharding as T_S and the end time of automatic sharding as T_e The start time of artificial sharding is set as t_s, and the end time of artificial sharding is set as t_e, and the error formula (1) between the two is,

$$e=|(T\_e-T\_s)-(t\_e-t\_s)|\times 1000(ms), \quad (1)$$

This formula calculates the error range. [6]

The following table shows the experimental results of automatic segmentation

Table.1. The error range of phoneme segmentation

Error / **correct**	**Error range**				
	e≤5(ms)	*e≤10(ms)*	*e≤15(ms)*	*e≤30(ms)*	*e≤50(ms)*
percentage	35.64%	46.79%	68.77%	71.31%	80.19%

In *Table.1*, e is the error, and the units of error are milliseconds.

According to the experimental results, the error is less than or equal to 5ms only accounts for 35.64%. From the perspective of accuracy, correct phoneme segmentation only accounts for 30% of the total factors. Phoneme segmentation accuracy is far from enough; the phoneme segmentation error less than or equal to 10ms (including the part less than 5ms) has not reached half of the experimental data. From this point of view, the matching degree between phoneme and model is not enough, which indicates that the training of HMM model is far from enough. When the error is less than or equal to 15ms, the accuracy rate is more than 60%, indicating that most of the phoneme matching errors are within 30ms, and the accuracy rate of segmentation errors increases slowly. This indicates that the corpus training is effective, but the corpus training is not enough. In the future, the number of manually annotated corpus should be increased to dozens of hours, and the training of HMM model should be strengthened. Further improve the accuracy of automatic segmentation.

8. Prospect

The experimental results have not reached the ideal state. Because the corpus size of training is far from enough, the training degree of the model cannot reach the accuracy of manual segmentation. According to the characteristics of the shard language, the Hvite tool in HTK can be adjusted to improve the shard efficiency. With the mandatory alignment function of Hvite of HTK, all phonemes in the speech signal can be marked quickly, maintaining the consistency of speech segmentation, but the accuracy is inferior to manual work. Combining the advantages of the two, HTK can be combined with Praat, a manually labeled platform. HTK can conduct model training and automatic segmentation, and the segmented voice signals of HTK can be sent to Praat platform for manual modification and adjustment. Combined with the advantages of the two, the unity and accuracy of speech segmentation can be greatly optimized.

Acknowledgements

This work was financially supported ByNorthwest University for Nationalities 2019 Graduate Research Innovation Project (No.Yxm2019120) and NSFC grant fund (No. 11564035). Corresponding author: Yonghong Li

References

[1] Zhang jinxi, li yonghong, Shan guangrong, li guangguang, jiang jing. “Research on automatic segmentation algorithm of single and triphoneme in Tibetan for speech synthesis” [J]. Computer application research, 2013, 30 (11): 3272-3275.

[2] Htkbook.http://users.ece.gatech.edu/antonio/htkbook/htkbook.html.

[3] GAO Lu, YU Hong-zhi, LI Yong-hong, et al.”Study on SAMPA_ST for Lhasa Tibetan and realization of automatic labelling system. Proc of International Conference on Image Analysis and Signal Processing” “.2010

[4] Teller V. Review of "Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition" by Daniel Jurafsky and James H. Martin. Prentice Hall 2000 [J]. Computational Linguistics, 2000, 26 (4): 638-641.

[5] Dujia.”Application of HMM in parametric speech synthesis system [D]. Shanghai jiaotong university” 2008

[6] Wu yijian.”Research on speech synthesis based on hidden markov model” [D]. University of science and technology of China, 2006.