The Phoneme Automatic Segmentation Algorithms Study of Tibetan Lhasa Words Continuous Speech Stream

ZHANG Jin-xi[1, a], YU Hong-zhi[1], MA Ning[1], LI Zhao-yao[1]

[1]Key Lab of China's National Linguistic Information Technology, Northwest University for Nationalities, Lanzhou, 730030, China

[a]happy0happy@163.com

Keywords: Tibetan; Corpus; Phoneme automatic segmentation

Abstract: In this paper, we adopt two methods to voice phoneme segmentation when building Tibetan corpus: One is the traditional artificial segmentation method, one is the automatic segmentation method based on the Mono prime HMM model. And experiments are performed to analyze the accuracy of both methods of segmentations. The results showed: Automatic segmentation method based tone prime HMM model helps to shorten the cycle of building Tibetan corpus, especially in building a large corpus segmentation and labeling a lot of time and manpower cost savings, and have greatly improved the accuracy and consistency of speech corpus annotation information.

Introduction

Corpus building including the corpus design, sound recording and sound segments cut grading. The only design a complete corpus to different engineering phonetics service, and a corpus designed scientific performance in the corpus contents of selected, for at Tibetan Lhasa words, Tibetan sound associated structure to consider, tones, intonation pattern, phonetic Changes phenomena and other aspects. In the Tibetan speech synthesis system, speech unit segmentation label is based on syllable or smaller than the syllable primitives.

Literature [1] for English voice library phoneme boundary cut points, Selected characteristics of the HMM, optimized model parameters and model clustering. Literature[2] has proposed a method for mute automatically added, after a rough segmentation, by rules set, short-term energy, the short-time zero-crossing rate and algorithm correction several steps, added the mute to the corresponding text, achieved good results. These improvements improved the accuracy of the phoneme segmentation in varying degrees.

In this paper, we consider the more complex structure of Tibetan speech; have consonant cluster、the consonant ending、compound vowel phonetic phenomena, etc. However, the artificial segmentation and labeling work extremely cumbersome、time-consuming, the nature of the work will be distributed label attention, if using the method of the artificial segmentation and annotation, then build library need quite a long period, it will affect the system of research and development time. Therefore, looking for a highly efficient automatic segmentation method is particularly important.

Data Preparation

Corpus Text Design

How to select a valid recording text is the key to the design of the corpus. The corpus designed to take into account the two levels of sound segments and rhythm, considering the combination of tones, the tone of the sound segments associated phenomena, voicing with duration of statement in corpus select. In the design of Corpus, an important principle is the exhaustion small corpus covering as much of the natural language phenomenon. The original text corpus design selected from the 2007 "Tibet Daily" text, corpus coverage between different syllables of two-syllable and most of the triphone. In rhythm, the corpus to meet the various combinations of two, three or four

Published by Atlantis Press, Paris, France.

syllables group contains a wide range of Tibetan sentence. We temporarily adopted the Greedy algorithm popular in the industry as we choose to Corpus basis, to maximize natural language coverage levels [3].

Recording of the corpus

We use 2000 sentences of 100 press releases of Tibet daily in 2007 as transcriptions. Recording requirements are as follows[3,4]:The announcer recording officer for the professional level, the young people of the Lhasa dialect accent；Sampling rate of 16K, precision 16-bit, single-channel recording, recording software Audition, the audio file wav format；Require normal speed, average 6-7 syllables per sec；Normal tone, not with any emotional recording.

The phoneme list of determined

The determination of the phoneme list is designed based on Key Laboratory of China's National Linguistic Information Technology Northwest University for Nationalities—SAMPA_ST [5], on this basis, coupled with length mute sil (pausing between sentences) and sp (sentence pause), forms the phoneme list which HTS system needs, As shown in Table 1.

Table 1 the phoneme list of its corresponding representative symbol in the automatic voice segment segmentation

Phoneme	Symbols	Phoneme	Symbols	Phoneme	Symbols	Phoneme	Symbols	Phoneme	Symbols
c	ca	th	th	ap	ap	up	up	iu	Iu
ch	ch	t□	td	au	au	uŋ	un	iŋ	in
….	….	….	….	….	….	….	….	….	….
s	sa	ak	ak	uk	uk	im	Im	Mute	sil
t	ta	am	am	um	um	ip	Ip	Pause	sp

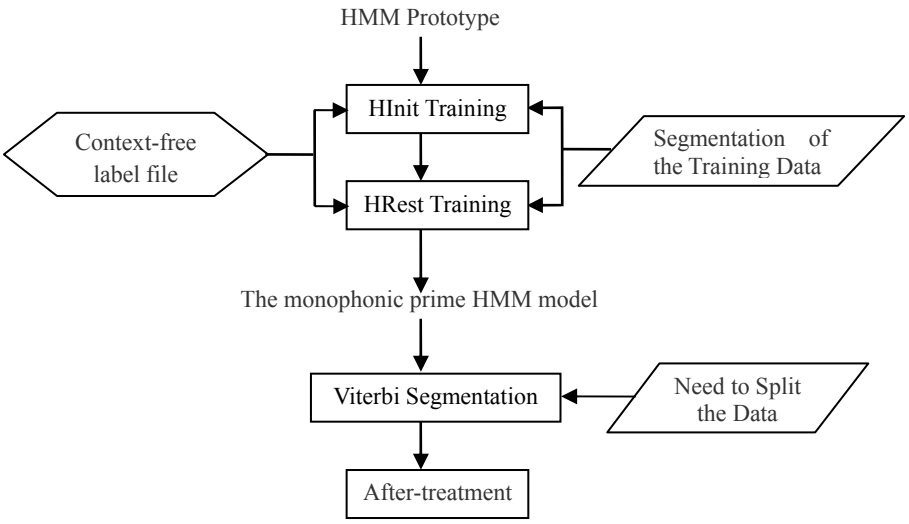
Segmentation Method

Artificial Segmentation Method

Artificial segmentation method is the most traditional speech segmentation method. This article artificial segmentation is done independent by two students with some Tibetan language and linguistics foundation, through the analysis of the characteristics of voice, based on their own experience to determine the voice of the cut-off point. The two will be different segmentation results with a voice, the same person there will be deviations with segmentation results of a voice in different time periods. But the advantage is that manual segmentation to a certain extent able to guarantee the accuracy of the voice data segmentation. We use Praat speech analysis software to mark phoneme (segment), write scripting program to batch segmentation.

