ATLANTIS PRESS

Advances in Engineering Research, volume 122

7th International Conference on Applied Science, Engineering and Technology (ICASET 2017)

Research and Realization of Intelligent English Phonetic Pronunciation Training System

Hong Zheng

Foreign Language Learning Center, QiLu University of Industry Jinan, China. 250353

zhh@qlu.edu.cn, 267353557@qq.com

Abstract: This study, based on advanced mathematical algorithms, is focused on computer technology and intelligent speech recognition technology. It provides a useful tool for English phonetics teaching by realizing intelligent English phonetic pronunciation training system on the computer platform, which can greatly reduce the workload of oral English teachers to correct students' phonetic pronunciation and improves teaching efficiency and quality.

Keywords: Speech signal, Intelligent speech recognition, Pronunciation

Introduction

Correct standard pronunciation is the foundation of learning English, however, for non-native speakers of English learners, effectively correcting mistakes in English pronunciation, especially correcting that of phonetic pronunciation, is one of the difficulties in English learning, thus also a key issue in oral English teaching.

With the rapid development of computer technology, the role of computer technology is more and more important in English teaching. It is of great significance practically and socially to use computer technology and speech recognition technology to perform assisted English pronunciation learning.

The system overview

Pronunciation detection methods

With the development of speech-recognition technology, many research institutions has performed in-depth research on speech recognition technology and developed some advanced speech-recognition system, including FLUENCY developed by the language technologies Institute at Carnegie Mellon University, EduSpeak by SRI in United States. In China, studies are also carried on, including that of department of electronic engineering of Tsinghua University, Harbin Institute of technology computer science Department by. Most of the studies is Chinese language-oriented.

Speech recognition methods can be divided into three types. The first method, based on vocal tract model , was started earlier, but has not reached a practical stage due to complex acoustic and phonetic knowledge . The second method, is a pattern matching method, which is relatively mature, and has reached a practical stage and common technologies involved include dynamic time warping (DTW), hidden Markov (HMM), vector quantization (VQ) and so on . The third method, one of artificial neural network, is still at the experimental stage since its implementation is rather complicated.

The system as proposed in this paper is based on a more sophisticated pattern matching method for detection of speech recognition. The known parameters of the Speech feature vector are saved

into the template gallery for further extraction and comparison, i.e., inputting parameters of speech feature vector to be measured with the template parameters for similarity comparison to obtain recognition results.

System process

The first step is voice input. Phonetic pronunciation sounds are collected and pre-processed to meet the data stream requirements of speech recognition system. The feature data are extracted from the speech signal data stream, which are matched with the standard reference model. Results matched are then evaluated for the final output.

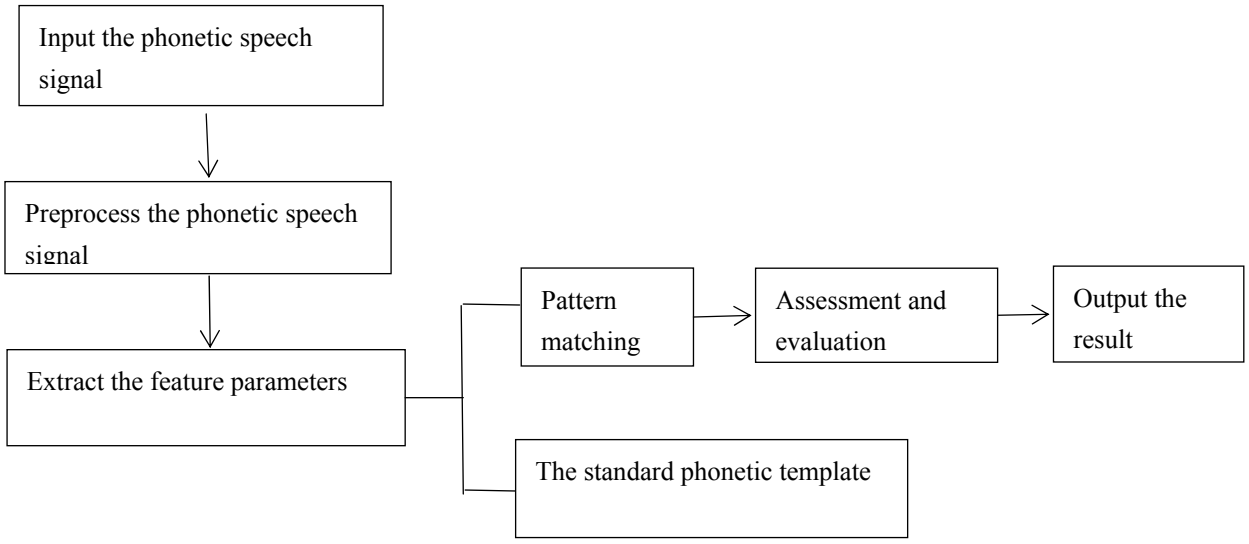

Figure 1. frame of intelligent recognition system of phonetic pronunciation

Realization of Intelligent English phonetic pronunciation identification system

Intelligent phonetic pronunciation identification system provides learners with animations, pictures, sounds, texts and other multimedia materials. It is capable of effective analysis and feedback in phonetic pronunciation learning, guiding learners to improve phonetic pronunciation accuracy, therefore enhancing the level of spoken English.

