Speech Signal Feature Extraction Method of Tibetan Speech Synthesis System Based on Machine Learning

Ze-guo Liu(✉)

Key Lab of China's National Linguistic Information Technology, Northwest Minzu University, Lanzhou 730000, China

Abstract. In order to improve the accuracy of Tibetan speech synthesis, a feature extraction method of Tibetan speech synthesis system based on machine learning is proposed. Based on the analysis of Tibetan speech text content, the construction of speech synthesis system is realized. By judging the level of Tibetan prosody, a synthetic encoder is designed to realize the feature extraction of Tibetan speech signal. According to the experimental results, under the condition of normal speaking speed and identical Tibetan speech content, the Tibetan speech synthesized by the speech signal feature extraction method of Tibetan speech synthesis system based on machine learning is more accurate.

Keywords: Machine learning · Tibetan · Synthesis system · Signal extraction

1 Introduction

Tibetan language is a national language with a long history. Tibetan language is mainly divided into: Wei Tibetan dialect, Anduo and Kangxi dialect. Tibetan language is not only used in Tibet, Qinghai, Gansu, Sichuan, Yunnan and other parts of China, but also used in Nepal, India and other countries. Tibetan Lhasa dialect is a kind of Wei Tibetan dialect, which is mainly used by Tibetan people in Lhasa city and its surrounding areas [1]. Lhasa dialect is the most used and influential Tibetan dialect in Tibetan areas. Therefore, Tibetan Lhasa dialect is also known as Tibetan "Putonghua", and its pronunciation has some basic characteristics, such as: there are no voiced initials and blocking initials in Lhasa dialect, and complex consonant initials are relatively rare [2, 3]. In reference [4], a speech signal feature extraction method based on EMD for Tibetan speech synthesis system is proposed. Firstly, the speech signal is decomposed into several intrinsic mode components by using empirical mode decomposition (EMD), and then these components are processed by FFT to obtain more detailed signal division, The obtained speech sequence of Tibetan speech synthesis system is mixed with the first-order difference and short-term energy features of Tibetan speech synthesis system, and then used in the following experiments. The experimental data show that the speech recognition rate of this algorithm is significantly higher than that of traditional Tibetan speech synthesis system

S. Liu and X. Ma (Eds.): ADHIP 2021, LNICST 416, pp. 470–479, 2022.
https://doi.org/10.1007/978-3-030-94551-0_37

under different test environments. However, the effect of Tibetan speech synthesized by this method is not good.

In view of the above problems, this paper proposes a speech signal feature extraction method based on machine learning for Tibetan speech synthesis system. Tibetan speech synthesis system based on machine learning is a technology that uses computer and some devices to realize text to speech. The aim of the feature extraction method of Tibetan speech synthesis system based on machine learning is to endow human machine with the ability of speech, so as to realize the speech communication between human and machine. Speech signal feature extraction of Tibetan speech synthesis system is an interdisciplinary subject, which involves linguistics, psychology, machine learning and many other fields. The typical Tibetan speech synthesis system mainly includes two parts: the front-end and the back-end.

2 Speech Signal Feature Extraction Method for Tibetan Speech Synthesis System

2.1 Tibetan Text Analysis

Based on machine learning Tibetan speech synthesis system, a prototype system with Tibetan speech synthesis is designed in the embedded platform, and Tibetan text content is input to the system, and the output will be continuous and fluent Chinese Tibetan speech with a certain degree of naturalness [5]. The feature extraction of Tibetan speech synthesis system based on machine learning adopts Xscale px255 of inter company. The specific structure of Tibetan speech synthesis system based on machine learning is shown in Fig. 1.

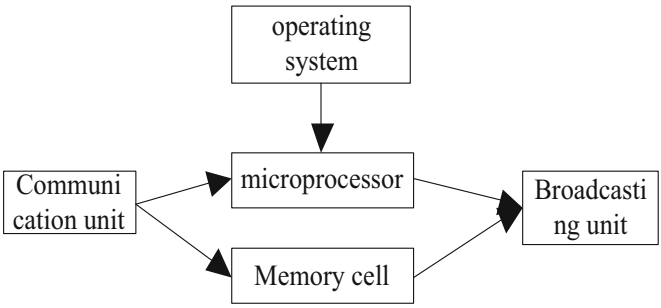

Fig. 1. Tibetan speech synthesis system

Based on machine learning Tibetan speech synthesis system, the feature extraction of speech signal is realized by arm 5te kernel technology. It provides 16 DMA channels to provide data for peripheral devices, and each channel has a special FIFO. When more than half of the FIFO data is received, DMA is triggered for data transmission, which can not only transfer the data to the memory quickly, but also greatly improve the efficiency [6].

