Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(6):654-660

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Research on the optimization of voice quality of network English teaching system

Zhu Zhimei

Department of Foreign Languages, Heze University, Shandong, China

ABSTRACT

In order to improve the voice quality of network English teaching system, the paper operated the study on the improvement of the characteristic parameter of MFCC and LSP, which reduced noise, optimized the voice quality and improved the accuracy of voice judgment in some extent. How to improve the quality of voice identification is the key to optimize the voice quality of network English teaching system. The study first analyzed the key factors of improving voice quality from the following three aspects, which are the preprocessing of voice signal, the extraction of parameters of voice characteristics and the measurement of similarity. And then took the parameters of voice characteristics as the entry point and finished the parameter extraction of voice characteristics by combining MFCC and LSP. The experiment shows that such method not only restores the voice reality of speakers effectively, but also reduces the misjudgment rate of voice matching. The above functions of such method make it own some research value.

Keywords：network English teaching, voice identification, characteristic parameters’ extraction, MFCC, LSP

INTRODUCTION

Network English teaching is based on the modern information technology, especially network technology, which makes English learning develops for individual, unlimited in time and space (Dong, 2012). Such new learning mode can arouse the enthusiasm of teachers and students fully, especially can train the self-learning ability of students, which can ensure the dominant role of students in English teaching. At the same time, the teaching method of network English teaching system increases the learning initiative of students more distinctly (Zhao, 2011). From a survey of college students, 88 percent of college students think that the voice optimization of current network English teaching system is necessary (see Figure 1)

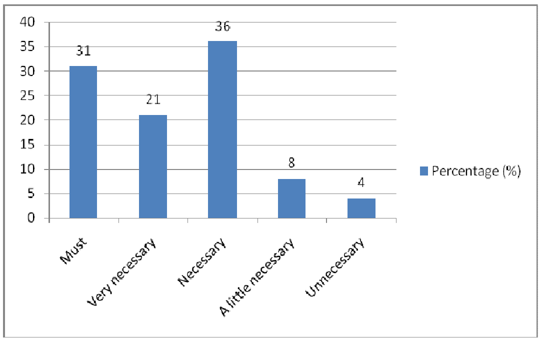

Figure 1. the necessary of voice optimization of current network English teaching system

The abilities of English listening, speaking, reading and writing are the main function of network English teaching system. Listen and repeat, test on oral English and test on English listening all needs the guarantee of high quality voice system, which is traditional English teaching system cannot compared with. The listening and speaking system is the core component of network English teaching system. In other words, voice system takes a very important part in network English teaching system.

Nowadays, network English teaching system has become very popular in current English teaching. More and more students are using the network English teaching system to learn English by themselves. There should be three systems in the network teaching platform of college English, which are teaching/learning system, teaching/learning resources and teaching/learning management system. In the three big systems, the network English teaching system is the major system, which includes five sub-systems, that is multi-media coursework system, real time guidance system, unreal time discussion system, homework submission/management system and online test system. Such system is the most common system used in college English teaching. The network classroom integrates the scattered resource and makes utilization, which can provide platform for self learning and build new English teaching method. The self learning mode of current network English teaching system can be seen in Figure 2.

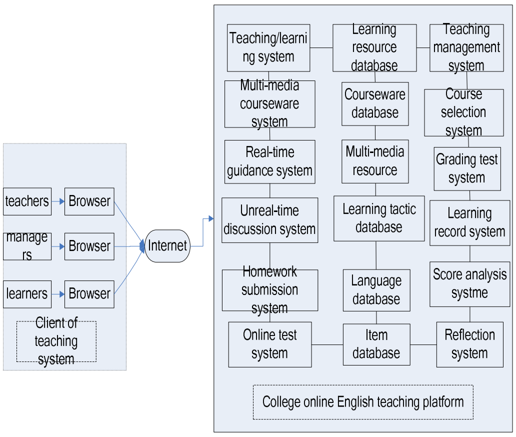

Figure 2 The learning mode of current network English teaching system

2 The voice system of network English teaching

2.1 The structure of the system of voice collection, processing and identification

The voice system of network English teaching mainly divides into two parts, which are accepting system of voice and collection system of voice. The technology of accepting system of voice is relatively mature, which can ensure the high quality of accepting voice. However, the technology of collection system of voice develops relatively slow and the collection, identification, measurement of similarity of voice is all the core components of the voice optimization of network English teaching system.