Automatic Segmentation Method

HMM (Hidden Markov Model) is a more mature voice field, wide application of a method. The Tibetan phonemes automatic segmentation algorithm on the basis of the HMM algorithm, training and segmentation of the speech model with the help of HTK toolbox. First, HMM training based on artificial segmentation data, then using HMM model training to segment voice data. HMM training and automatic segmentation process is shown in Fig. 1.



Published by Atlantis Press, Paris, France.

Fig.1 HMM training and the flow chart of the automatic segmentation

Analysis of experimental results

The segmentation experiments is based on a Tibetan Lhasa dialect corpus, corpus is mainly declarative, select from 100 as a test subset. We regard two annotation consistent results as the reference standard for the evaluation, automatically between the boundaries of the segmentation boundary points and the reference standard deviation is within the range of ± T, we think is correct, otherwise it is wrong, where T is called the fault tolerance threshold. The total number of correct automatic segmentation boundaries percentage of segmentation boundary called this fault tolerance threshold under automatic segmentation accuracy. We usually study is fault tolerant threshold for 20ms automatic segmentation accuracy [6].

This experiment adopts 1500 sentences as training data, 100 sentences as a test for automatic segmentation. The first is to establish a single phoneme model, then compare and analyze the automatic segmentation results with manual segmentation results in detail, that is comparison each phoneme period one to one correspondence, calculated error range.For example, the first letter of a sentence as "na",in artificial segmentation, the start time t1, the end time t2, in automatic segmentation, start time T1, the end time T2,then the error range of the phoneme W available formula $W = |(T_2 - T_1) - (t_2 - t_1)| \times 1000(ms)$ .Single phoneme automatic segmentation error range of statistics as shown in Table Ⅱ.

Table 2 mono prime automatic segmentation error range Tables

Sentence Serial Number	Phoneme Corresponding Symbol	Artificial Segmentation Time (s)		Automatic Segmentation Time (s)		Error Range W(ms)
		Start Time t1	End Time t2	Start Time T1	End Time T2	
001	sa	1.000	1.865	1.100	2.000	35
	fw	1.865	2.945	2.000	3.100	20
	ca	2.945	3.330	3.100	3.500	15
	yu	3.330	4.324	3.500	4.500	6
	xa	4.324	5.514	4.500	5.700	10
.....						
100	ng	17.300	17.895	17.700	18.300	5
	ta	17.895	18.255	18.300	18.800	140
	yu	18.255	19.405	18.800	20.000	50
	pa	19.405	19.740	20.000	20.400	65
	aj	19.740	21.030	20.400	21.700	10

From the statistical results, 100 sentences contain a total of 4783 phonemes. Statistical results as shown in Table 3.

Table 3 monosyllabic prime HMM model-based automatic segmentation results

Error Range	≤5ms	≤10ms	≤20ms	≤30ms	≤40ms	＞40ms
Phoneme Number	2176	2987	3689	3919	4135	648
percentage of the total number	45.49%	62.45%	77.13%	81.94%	86.45%	13.55%

Average allowable error range is expressed with W, that is $\overline{W} = \sum_{i=0}^{4783} W_i / 4783 = 21.83ms$.

Automatic segmentation accuracy $R = \frac{m}{M} \times 100\%$, Where in M represents a sentence number of the phoneme, m represents the sentence automatic segmentation with manual segmentation consistent with the number of phonemes, the total average segmentation accuracy $\overline{R} = \sum_{i=1}^{100} R_i / 100 = 80.69\%$.

Summaries

This article by experimental comparison of two segmentation methods, found that automatic segmentation method based on the model of HMM monosyllabic prime the Tibetan speech synthesis corpus phoneme layer is superior to manual segmentation method. However, due to the richness of the voice signal voice phenomenon will take different measures, so HMM modeling

Published by Atlantis Press, Paris, France.

spectral parameters only far from being able to solve some actual voice. Therefore, there are some problems in the number of HMM-based auto-cut separation. In order to make higher segmentation accuracy, we need to take certain measures to cope with the late part of the treatment on the results of the segmentation.

References

[1] WANG Li-juan, CAO Zhi-gang. Automatic Phonetic Segmentation Using HMM Model [J]. Journal of Data Acquisition & Processing, 2005, 20(4):381-384.

[2] CHEN Kai, CAI Pei-qi. Silence Insertion in HMM-based Chinese Automatic Segmentation [J]. Computer Engineering.2004, 30(9):40-41.

[3] ASKAR Rozi. Research and Implementation of HMM based Uyghur Speech Synthesis System [D]. Xinjiang University.2008.

[4] Gao Lu,Yu Hongzhi, et al. Study on SAMPA_ST for Lhasa Tibetan and Realization of Automatic Labeling System[C].IASP 2010.Vol I, pp.133-137.

[5] WANG Li-juan, CAO Zhi-gang. Automatic Segmentation for TTS Units [J]. Microelectronics and Computer, 2005, 22(12):8-11.

[6] Htkbookhttp://users.ece.gatech.edu/~antonio/htkbook/htkbook.html[OL]

Published by Atlantis Press, Paris, France.