Advances in Computer Science Research (ACSR), volume 61
7th International Conference on Education, Management, Computer and Society (EMCS 2017)

Analysis of Tibetan-language Speech Technology

Xiaowei Bai[1, a], Xingyu Tao[2, b], Yiwen Wu[1, c] and Wei Xiang[1, d*]

[1]College of Electrical and Information Engineering, Southwest University for Nationalities, Chengdu, Sichuan, 610225, P. R. China

[2]glasgow college, University Of Electronic Science And Technology Of China

[a]634811696@qq.com, [b]1660261925@qq.com, [c]546317563@qq.com, [d]3730544@qq.com

*The corresponding author

Keywords: Tibetan character recognition; TTS; RS; Speech technology; Speech Recognition

Abstract. This paper studies the speech technology (Speech Recognition and Text To Speech) for Tibetan. Recognition of Tibetan characters is a significant module of multi-language information processing system in China. Speech is the most convenient and natural way of communication. Owing to the special structure of Tibetan characters, the SR and TTS of traditional Tibetan characters face problems of low recognition rates and poor recognition effects. Through an in-depth study on the features of Tibetan characters, we finished this paper. This paper briefly introduces the development and the basic principles of speech technology, and then we provide an analysis of SR and TTS of Tibetan. We give the structure diagrams of the SR and the TTS technologies and introduction of the key module. Finally, we discuss the prospects of application in academic writing, demands and challenges of the Tibetan speech technology briefly in many areas.

Introduction

Tibetan traditional culture is one of the most wonderful works among cultural treasures of the human world[1]. One out of a multitude of outstanding achievements in Tibetan culture is the Tibetan language, which is the precious heritage in the history of world culture[2]. Speech Technology has gradually expanded and spread quickly in the field of telecommunications, such as voice information service applications. Speech Technology has two branches, one is the technology that converts human speech into text, known as Speech Recognition (SR) and the other is to convert text into natural speech, called Text-to-Speech (TTS). Like other languages recognition technologies, the Tibetan printed or written can be distinguished and transacted automatically by computers and people's laborious intensity can be greatly reduced through this kind of ideal tool with high-speed text input. We can describe Tibetan recognition visually as a technique that lets the computer ‘know’ the Tibetan language[3]. It will greatly improve the stability and development in economies, culture and education for the Tibetan compatriots. Speech Technology relates to many interdisciplinary sectors, including acoustics, phonetics, linguistics, computer science theory, information theory and digital signal processing. The requirements of speech recognition in speech corpus are not as strict as speech synthesis stands at implement. For speech recognition influenced by environment, language and enunciation, the recognition rate is not high in the latter part of the study process. Therefore, the technology of speech recognition can not be generally applied and needs to be improved. Speech synthesis, requirements for voice libraries are relatively high and needs to annotate to corpus. Speech synthesis does not need a lot of manpower or resources but does have higher requirements for the basics of language and speech[4]. In recent years, the rapid development of information processing technology in computers provides wide application prospects and development directions for speech technologies of Tibetan, such as speech recognition and speech synthesis, but it also faces many challenges at the same time.

Application of Voice Technology

Speech Translation has always been the goal pursued by people. After the invention of computers in the 1950s, people have expected mutual translations between different languages by computers[5]. Speech Translation System's basic function is to translate one language into another language. In order to achieve this goal, the speech translation system should contain three basic parts: speech recognition and understanding, machine translation, speech synthesis. Speech recognition and understanding is to identify, understand the source language input as well as get the text description and formal description of the source language. Machine Translation is the translation of the speech recognition with the text of target language output. Speech synthesis is based on the results of machine translation, producing discourses of the target language [6].

Through voice processing and filtering out some of the interferences or noises in the environment, we can improve the clear and identifiable degree of voices.

The Analysis of Key Technologies for Tibetan Speech

Recognition, machine translation, and speech synthesis are prime technologies of translation systems. Achieving speech input and another speech output are essential to any one of these technologies. At present, these three technologies have made great progress and obtained remarkable results which brought utmost convenience to human study, travel, work and life.