According to the need of functionality and speech recognition process, the core modules of the system are confirmed as follows:

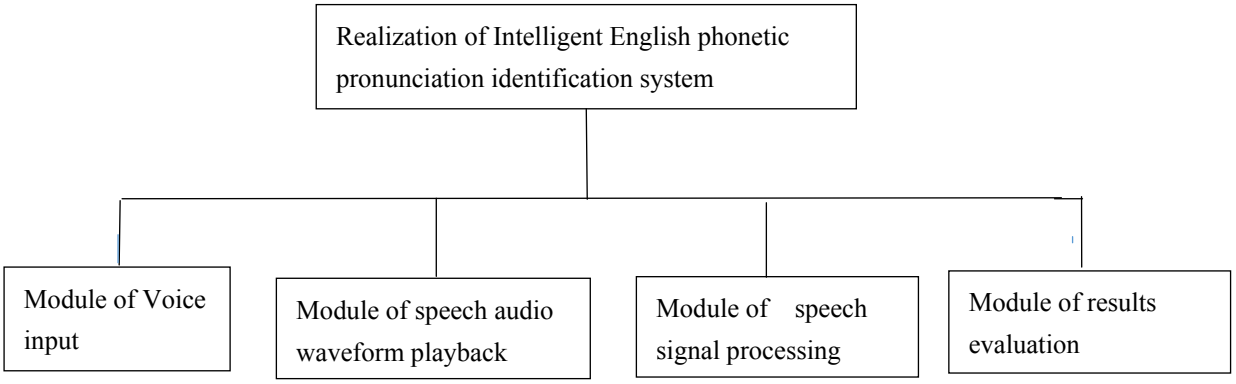

Figure 2. Block diagram of phonetic pronunciation intelligent recognition system

Voice input and waveform playback

English phonetic pronunciation of speech signal input is the basis of the system. Humans can hear sounds with a frequency ranging 20 Hz~20000 Hz, while voice frequencies can be up to 3400 Hz. Speech signal collecting refers to the process during which speech signal is generated as analog electrical signal by microphone and signal conditioning circuits, and further converted into acceptable digital signals by AD DAC chip.

As prescribed by the "Nyquist sampling theorem" , the sampling frequency must be two times greater than the highest frequency of the analog signal. Since voice signal's frequency is ranging 300~3400 Hz, the sampling frequency is set at 10 kHz audio, which is suitable for most AD DAC chip. The AD DAC chip in the research is 24-bit Δ－Σ AD converters CS5361, which is introduced by CRYSTAL company, with sampling frequency of 114dB, 192kHz. Main features of CS5361 are as follows:

multi-bit δ-σ-structures;

With 24-bit accuracy;

114dB dynamic range;

Total harmonic distortion (THD) plus noise better than -105dB;

Sampling rate up to 192kHz;

Power consumption less than 150mW;

Internal high-pass filter circuits or internal calibration offset voltage DC;

In-band linear phase digital anti alias filter;

Supporting 5V to 2.5V logic levels;

Using differential input structure;

overflow detection capabilities;

CS5361 is completed a/d converter for digital audio system, capable of sampling, ADC, antialiasing filters, and other functions. The resulting output is in serial mode, corresponding to 24-bit sample data of two input channels, and the highest data output rate can be as high as 192kHz.

CS5361 uses differential input structure with excellent noise suppression and 5 scale multi-bit Δ－Σ modulator, combined with a digital filter and sampler, which avoids the trouble for an external anti alias filter .

Judging from the above technical parameters, CS5361 is able to fulfill the needs of speech signal acquisition. Choosing a suitable AD converter effectively is the key to a practical speech recognition system.

speech signal of Phonetic pronunciation, one inputted into a computer system,is stored in the system memory as an array for easy speech waveform playback as well as the subsequent signal processing. Speech waveform is played in the system.