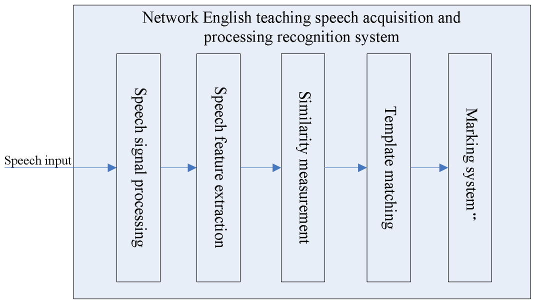
Network English teaching speech acquisition and processing recognition system
Speech input
Speech signal processing
Speech feature extraction
Similarity measurement
Template matching
Marking system

Figure 3 the structure of voice collection, processing and identification system

After the entering through microphone, the voice signal first will be preprocessed, and then the network English teaching system will extract the parameters of voice characteristics, after that, the system will make the similarity measurement of the voice signal and match them with the module of database. Finally, the network English teaching system will judge the level of similarity between entering voice signal and the module of database according to marking system.

The function of listening and reading in network English teaching system is to match the read after voice of learners and the standard voice of database. By doing this, the network English teaching system can judge the voice of learners. If the accuracy of the voice system is not high enough, it will influence the quality of listen and read in a big extent even it may cause misjudgment (Kang, 2011). Oral English test system has a very rigorous demand to voice quality and marking system, which puts forward a higher demand for the optimization of voice quality. Take English pronunciation for instance, it is very close between vowel and some non-vowel alphabetic, which demands voice system to have an accurate identification and judgment.

2.2 voice identification

Voice identification has become an important research instrument of artificial intelligence and mode identification. Voice identification system mainly includes voice entering, extraction of characteristics parameters, model of acoustical standard, database dictionaries, voice models of grammar and identification programs (Wang and Liu, 2012). In addition to above factors, the environmental factors of voice entering also should be considered. Voice identification system must own the technology to resist environmental disorders in order to eliminate environmental noises. At the same time, voice identification system should be supported by the input and output interface technology of voice. Thus, the technology of voice identification must interact with various kinds of external technologies. Only by this, voice identification system can realize its function smoothly.

Voice identification is one type of module identification, which mainly includes the following three modules, preprocessing of voice signal, and extraction of voice characteristics and measurement of similarity. According to the different applications of practice, voice identification system can be divided into special person identification and non-special person identification, independent and continuous words identification, small quantity, big quantity and unlimited quantity of words identification. Among those, network English teaching system mainly considers the identification of non-special person, continuous words and unlimited words.

2.2.1 Voice preprocessing

Before analyzing and processing voice signals, voice identification system must make preprocessing for them, which mainly includes digitization, anti-aliasing filter, pre-emphasis, framing and windowing, and endpoint detection (Song et al., 2012).

(1) Pre-emphasis

The main function of pre-emphasis is to compensate and emphasis the frequency spectrum with low ingredients. This is because that it is very difficult to speculate frequency spectrum with low ingredients. After compensating and emphasizing, it is easier for seeking them. The characteristic of voice signals is that there are fewer ingredients with high frequency and most of them are belonged to low frequency ingredients. It is very difficult to speculate the frequency spectrum as most of them are belonged to low frequency ingredients. Thus, it is necessary to emphasis the ingredients with low frequency in order to speculate them and make analysis of frequency spectrum and track parameters. In this study, the pre-emphasis of voice signals is finished by improving the digital filter of high frequency through 6dB frequency.

(2) Windowing

Voice signal has the characteristic of short time stationary. From long term, voice signal is no stationary, but in short time, it is stationary. Considering such characteristic of voice signal, the whole non-stationary process of voice signal can be divided into several short time stationary processes. In the short time stationary processes, the characteristics parameters of can be analyzed. Such short time processes are called frame. The above process is called framing of voice signal. The realization of framing is mainly through adding window function and the frame size always take between 10 and 30ms. The process of framing can divide time axis continuously. However, the common method of framing is to make overlapping periods processing through sliding window. The advantage of such method is that it kept the smooth transition among different frames. Nowadays, there are three kinds of windows that are being used frequently, which are rectangular window, Hamming window and Hanning window. The definitions are as follows (N indicates the length of windows),

Rectangular window

$$\omega(n)=\begin{cases}1, 0\le n\le N-1\\ \quad 0, n=else\end{cases} \tag{1}$$

Hamming window

$$\omega(n)=\begin{cases}0.54-0.46\cos[2\pi n/(N-1)],0\le n\le N-1\\ 0,n=else\end{cases} \qquad (2)$$

Hanning window

$$\omega(n)=\begin{cases}0.5[1-\cos(2\pi n/(N-1))],0\le n\le N-1\\ 0,n=else\end{cases} \qquad (3)$$

2.2.2 The extraction of voice characteristics parameters

The extraction of voice characteristics parameter is a key procedure of voice identification process. The good or bad of the parameters extracted directly influence the performance of the identification system. After pre-processing, it can make extraction and analysis of characteristics parameters for voice signal. The extractive principle is to make sure that the internal distance is as small as possible and the between class distance is as big as possible. There are many parameters that describe voice characteristics, such as average energy, zero-crossing rate, frequency spectrum, resonance peak, cepstrum, linear prediction coefficients, PARCOR coefficients, track characteristics, voice length, pitch and tone. Voice identification system can choose part parameters of voice characteristics to extract and optimize according to practical demand and height of accuracy.