SR. SR module is the foundation of the whole speech Translation System, which is responsible for receiving speech signal after preprocessing and making voice signals into speech primitives (such as syllables, phonemes, etc). Then we convert the speech primitives to source language text stream based on knowledge of grammar and constitutive rules of voice. Speech recognition is very important for whole translation systems, especially accuracy of recognition. For years, many scientists have been concerned with speech recognition technology. In recent years, to make a significant progress at home and abroad on building a large vocabulary of isolated words recognition and the recognition rate is higher, many people make great effort on it. Although great progress in continuous speech recognition has been made, but far from the recognition rate of isolated words. Main reasons as follows:

(1) User diversity: differences in gender, age, social backgrounds, dialect, channel length, lead to different users having different pronunciation of the words.

(2) Variability of the same users: factors such as different emotional conditions, tones and states of health lead to different pronunciation by the same user.

(3) Vagueness of speech: patronymic and homonym, with the environment and different locations, their voice characteristics make a big difference, including volume, speed, accent, pronunciation, and approach to reading.

(4) Noise interference: although we preprocessing the speech before speech recognition, we can’t completely eliminate noise interference.

Aiming to solve these problems, the current continuous speech recognition processing methods are proposed:

(1) Segmentation Method: The recognition technology has made great progress in isolated word area. Therefore, cut the speech signals flow into simple basic units of speech and the isolated words recognition technology can be applied to continuous speech recognition. The whole process is shown in Fig. 1.

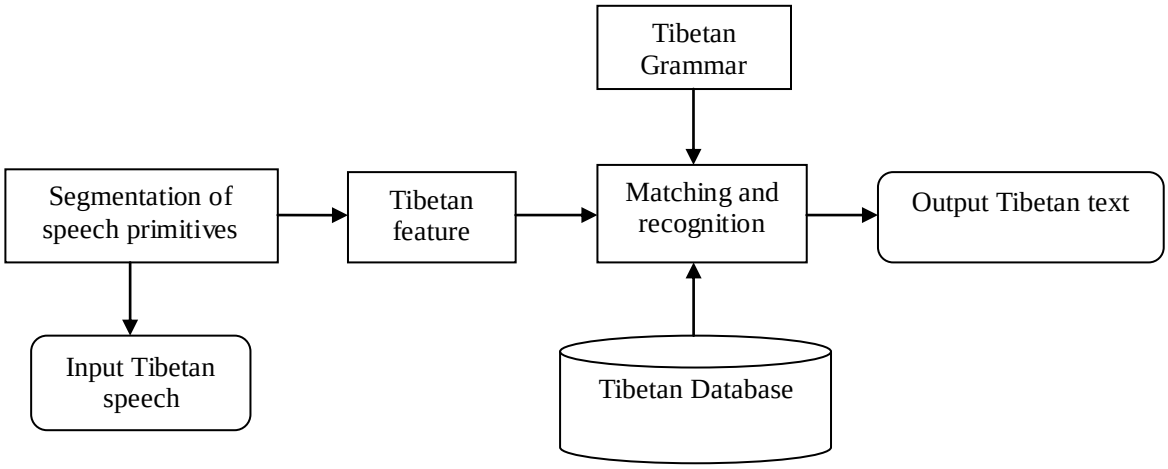

Figure 1. Cutting method for Tibetan speech recognition

This approach is applicable to Chinese as it is monosyllabic. For the western languages such as English, the division error rate is high. So more linking phenomena in English and other western languages can't be cut down to words by voice alone.

(2) The way of keywords: first, choose some sentence patterns and set a thesaurus. Then Identify keywords in the whole sentence and recognize the keywords. Once the keywords recognition is successful, we will no longer consider other components. About objective of the whole sentence recognition, the method has the advantage of allowing input sentences not standardized, but the drawback is the result of recognition may not be accurate. Because of a large number of non - standard and oral sentences, the keyword method used in current speech Translation System is more appropriate.

