Research on Speech Feature Extraction System in Oral English

Jinhuan Wang,Yan Li, Zehui Xue, Chunxiang Li
Intelligent Science & Information Engineering College, Xi'an Peihua University
Xi'an, China

***Abstract*—This paper analyzes and studies the problems existing in oral practice, proposes a study of speech feature extraction system in oral practice, and uses MALTAB to analyze cepstrum on the extraction of speech signal, to mainly research the same type of feature parameters for reference speech and follow-up speech. Make reference voice and follow-up voice in the same framework of performance according to voice feature, dynamically present in the form of waveforms and graphs, showing differences at multiple angles and levels, and encouraging oral practitioners to continue to improve pronunciation, narrow differences for corpus marking.**

Keywords—corpus marking; LPC algorithm; MATLAB

I. INTRODUCTION

Speaking ability is an important skill of any language learner and one of the important skills involved in communicative competence. Many language learners agree to put acquiring oral English as the primary goal of learning. A speech feature automatic extraction system extracts the same type of feature parameters for reference speech and follow-up speech for perceptual comparison. The initial idea is to use deep learning technology to extract the characteristics of the speech signal by layers, and combine with the basic parameters such as time-frequency of the speech signal to use. Since the follow-up speech and the reference speech, as well as the accompanying speech, are difficult to achieve the same duration, it is necessary to take some measures to process the speech signal, to present the comparison between the pronunciation and the reference speech. According to the voice feature to make the reference voice and the follow-up voice in the same framework of performance, dynamically present in the form of waveforms and graphs, showing differences at multiple angles and levels, and encouraging oral practitioners to continue to improve pronunciation, to narrow differences, for corpus marking. Speaking training requires a large amount of reference speech, including not only the original speech signal, but also some new attributes required for the research of the subject, such as the synchronous matching of pronunciation and words in time.

II. DESIGN OF SPEECH FEATURE EXTRACTION SYSTEM

Two typical ways of speech signal processing are waveform display and parameter display. The waveform display can intuitively make people recognize the voice signal; the parameter display provides in-depth analysis of the speech signal and matches the features of the speech recognition system and speaker recognition. Spoken recognition technology is essentially a process of pattern recognition. The pattern of unknown speech is compared with the reference mode of known speech one by one. The best matching reference pattern is used as the recognition result.

A. LPC algorithm

LPC is an abbreviation for linear predictive coding (LPC), which is a commonly used and important coding method. In principle, the parameter that generates the voice channel excitation and transfer function by analyzing the voice waveform is LPC. In fact, the encoding of the sound waveform is converted into the encoding of these parameters, so that the amount of data of the voice is greatly reduced. The parameters obtained at the receiving end are analyzed by using LPC and the speech is reconstructed by a speech synthesizer. In fact, the synthesizer is a discrete time-varying linear filter that changes over time, and represents generation system model of the human voice. Time-varying linear filters can be used not only as predictors, but also can be used as synthesizers. But when analyzing voice waveforms, we mainly use it as a predictor. When synthesizing speech, we use it as a speech generation model. The parameters and excitation conditions of the model can be periodically adapted to new needs as the changes of voice waveform.

B. MFCC algorithm

MFCC (Mel Cepstrum Coefficient) vividly simulate human auditory characteristics, which is a kind of speech characteristic parameter that conforms to human auditory characteristics. After preprocessing the speech signal, we will extract the characteristic parameters of the speech signal. In general, we divide the characteristic parameters of the speech signal into two types: The first type is the parameter of time domain feature. Usually, each time-domain sample in a frame of speech signal will directly constitute a parameter vector. The second type is the parameter of variation domain characteristic. Currently, the most commonly used characteristic parameter is the frequency domain.