2.2.3 The measurement of similarity

After the extraction of voice characteristics parameters, the network English teaching system needs to compare the parameters of characteristics and formwork of voice in order to judge the similarity of them. However, in reality it can not compare the parameters of voice characteristics and formwork directly. This is because voice signal has very big randomness and in different time even the same person speak the same word with the same pronunciation, it can not get the same length with before. Thus, in the comparison of formworks, it must consider the problem of time flexible process and reduce the influence of the change of time length for measurement as much as possible in order to improve the rate of identification.

The commonest method is to consult the length of the reference template and make the elongation and shorten processing of the collected voice signal in order to keep the same with reference template to the greatest extent. However, such method has a disadvantage, which is difficult to justify correctly of the collected voice signal and reference template and this will further cause the low efficiency of identification. In order to solve the problem, this paper applied the dynamic time corrected method of nonlinear etiquette technology to finish the elongation and shorten processing of time. The principle can be seen in figure 2.

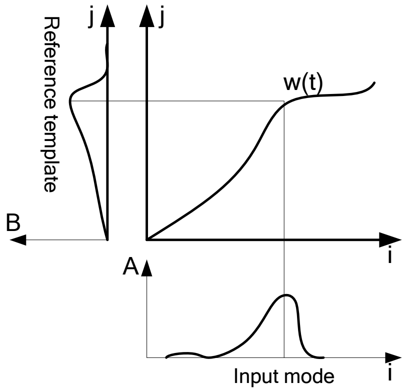

Figure 4 DTW principle

Suppose the parameter of voice test has I frame vector and the reference model has J frame vector, I ≠ J, order time neat function $j=\omega(i)$, its function is to map the time axis i of voice vector which was being tested to the time axis j non-linearly. The distance of D is as follows,

$$D=\min_{\omega(i)}\sum_{i=1}^{I}d[T(i),R(\omega(i))] \qquad (4)$$

In the above formula, $d[T(i),R(\omega(i))]$ indicates the distance measurement between the T(i) and R(j). T(i) indicates the measurement vector of i frame. R(j) indicates the j vector of the model. D indicates the distance between the voice vector being tested and the vector of reference template under the condition of optimization time.

DTW is realized by applying the technology of Data Processing. Data Processing is a kind of optimization algorithm. Its principle can be seen in figure 3.

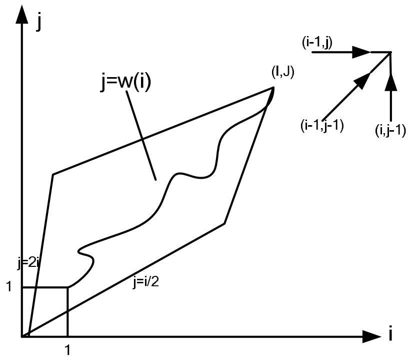

Figure 5 Diagram of DTW Path

The distance speculation of DTW is applied the process of inverse division, which gets the optimization path from the initial state (I, J) to final state (1, 1). Each state (I, J) has relevance with its adjacent states (I-1, J-1), (I-1, J) and (I, J-1). It makes time wrapping from adjacent states in order to get optimization path.

3 Optimization of voice quality

3.1 combination of MFCC and difference parameter

There is some advantage of MFCC (Mel Frequency Cepstral Coefficients) in reflecting the characteristics parameter of hearing mechanism of human ears (Wang, 2012; He and Pan, 2011). The detail speculation process mainly divides into the following four procedures.

(1) Determine the points of the voice order of each frame. The experimental frame length and frame movement applies 256 point and 120 point separately. And then make change of frequency spectrum to all frame signals in order to get the short time power spectrum $P_n(k)$. The solving process is mainly based on the relation between power spectrum and Fourier transform in order to get power spectrum indirectly.

$$X_n(k)=\sum_{m=0}^{N-1}x_n(m)e^{-j2\pi km},0\le k\le N-1 \qquad (5)$$

$$P_n(k)=X_n(k)\cdot X_n^*(k)=|X_n(k)|^2 \qquad (6)$$

To formula (2), its magnitude of power becomes the following after entering M filter,

$$P_m=\sum_{k=0}^{K-1}P_n(k)H_m(k) \qquad (7)$$

Then formula (3) becomes the following part,

$$c(i)=\sqrt{\frac{2}{N}}\sum_{m=0}^{M-1}\lg P_m\cos[(m+0.5)\frac{i\pi}{M}],i=1,2,\cdots,M-1 \qquad (8)$$