Space TTS. The speech synthesis module is to transform target languages from machine translation module into the corresponding speech. The main design philosophy of the methods includes synthesis rules-driven (rule-based) method and data-driven (data-based) method [7] and is shown in Fig. 2

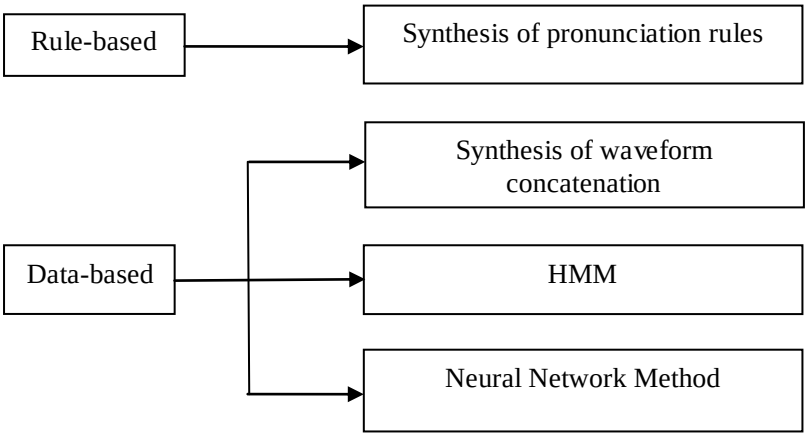

Figure 2. Speech synthesis method

The former is based on the parameters of the vocal organs and physical processes. Direct simulation of the pronunciation of people[8] is complex and the nature of synthesized speech is low-levered and rarely used. Waveform technology is based on text analysis information, electing appropriate unit from prerecorded voice library and we can get the final synthesized speech from it. As the final units are copied directly from the recording voice library, the method can maintain the original pronunciation from people and achieve high natural. The whole process is shown in Fig. 3.

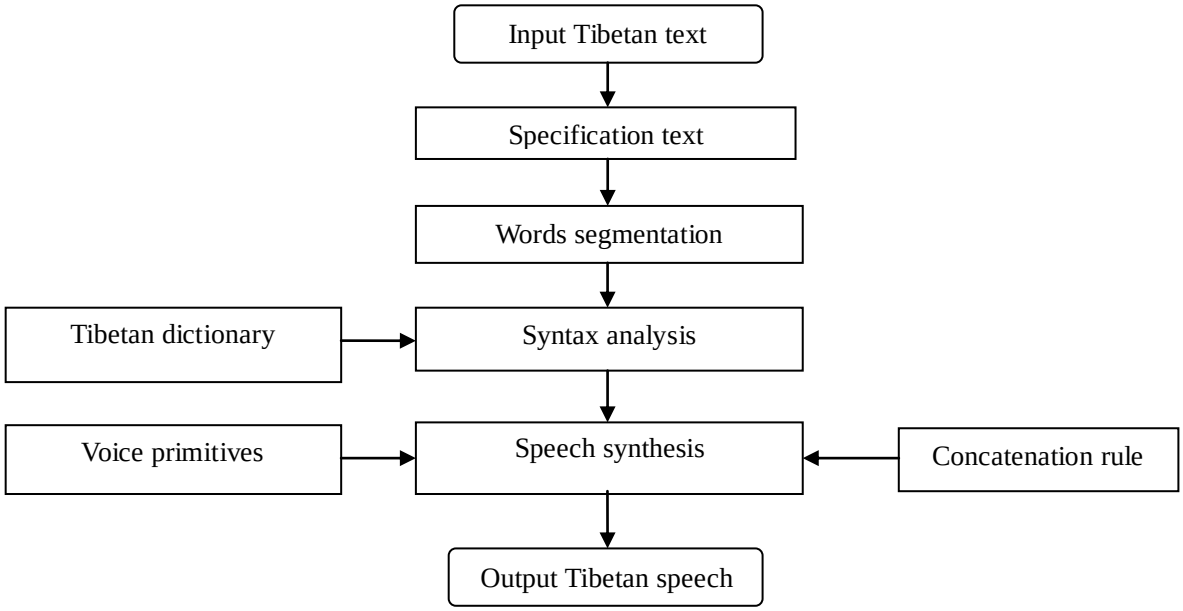

Figure 3. TTS of Waveform splicing method

The synthesis process is divided into 3 steps:

(1)Text process: process the input text and remove the errors or unpronounceable characters.

(2)Philology disposal: makes sure the severity of the transforming tone expending on the text structure, composition and punctuation marks on different locations.