C. MATLABGUI technology

To enable the system to achieve real-time display of waveforms and typical parameter characteristics of speech signals, design a GUI interface:

1) Specific steps and analysis of signal feature extraction

a）MALTAB's extraction for speech signals

When we enter the system interface, we first need to retrieve a piece of audio. So, before getting the audio, we need to initialize the system interface, and click the Get Audio button in the menu bar, to retrieve an audio file with a file size smaller than 8K, the format of the audio file retrieved must be *wav format. We can use the wavread function to sample the speech signal, and then directly set the sampling frequency and sampling points

b）Cepstrum analysis

The cepstrum parameter is an important speech feature parameter, which is the product of homomorphic processing speech. The hemimorphic processing is also called homomorphic filtering, which realizes the separation process of transforming the convolution relationship into the summation relationship, that is, deconvolution. Deconvluting the speech signal can separate the glottal excitation information with the channel response information of the speech signal, thereby obtaining the channel resonance characteristic and the pitch period, for speech coding, synthesis and recognition.

Based on MATLABGUI technology, the interface design and algorithm design of the speech signal feature extraction system are relatively intuitive. Through the system interface, users can directly upload and audit the audio in the voice library.

2) Fluency comparison of simultaneous interpretation language output

This experiment draws on Kormos's (2006: 163) measurement standard of fluency of language output, which is mainly used to quantify the following three factors:

(1) Speed of speech: the average number of words produced per minute;

(2) The number of filled words such as "uhm/er/mm" appearing every minute;

(3) The number of language corrections (ie, repetitions or modifications) that occur on average per minute.

When analyzing the data, firstly transfer the simultaneous interpretation recording of the experimental object, and calculate the time used by each participant to translate the same.

III. FOLLOW THE RULES BELOW FOR TRANSCODING AND STATISTICS

Whether it is calculating the time or the number of words, it does not include translating the title of the article; the digitals are converted into English word, with no hyphens added in the middle; whether to use abbreviations when transcoding, depending on the recording.

A. Speed of speech

During the analysis, calculating the total number of words used in the simultaneous interpretation of the experimental object, and divided by the time used, thus the speech speed of simultaneous interpretation of the experimental object is obtained, which is:

Speech speed of Simultaneous interpretation = total number of simultaneous interpretation words / time of simultaneous interpretation

When counting the total number of simultaneous interpretation words, we do not consider the translation of the title, just starting from the text translation. The filling words that appear during translation and the words that are repeated or corrected are also counted. The time of simultaneous interpretation is calculated from the text translation. When the second is converted to minutes, the time with remainder is reserved to three decimal places, following the principle of rounding. The final calculated speech speed of simultaneous interpretation is in units of "words/minutes", and the time with remainder is reserved to two decimal places, following the principle of rounding.

Calculate simultaneous interpretation's speech speed of each subject respectively, and compare the average value of the simultaneous interpretation's speech speed of the two groups. The data comparison is as follows:

TABLE I COMPARISON TABLE OF SIMULTANEOUS INTERPRETATION'S SPEECH SPEED (UNIT: WORD / MINUTE)

Sequence number	Follow the group	Listening group
1	2.07	2.07
2	0.32	2.53
3	2.55	0.79
4	2.55	2.70
5	1.27	3.35
6	3.36	0.95
7	1.76	2.38
8	0.48	1.58
9	2.53	1.11
Average value	1.907	1.940

B. Filling words

When transliterating text, mark E where the filling word appears in the recording. First, calculate the number of times that the filling words appear in the simultaneous interpretation of the experiment object, and divide by the time (minutes) used in the simultaneous interpretation, to get the number of times of filling words that appear every minute when the experimental object is interpreting simultaneously, that is, the frequency of the filling word when the object is interpreting simultaneously. The formula is as follows:

Frequency of filling word = number of filling words / time of simultaneous interpretation (1)

The final calculated frequency of filling words of each subject is in units of "times/minutes", and the remainder is retained to two decimal places, following the principle of rounding. On the basis of calculating the frequency of the filling words of each subject, the average value of the frequency of the filling words in the two groups is further compared. In order to compare the frequency of the filling words of the two groups in more detail, the average frequency of each group is retained to three decimal places, following the principle of rounding. The data comparison is as follows:

TABLE II COMPARISON TABLE OF FILLING WORD FREQUENCY (UNIT: TIMES / MINUTE)