Then after speculating the parameter of MFCC and weighting, formula (4) becomes the following part,

$$\omega_i=1+\frac{M}{2}\sin(\frac{\pi i}{M}),1\le i\le M \qquad (9)$$

Through experiment, it shows that the improved MFCC through cepstrum solves the problem of voice dynamic identification commendably, which is original MFCC can not solve. Original MFCC showed good feature in processing statistic characteristic of voice signal. However, in practical noisy environment of voice, original MFCC becomes outshone comparing with the improved MFCC with cepstrum. The application of improved MFCC in voice system will improve the performance of voice identification to great extent. Commonly, it applies the method of combination of improved MFCC with ceptrum and difference parameter to train. The main computational formulas are as follows,

$$d_n = \frac{1}{\sqrt{\sum_{j=-k}^{k} j^2}} \sum_{j=-k}^{k} j \cdot c_{n+j} \qquad （10）$$

Inside, d_n indicates the difference cepstrum parameter of the voice signal of no n. K is a constant, and c_n indicates the cepstrum parameter of the voice signal of no n.

3.2 The comprehensive judgment of MFCC and LSP

Classical studies show that voice identification is a system all linked with one another. Each computation result of one cyclic will directly influence the identification quality of next cyclic, which will further influence the final judgment result. The most important procedure in processing is the extraction of voice parameter. The difference of characteristics parameter extracted will produce direct impact on the precision of judgment. The most popular characteristics parameter being used is MFCC. There are many experiments show that MFCC can express the characteristics of listening mechanism of human ears comparing with other kinds of parameters. The other improvement method in this paper is that it focuses on the non-linear characteristics of MFCC and combines another important parameter of voice signal which is Linear Prediction. It proposed a method of mixed use of voice characteristics parameter, which improves the accuracy of judgment system of voice quality.

The spectrum features of voice signals are all contained in LPC except for tone period. As a deduction parameter, LSP is defined as the root of formula (11) and (12).

$$P(z) = A_p(z) - z^{-(p+1)} A_p(z^{-1}) \qquad （11）$$

$$Q(z) = A_p(z) + z^{-(p+1)} A_p(z^{-1}) \qquad （12）$$

The frequency distribution of the above two polynomials are as follows, $0 < \omega_1 < \theta_1 < ... < \omega_{p/z} < \theta_{p/z} < \pi$, ω_i and θ_i are the no i zero point of P（z）and Q（z）. Their appearance reflects the characteristic of frequency spectrum of voice signal in some extent. LSP reflects the formant characteristic of magnitude spectra clearly and play some compensate role for characteristics (Xuefeng, Zhang et al., 2011).

This paper proposed the method of combination of MFCC and LSP parameter to make comprehensive judgment. The main basis is that voice signal is a very complex random process and on account of the listening principle of human ears, MFCC stands for the non-linear characteristic of voice signal, LSP parameter stands for the linear characteristic of voice signal. There are both connection and difference between MFCC and LSP. MFCC is often being used to identify the characteristics parameter of models and LSP is often being used to be the judgment basis after identification. In this network English teaching system, there is no special judgment module, whereas combine the parameters of MFCC and LSP to become the characteristics parameter of model identification. Thus, it can not only reduce the quantity of processing modules of the system which can further reduce the complexity of computation, but also can improve the accuracy of the system effectively.

CONCLUSION

As there are many problems in the voice quality of network English teaching system, how to optimize the voice quality of current network English teaching system has become an important issue waiting to be solved. This paper just tends to solve the above problem. The result of the paper can give some guidance for the improvement of the voice quality of current network English teaching system.

REFERENCES

[1] Wang, B., **2012**. *Comput. Digit. Eng.*, 4: 19-21.
[2] Zhao, C., **2011**. *Educ. Occupation*, 11: 151-152.

[3] Wang, S. and W. Liu, **2012**. *Comput. Eng. Appli*., 11: 71-74.
[4] Kang, X., **2011**. Discussion on combination of network english teaching and traditional english teaching. Vol. 18. Educ. Teaching Forum.
[5] 13517op
[6] Xuefeng, Zhang, F. Wang and P. Xia, **2011**. *Compuet. Eng*., 4: 216-217.
[7] Dong, Y., **2012**. *Chinese Newspaper*, 16: 187-188.
[8] Song, Z., L. Ma, S. Liu and Q. Li, **2012**. *Comput. Simulation*, 5: 152-155.
[9] He, Z. and P. Pan, **2011**. *Scientific Technol. Eng*., 18: 4215- 4227.