The feature extraction of typical Tibetan speech synthesis system mainly includes two modules: front-end text analysis and back-end speech synthesis [7]. The front-end of Tibetan speech synthesis system is mainly responsible for analyzing the input Tibetan text, and then extracting the information needed by the back-end modeling. Therefore, the back-end of Tibetan speech synthesis system builds an acoustic model according to the front-end text processing results. The speech signal feature extraction method of Tibetan speech synthesis system based on machine learning. The specific speech signal feature extraction structure of Tibetan speech synthesis system is shown in Fig. 2.

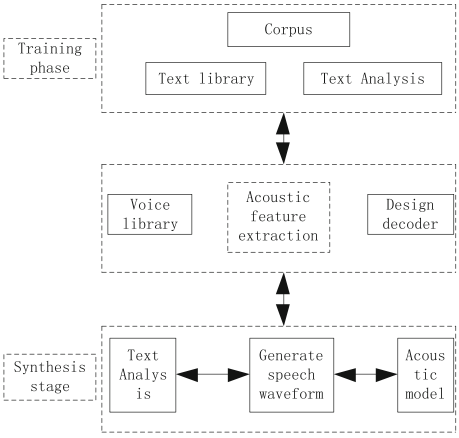

Fig. 2. Structure diagram of speech signal feature extraction in synthesis system

Based on machine learning Tibetan speech synthesis system, speech signal feature extraction is designed into two stages: training and synthesis [8]. The training stage is to analyze the Tibetan text in the corpus, send it to the encoder module, extract the acoustic characteristic parameters from the speech database, and then send it to the decoder; get the acoustic model through the decoder; in the Tibetan speech synthesis stage, send the text to be synthesized into the acoustic model through text analysis, and then generate the speech waveform corresponding to the text sequence through the acoustic model.

2.2 The Judgment of Tibetan Prosodic Level

Prosody is mainly an auditory feature and a psychological quantity. Prosody contains the speaker's intention information and the hearer's perception information, which is very useful in helping the hearer understand the language and intention [9]. The speech signal feature extraction of Tibetan speech synthesis system based on machine learning is described by its corresponding acoustic features, such as fundamental frequency, duration, amplitude and frequency spectrum, and four speech auditory features, such as pitch, duration, intensity and timbre. In addition, the appropriate pause in Tibetan speech synthesis is also a very important component of prosody [10]. The basic structure of Tibetan texts is shown in Table 1.

Table 1. Classification of Tibetan characters

Positive addition			Add after		
Character character	Timbre	Pronunciation	Character character	Timbre	Pronunciation
Positive	Strong	Airless sound	Positive	Strong	Airless sound
Neutral	Neutralization	Voiceless aspirated	Neutral	Neutralization	Voiceless aspirated
Negative	Weak	Voiceless consonant	Negative	Weak	/
	Very weak	Secondary voiceless		/	/

The prosodic feature of Tibetan speech synthesis system is not only to complete the pronunciation of consonants and vowels, but also to pay attention to the factors of Tibetan tone, strength and duration. However, these factors can not exist alone, but are attached to the consonants and vowels in Tibetan speech synthesis system [11]. The prediction and judgment of Tibetan prosodic level is to better realize the feature extraction of Tibetan speech synthesis system based on machine learning.

The characteristics of Tibetan prosody are to convey information about fundamental frequency, duration change and amplitude [11]. It is difficult to measure, model and simulate the intonation and stress in Tibetan speech synthesis system. In the speech signal feature extraction of Tibetan speech synthesis system based on machine learning, the prosodic levels from small to large are: last position, syllable, step, phonological copula, attached morpheme phrase, phonological phrase, intonation phrase and prosodic sentence.