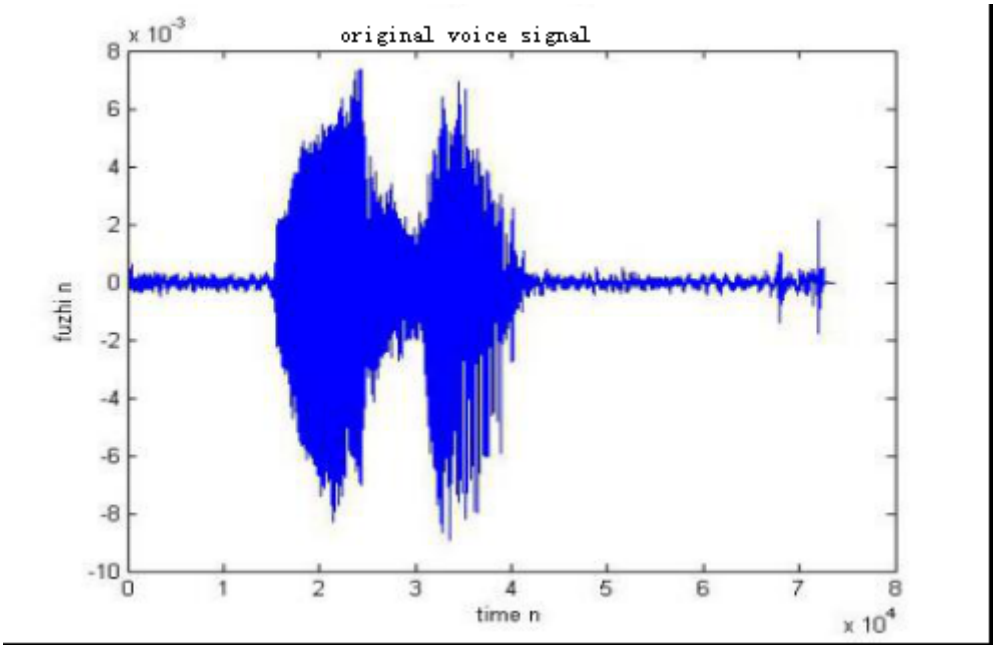

Figure 3 Sampled waveforms of voice signal

Speech signal processing

After the acquisition of speech signals and conversion of them into digital signals, characteristic parameters needs to be extracted. Data compression of original speech signal collected is needed to remove the irrelevant information and preserve key relevant information for speech recognition.

The higher the speech frequency is, the lower the amplitude is. Therefore, it is of necessity to carry a series of processing including high-frequency pre-emphasis, framing, window treatment before extracting characteristic parameters. Speech signal processed has smoothing waveform and is ready for extraction of characteristic values. Characteristic parameters commonly used include linear prediction coefficients (LPCC), Mel Cepstral coefficients (MFCC), and accent-sensitive parameters (ASCC) algorithms. These three methods have their advantages and disadvantages, of which linear prediction coefficients (LPCC), characterized with small amount of calculation and ease for realization, is used in this system.

The system uses linear predictive Cepstral coefficients (LPCC) to analyze the speech frames, resulting in mapping of prediction coefficient in cepstrum domain. Since extraction algorithms of linear predictive Cepstral parameters is carried out in the time domain, so the calculating strength is low. For spectrum of vowels in the speech, LPCC algorithm has a good ability to extract, which can guarantee the integrity of the main character of the voice, ensuring that the system has high recognition rate.

Voice signal matching process

Speech signal is different from other regular signals for its remarkable randomness; even for the same person with the same tone, length of his every utterance will not be completely equal. if the match is only performed at the linear level of time matching, phonemes might not be accurately paralleled. recognition and identification of is crucial for the efficiency of phonetic speech recognition algorithms. Commonly used algorithm are the method of dynamic time warping (DTW), the method of vector quantization (vQ) and so on. This system uses the method of dynamic time warping (DTW).

Method of dynamic time warping (DTW) was proposed by Japanese Itakura in the 1960 of the 20th century. the basic idea of the algorithm is to unevenly twist or bend the unknown quantities and correct their characteristics and the characteristics of the reference model. Dynamic time warping has made a breakthrough into the problems of inconsistency of speech characteristic parameter and has been a great success.in isolated word recognition system. the DTW algorithm is more applicable to Phonetic phonetic pronunciation.

DTW is an optimization algorithm which describes the time correspondence between test and reference templates with w(n) time warping function that meets certain conditions, thereby solving matching distance between the two templates, which means finding time warping functions m = w (n), making an non-linear mapping of the time-line of a test template onto that of the reference templates , thus minimizing the total matching distance between test and the reference templates.

Dynamic time warping (DTW) technology is essentially an algorithm for dynamic optimization. This algorithm irregularly distorts signal in the time domain, align the characteristic parameters with the template gallery. In the process of alignment, it continuously calculates the distances between feature vectors in order to obtain a best matching path minimizing the total distance. it can minimizes the time difference distortion and maximizes the acoustic similar characteristics, thus becoming one of the most commonly used algorithm for speech recognition.