Sequence number	Follow the group	Listening group
1	1.91	0.32
2	0.16	0.16
3	1.28	0.16
4	2.07	2.22
5	0.16	0.32
6	3.15	0
7	0	0
8	0	0.63
9	0.16	1.91
Average value	0.988	0.636

C. Language correction

When translating text, a language correction appears in the recording, that is, where the repetition or modification is marked R. First, calculate the number of language corrections that occur when the experimental subject is interpreting simultaneously, and then divide by the time (in minutes) used in the simultaneous interpretation, which can obtain the number of language corrections that occur every minute when the subject is interpreting simultaneously, that is, the frequency of correction when the object is interpreting simultaneously. The formula is as follows:

Correction frequency = number of corrections / simultaneous interpretation time (2)

Calculate the correction frequency of each subject separately, and then compare the average value of the two groups of correction frequencies, and the final correction frequency of each subject is measured in units of "times/minutes", and the remainder is retained to two decimal places, following the principle of rounding. In order to compare the correction frequencies of the two groups in more detail, the average frequency of each group is retained to three decimal places, following the principle of rounding. The data comparison is as follows:

TABLE III COMPARISON TABLE OF CORRECTION FREQUENCY (UNIT: TIMES / MINUTE)

Sequence number	Follow the group	Listening group
1	100.91	116.64
2	100.16	83.06
3	99.41	98.46
4	94.78	89.68
5	94.82	102.76
6	96.01	93.43
7	95.78	916.51
8	75.59	104.32
9	97.99	78.31
Average value	95.05	95.91

D. Corpus marking

During the experimental analysis, the recording of the subject was transferred. By comparing the language expressions marked by the participants with the language expressions used in the actual simultaneous interpretation, the frequency of use of the noted expressions in simultaneous interpretation is obtained. The frequency expression used is as follows:

Frequency of use = number of expressions used in simultaneous interpretation / number of expressions maked.

Calculate the frequency of use of each participant separately. When calculating the frequency, follow these principles: Regardless of the grammatical accuracy, that is, if the subject uses the expression noted or listened to, even if the grammar is wrong, it is considered to use the expression; Although the simultaneous interpretation is not completely in accordance with the original expression, use variant of the vocabulary or phrase, which is still considered to use the expression; When a part of words in a phrase is used, it is considered to use the expression.

IV. CONCLUSION

This paper mainly analyzes the specific steps and analysis of signal feature extraction, the fluency comparison of simultaneous interpretation language output, speech speed, filling words, language correction, and corpus marking of the LPC algorithm, MFCC algorithm, MATLABGUI technology. The research of the curriculum design system provides reference for the development of speech recognition in the future, and also provides valuable reference for English learners.

ACKNOWLEDGMENT

This work was supported by the thirteenth Five-Year Plan" for Education Science in Shaanxi Province in 2018. No. SGH18H509 and Special research project of Shanxi Provincial Department of Education in 2018. No.18JK1084.

REFERENCES

[1] Sun Shanghong, Bai Zhen. Spectrum Analysis of Speech Signal in MATLAB [J].Journal of Hetao College,2016,13(01):72-75.

[2] Wang Guangyan, Zhao Xiaoqun, Wang Xia. Design of speech signal feature extraction system based on MATLABGUI [J].Journal of Hebei University of Technology,2010,39(04):14-18.

[3] Li Jing. Design of Speech Signal Acquisition and Processing System Based on MATLAB [J].Journal of Shanxi Datong University (Natural Science Edition), 2016,32(02):30-33.

[4] Li Jing. Design of Speech Signal Acquisition and Processing System Based on MATLAB [J].Journal of Shanxi Datong University (Natural Science Edition), 2016,32(02):30-33.

[5] Hu Jiarong, Liao Baisen. 2009. Discussion on the Current Situation of Chinese-English Interpreting Teaching in Taiwan Colleges and Universities [A]. Compilation theory (Volume 2, Issue No. 1) [C], 151-178. "National Compilation Hall".