Generally, they are simplified into prosodic words, prosodic phrases, intonation phrases, and prosodic sentences [12]. A smaller prosodic component is contained in a larger prosodic component, which forms the prosodic hierarchical structure of Tibetan

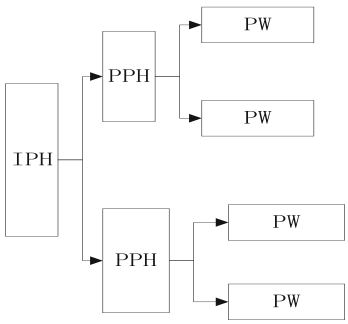

Fig. 3. The prosodic level of Tibetan speech

speech. PW is a prosodic word; PPH is a prosodic phrase; IPH is a intonation phrase. The specific prosodic hierarchy is shown in Fig. 3.

The speech signal feature extraction of Tibetan speech synthesis system based on machine learning is from the point of view of the pronunciation position and pronunciation method of Tibetan natural speech, but the prosodic structure and grammatical structure of Tibetan are not completely consistent. In the Tibetan speech synthesis system, there is a certain close relationship between the components of Tibetan prosodic words and the part of speech features, between the pronunciation position of stress and the syntactic structure, and between the pause of prosodic boundary and the characteristics of syntactic structure and the part of speech [13]. Therefore, grammar in prosodic structure plays an important role in feature extraction of Tibetan speech synthesis system based on machine learning.

In the speech signal feature extraction of Tibetan speech synthesis system based on machine learning, with the continuous development of computer, the in-depth exploration of Tibetan text information is also developing. Similar to Chinese speech synthesis information processing, Tibetan word segmentation also plays an important role in Tibetan information processing [14]. The speech signal feature extraction of Tibetan speech synthesis system based on machine learning can analyze and process the words, words, phrases, sentences, semantics, even the whole text or language in Tibetan prosody. It can also realize the automatic segmentation of Tibetan prosody. It can not only mark some grammatical features of grammatical words such as part of speech and word length automatically, but also do not need too much manual repair The efficiency is greatly improved.

2.3 Feature Extraction of Speech Signal Based on Machine Learning

The speech signal feature extraction of Tibetan speech synthesis system based on machine learning, the feature of Tibetan speech signal obtained by its encoder is linear prediction residual signal approximated layer by layer, and the synthesized speech quality increases with the rate. The Tibetan speech synthesis system based on machine

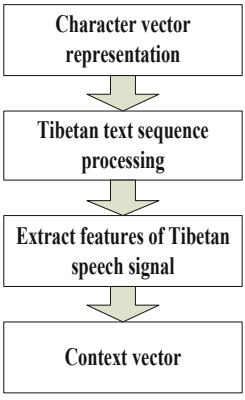

Fig. 4. Framework of Tibetan speech synthesis encoder

learning transforms a variable length Tibetan text sequence into a fixed length feature vector [15]. At the same time, the encoder in the system obtains the output value context vector of the encoder by inputting the character vector, preprocessing the Tibetan text sequence in the hidden layer, and extracting the text feature network (Fig. 4).

In the Tibetan speech synthesis system based on machine learning, the design of adaptive codebook structure plays an important role in improving the Tibetan speech synthesis encoder. The previous Tibetan speech synthesis algorithm only uses the excitation signal of the core layer to update the adaptive codebook buffer, thus abandoning the excitation signal of the enhancement layer 1 or the enhancement layer 2 with better quality [16]. Therefore, the new algorithm for feature extraction of Tibetan speech synthesis system based on machine learning is the core layer, enhancement layer 1 and enhancement layer 2, and a buffer is set respectively to update the corresponding adaptive codebook with the excitation signal of each layer, so as to achieve the best matching of coding parameters of each layer.

In order to reduce the complexity of the algorithm, the first level target vector of the core layer is analyzed by closed-loop pitch analysis, and then the best integer and fractional pitch delay are obtained. According to the analysis results of common core layer, enhancement layer 1 and enhancement layer 2, it is clear that the optimal pitch delay used in the three layers of interpolation is the same in the Tibetan speech synthesis system based on machine learning.