During the process of pattern matching, the system uses loose-end alignment for the endpoint sensitivity and larger amount of computation, improves the sensitivity of its endpoints and effectively reduces computational complexity, thus performing well in phonetic speech recognition.

System assessment on the speech signal

Evaluating judgment on sound is the key function and core part of of Intelligent English Phonetic Pronunciation Training System. learners' pronunciation is scored to achieve a quantitative evaluation; accurate pronunciation scoring enable learners to have a clear understanding of their accomplishments and improve pronunciation continuously.

This system measures the pronunciation level according to the standard pronunciation as reference templates. pronunciation scoring algorithm is expected to have a high degree of reliability and accuracy, and can accurately evaluate the learners ' pronunciation, moreover, it need to meet the requirements of real-time calculation and system functions. No additional training is needed since the amount of computation of standard Phonetics reference template is not large. The method possesses high reliability as for the syllable and small vocabulary speech sound, so it is applicable to English phonetic pronunciation evaluation.

Signal of the test voice and standard reference templates are pre-processed for character extraction, then the test templates and reference template can be matched to obtain the matching distance Dmin (N,M), which serves as measurement for the pronunciation difference between reference template and test templates. the system can reveals the similarity of language features in a reliable and comprehensive way.

In Pronunciation quality assessment system, the scoring mechanism is focused on mapping relationship between matching distance and pronunciation scoring. It gives a matching score calculation method between distance and the pronunciation. the total distance is Defined as the sum of vector distance between test templates and reference templates, and it is reasonable that the larger the frame length is, the larger the total distance since different speech frame length corresponds to the pronunciation is not the same. The average distance can be obtained by dividing total distance by the frame length. The introduction of the average distance into pronunciation evaluation can eliminates the influence of sound length, and it is a reliable index for pronunciation level.

Conclusions and Outlook

The pronunciation training system collects sound signals by using single-chip microcomputer embedded systems at the front end of PLC and realizes waveform display, repeating, real-time evaluation and correction for the English phonetic pronunciation by using VC++ at the back-end of the upper computer.

during the daily application of this system, we have found many places to improve. further research and improvement can be devoted to some aspects including optimizing algorithms, reducing the amount of calculation, improving operational efficiency and response speed, enhancing optimization of interface, adding more and more useful features, such as the contrast of lip movements and the system's practical value.

With the development of the mobile Internet, speech recognition technology is more and more popularized. Under the guidance of Google, Internet and communications companies take speech recognition as important research field, for example, Google has speech translation, Iphone has Siri speech recognition software. as Baidu, Tencent, Huawei has also joined the field of speech recognition, speech recognition technology in China has been elevated to the world's advanced level. with the increasing penetration of smart phones, Android system as a good smart phone operating system has developed rapidly in recent years, speech recognition systems based on the Android system will have a lot of room to grow. portable English phonetic pronunciation training system based on Android-based phones will be the new orientation of further research.

References

[1] Wang Shuo. The design and implementation of intelligent English pronunciation training system based on Android platform [D], Nanjing University of Posts and Telecommunications, Nanjing, China (2013)

[2] Li Hong-Yan, Huang Shen, Wang Shi-Jin, Liang Jia-En, Xu Bo. Automatic Mispronunciation Detection for English Learners by GMM-UBM and GLDS-SVM Methods, ACTA AUTOMATICA SINICA, Beijing, China. 332-336(2010)

[3] Tu Huiyan, Chen Yining. English Learning System of Oral Phonation Based on ASR and Smart Phone Platform [J], Computer Applications and Software, Shanghai,China.64-66 (2011)

[4] Li Cheng. Research and Implementation of Portable Speaker Verification System Based on SoPC Technology [D], Beijing, China.

[5] Li Jinguo. Design of Digit-Voice Recorder and Playback System[D], Hunan University of Arts and Science, Hunan, China (2010)

[6] Zheng Bo. Research on phoneme level correction algorithm for English pronunciation [D], NanKai University, TianJin, China.(2008)

[7] Zhao Li.Speech signal processing[M]，China Machine Press, Beijing, China.

[8] Tu Bingx,Qu Dan. Practical speech recognition Foundation[M]. National Defence Industry Press, Beijing, China.(2005)

[9] Adriana Garcia Kunzel. An Android Approach to The Web Services Resource Framework.[M] Florida Atlantic University (2010)

[10]Yao KS, Paliwal KK，Nakamura S. Noise adaptive speech recognition based on sequential noise parameter estimation[J], Speech Communication, 2004,42(1) :5-23