(3)Acoustic process: we can achieve speech synthesis by selecting voice from the voice library primitives and connect them based on connection rules.

The advantages of this approach are high quality, high similarity of tone, better to learn the pronunciation of the natural rhythm. The disadvantage is the demand for larger speech database support. Therefore, the way of waveform concatenation is more dependent on the quality of speech corpora, size, the smallest unit, etc. With the development of Data Mining technology in the computer field, many statistical methods and artificial neural network technologies in data processing applications are successful. In this context, there have been based on data - driven text analysis methods at home and abroad [9, 10]. For example, using hidden a Markov Model (HMM: Hidden Markov Model) and neural networks method (Neural Network Method) [11, 12].

Summary

In this paper, we introduce speech technology based on analyzing the features of Tibetan language thoroughly. Tibetan traditional culture is one of the most wonderful works among cultural treasures in the human world, and speech is the most convenient and natural way of communication speech technology in economies, education, commerce, tourism and other fields has made great progress and has broad prospects. Its application in translation system is to promote mutual learning and communication between users in different languages. Especially in recent years, there is a lot of translation software with voice functionality, providing a platform for language learning, listening, and speaking greatly improving the efficiency of people learning new languages. Of course, speech technology is still facing many problems and challenges, such as user adaptation for speech recognition, the precision of consecutive translation, tone of voice processing and naturalness for synthesis. But as speech technology is combined with other technologies, it will be gradually improved and be promoted and applied in broader fields.

Acknowledgements

This work was financially supported by Innovative Research Team of the department of Sichuan Province(15TD0050) and SWUN students' innovative training program (No.S201610656077).

References

[1] Chenxing Z, Bing Y. The Vast Vistas for Development of Tibetan Information Processing Technology [J]. Journal of Qinghai Normal University (Natural Science), 1999, (01): 9-15.

[2] Gang W, Xijiacuo D, Heming H. Printed Tibetan Character Recognition Technology[J]. Journal of Qinghai Normal University (Natural Science), 2006, (01): 32-37.

[3] Yuzhong C, Shiwen Y. The Tibetan Information Processing Technology Research Status and Prospect [J]. China Tibetology, 2003, (04): 97-107.

[4] Zhang Bin，Quan changqin，REN Fuji.Speech synthesis method and development summary [J]. Journal of Chinese Computer Systems, 2016, (1):186-192.

[5] Wei Maocheng, Zhang Sen, Zhang Fenghou. Analysis of key techniques of English - Chinese speech Translation System [J]. Journal of Shandong Institute of Building Materials, 1998, 12, (3): 226-230.

[6] Li Jing Jiao, Yao Tianshun.Medium vocabulary English - Chinese speech Translation System [J]. Mini - Micro Systems, 1998, 19, (8): 44-48.

[7] Youcef tabet, Mohamed boughazi. Speech synthesis techniques．a Survey[C].7th International Workshop on Systems，Signal Processing and their Applications (WOSSPA), 2011: 67-70.

[8] Huang Nanchuan, Deng Zhenjie, Wang, Zhang haojian. Research and Development of Speech Synthesis techniques [J].Journal of Nor th China Institute of Astro nautic Engineering, 2002, 12(3):36-39.

[9] SH IN YA NAKAJMIA. Automatic Synthesis Unit Generation for English Speech Synthesis Based on Multi - Layered Context Oriented Clustering [J]. Elsevier Science Publishers B V, 1994, 14 (4):313-324.

[10] TAKAYOSH IYOSHIMURA, KEIICHOKUDA,TAKASHIMASUKO,eta.l Incorporating a Mixed Excitation Model and Post filter into HMM -Based Text To-Speech Synthesis [J]. System s and Computers in Japan, 2005, 36 (12):43-50.

[11] ANDREJ LJOLJE, JULIA HIRSCHBERG, JAN P H VAN SANTEN. Automatic Speech Segmentation for Concatenative Inventory Selection [C] // Progress in Speech Synthesis. [S..l] :Springer Verlag, 1997:315-321.

[12] COLIN W W IGHTMAN, AV ID T TALKIN. The Aligner:Text To Speech Alignment Using Markov Models[C] // Progress in Speech Syn thesis. [S..l] : Springer-Verlag, 1997:322-332.