According to the different adaptive codebook buffers of each layer, the excitation vectors of each layer are obtained by interpolating the past excitation at the optimal pitch delay, which are recorded as vn, vn_{12} and vn_{16} respectively. At the same time, the feature extraction gain of Tibetan speech synthesis system based on machine learning is calculated. In order to save the number of coding bits, the adaptive codebook gain of the core layer is taken as the common gain of the adaptive codebook of each layer of the current subframe, which is recorded as g_p. Vector xn, xn_{12}, xn_{16}, get the specific expression, see Formula 1:

$$\left\{ \begin{array}{l} Xn_2(n) = xn(n) - g_p \cdot y_1(n) \\ Xn_{2-12}(n) = xn_{12}(n) - g_p \cdot y_{1-12}(n) \\ Xn_{2-16}(n) = xn_{16}(n) - g_p \cdot y_{1-16}(n) \end{array} \right\} \tag{1}$$

Where, $y_1(n) = vn * h(n)$ is the convolution of the core layer's adaptive codebook excitation vector and the perceptual weighted synthetic filter's unit impulse response; $y_{1-12}(n) = vn_{12} * h(n)$ is the convolution of the enhancement layer's adaptive codebook excitation vector and the perceptual weighted synthetic filter's unit impulse response; $y_{1-16}(n) = vn_{16} * h(n)$ is the convolution of the enhancement layer's adaptive codebook excitation vector and the perceptual weighted synthetic filter's unit impulse response.

Therefore, in the new algorithm of speech signal feature extraction of Tibetan speech synthesis system based on machine learning, the adaptive codebook structure makes full use of the feature that the Tibetan speech synthesis coding can generate multiple excitation signals; the excitation signals of different quality adaptive codebooks are obtained, and are respectively used for searching algebraic codebooks at different levels [17]. In the Tibetan speech synthesis coder based on machine learning, the algorithm complexity of the candidate coder is measured by the execution of millions of operations

per second, and the operations such as addition, subtraction, multiplication and division are weighted according to different weights. The specific statistical results are shown in Table 2.

Table 2. Storage complexity of candidate encoders

/	/	Storage complexity (kwords)
ROM	Coding end	14.92
	Analytic end	1.247
	Total	16.177
RAM	Coding end	13.185
	Analytic end	6.273
	Total	19.458

In the feature extraction of Tibetan speech synthesis system based on machine learning, in order to prevent over fitting of acoustic parameters in Tibetan speech synthesis system, a series of nonlinear transformations are carried out for each input frame through calculation. In the Tibetan speech synthesis coder based on machine learning, there are two hidden layers, and the layers are completely connected. Therefore, when the Tibetan speech output in the decoder is not directly converted to audio, it is necessary to introduce the machine learning output into the form of waveform expression.

3 Simulation Experiment Analysis

3.1 Experiment Preparation Activities

Speech signal feature extraction of Tibetan speech synthesis system based on machine learning, in which Tibetan speech synthesis refers to converting text information into speech output. The construction of Tibetan speech synthesis corpus plays an important role in the design of the system and the realization of feature extraction of Tibetan speech signal based on machine learning. The construction process of the corpus in the Tibetan speech synthesis system is shown in Fig. 5.

The corpus of the Tibetan speech synthesis system aims to study, develop and evaluate the synthesis and recognition of Tibetan. In this experiment, the feature extraction experiment of Tibetan speech synthesis system based on machine learning is carried out. Based on the initial manual segmentation of the half syllable boundary in the phonetic database, the HTK toolkit is used as the basic unit. After the special training, the personnel have manually checked and adjusted. At the same time, the basic frequency parameters of each syllable are modified, and the whole sound bank is marked with Tibetan prosody structure and stress level.

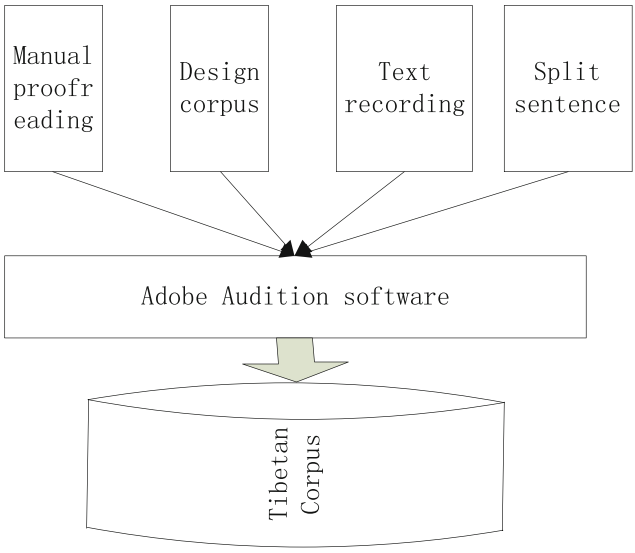

Fig. 5. Construction process of Tibetan speech synthesis corpus

3.2 Experimental Process

The front-end part of Tibetan speech synthesis system based on machine learning mainly includes text analysis module. It mainly extracts the information needed for the back-end modeling of the system from the text through the process of analyzing the language features of Tibetan text, checking the grammar rules, and searching the semantic data. Therefore, the back-end part of Tibetan speech synthesis system uses machine learning or neural network technology to build acoustic model of speech according to the front-end text analysis results, and generates target speech by using text information and trained acoustic model.

In the Tibetan speech synthesis system based on machine learning, the corpus used is the human-computer speech interaction group. In the process of traditional Tibetan corpus construction, every step of text design, text recording and sentence segmentation needs careful manual supervision and proofreading, which requires a lot of energy and time. In the Tibetan speech synthesis system based on machine learning, the application of modern technology shortens the synthesis time and improves the work efficiency.

3.3 Experimental Result

Under the condition of normal speaking rate and the same Tibetan speech content, through praat's speech analysis software, the traditional method 1 and traditional method 2 and the method of this paper are used to extract the voice signal characteristics of the Tibetan speech synthesis system, and the two are verified. The accuracy of Tibetan speech synthesis of the method, the specific comparison result is shown in Fig. 6:

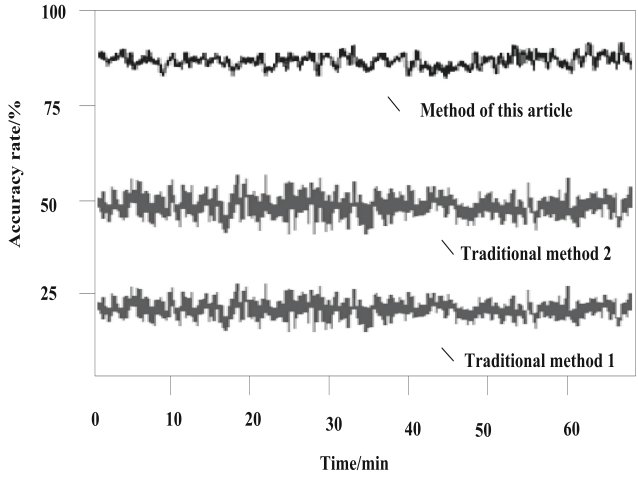

Fig. 6. Tibetan speech synthesis comparison results under different methods

According to the comparison results in the figure, the traditional method 1 extracts the features of the voice signal of the Tibetan speech synthesis system, and the accuracy of Tibetan speech synthesis is within 30%, and the traditional method 2 extracts the features of the voice signal of the Tibetan speech synthesis system. The accuracy is less than 60%. The Tibetan speech synthesis accuracy after extracting the speech signal features of the Tibetan speech synthesis system with the method in this paper is less than 90%, indicating that the Tibetan speech synthesized by the method in this paper is more accurate. Therefore, the method for extracting features of the speech signal of the Tibetan speech synthesis system based on machine learning, on the basis of ensuring the quality of Tibetan speech synthesis, better completes the content of Tibetan speech synthesis, which has more important practical significance.

4 Conclusion and Outlook

In traditional Tibetan speech synthesis, the prosodic control of Tibetan speech synthesis is not considered, so the control effect of the range of Tibetan speech synthesized according to timbre is not obvious. With the extension of Tibetan sentences and the narrowing of the range, the duration of vocal cord vibration is limited. Especially when the speech is close to the end of the sentence, the acoustic characteristics of natural speech are obviously affected by its physiological factors. However, the feature extraction of Tibetan speech synthesis system based on machine learning can improve this point, make Tibetan speech synthesis more flexible, and greatly meet the changing requirements of Tibetan speech rhythm. However, the method in this paper does not consider the calculation time, which leads to a longer time for the speech signal feature extraction of the Tibetan speech synthesis system. Therefore, the following research will focus on the calculation, aiming to improve the speech signal feature extraction efficiency.

Fund Projects. Funded by Graduate Research Innovation Project of Northwest Minzu University (Granted No. YXM2019010).

References

1. Müezzinoğlu, T., Karaköse, M.: An intelligent human–unmanned aerial vehicle interaction approach in real time based on machine learning using wearable gloves. Sensors **21**(5), 1766 (2021)
2. Yao, Y., Ding, J., Wang, S.: Soil salinization monitoring in the Werigan-Kuqa Oasis, China, based on a three-dimensional feature space model with machine learning algorithm. Remote Sens. Lett. **12**(3), 269–277 (2021)
3. Subramani, P., Srinivas, K., Kavitha Rani, B., Sujatha, R., Parameshachari, B.D.: Prediction of muscular paralysis disease based on hybrid feature extraction with machine learning technique for COVID-19 and post-COVID-19 patients. Pers. Ubiquitous Comput. (2021, prepublish)
4. Yilin, F., Yanming, H., Feng, J.: Research on MFCC speech signal feature extraction algorithm based on EMD. Electro. World **008**, 23–25 (2019)
5. Su, H.: Design of the online platform of intelligent library based on machine learning and image recognition. Microprocess. Microsyst. **82**, 103851 (2021)
6. Mei, Y., Ye, D.-P., Jiang, S.-Z., Liu, J.-R.: A particular character speech synthesis system based on deep learning. IETE Tech. Rev. **38**(1), 184–194 (2021)
7. Lei, Y., Zhang, B., Li, R.: Detection of unusual targets in traffic images based on one-class extreme machine learning. Traitement du Signal **37**(6), 1003–1008 (2020)
8. Gajurel, A., Chittoori, B., Mukherjee, P.S., Sadegh, M.: Machine learning methods to map stabilizer effectiveness based on common soil properties. Transp. Geotech. **27**, 100506 (2021)
9. Silvello, G.C., Bortoletto, A.M., Costa, M., de Castro, A., Alcarde, R.: New approach for barrel-aged distillates classification based on maturation level and machine learning: a study of cachaça. LWT **140**, 110836 (2021)
10. Yao, Y., Yu, L., Chen, Y.: Feature Extraction Method of Radiation Source in Deep Learning Based on Square Integral Bispectrum. J. Phys. Conf. Ser. **1678**(1), 012074 (2020)
11. Kumar, S., Singh, S., Agarwal, P., Acharya, U.K., Sethy, P.K., Pandey, C.: Speech quality evaluation for different pitch detection algorithms in LPC speech analysis–synthesis system. Int. J. Speech Technol. **24**(3), 545–551 (2020)
12. Xinyi, Y., Boyu, S., Qingyun, M., Kailin, H.: Design of the speech tone disorders intervention system based on speech synthesis. J. Phys. Conf. Ser. **1617**(1), 012078 (2020)
13. Zhu, X., Xue, L.: Building a controllable expressive speech synthesis system with multiple emotion strengths. Cogn. Syst. Res. **59**, 151–159 (2020)
14. Yue, W.: Research on feature point extraction and matching machine learning method based on light field imaging. Neural Comput. Appl. **31**(12), 8157–8169 (2019)
15. Li, G., Li, J., Zhaojie, J., Sun, Y., Kong, J.: A novel feature extraction method for machine learning based on surface electromyography from healthy brain. Neural Comput. Appl. **31**(12), 9013–9022 (2019)
16. Liu, S., Bai, W., Liu, G., et al.: Parallel fractal compression method for big video data. Complexity **2018**, 2016976 (2018)
17. Liu, S., Fu, W., He, L., Zhou, J., Ma, M.: Distribution of primary additional errors in fractal encoding method. Multimedia Tools Appl. **76**(4), 5787–5802 (2014)