MATEC Web of Conferences **22**, 01032 (2015)
DOI: 10.1051/matecconf/20152201032

Rating Algorithm for Pronunciation of English Based on Audio Feature Pattern Matching

Kun Li, Jing Li, Yufang Song & Hewei Fu
Qinhuangdao Institute of Technology, Qinhuangdao, Hebei, China

ABSTRACT: With the increasing internationalization of China, language communication has become an important channel for us to adapt to the political and economic environment. How to improve English learners' language learning efficiency in limited conditions has turned into a problem demanding prompt solution at present. This paper applies two pronunciation patterns according to the actual needs of English pronunciation rating: to-be-evaluated pronunciation pattern and standard pronunciation pattern. It will translate the patterns into English pronunciation rating results through European distance. Besides, this paper will introduce the design philosophy of the whole algorithm in combination with CHMM matching pattern. Each link of the CHMM pattern will be given selective analysis while a contrast experiment between the CHMM matching pattern and the other two patterns will be conducted. From the experiment results, it can be concluded that CHMM pattern is the best option.

Keywords: audio feature; pattern matching; pronunciation of English; CHMM

1 INTRODUCTION

English learning generally includes listening, speaking, reading and writing. Each part has its own features. However, we've been lacking of a good way to improve speaking for a long time. Even with places like English corners, not all English learners can be guaranteed to improve their speaking ability due to various reasons. Moreover, as "speaking" cannot be done individually, failure to practice speaking has caused English learners' declined enthusiasm towards English study. As a result, many English learners find it is too hard to well master this language.

With dozens of students in one class, a teacher can only give a general introduction to the skills and methods of English pronunciation. It is impossible for him/her to offer each student detailed training in combination with the characteristics of English language. As the practice for pronunciation of English needs long-term training, the forty minutes in the class is far less than enough for students to learn the right way of pronunciation. With the ever-growing language learning disability and increasing scientific development, *CAPT*, the computer-assisted paralanguage technology, has come into being. As a branch of *CAPT*, the pronunciation technology of computer language imposes no restriction on requirement to English pronunciation learners. There's a very important part in *CAPT* research that the computer can give ratings to the pronunciation after the English learners speak. This part is of great significance in *CAPT* development.

Many researchers have made great effort in studying computer English pronunciation and have obtained certain achievement abroad and at home. In foreign countries, Carlos Molina et al. have concluded a method to generate a dictionary for evaluating whether the pronunciation is right or wrong. The main feature of this method is the design of Bayes multilevel classifier which can make the subjectivity of the obtained results simpler without verification information. The results obtained through this method show that the coefficients of manual rating result and machine rating result can reach 0.82 and 0.67 respectively. Ambra Neri et al. think the problems existing in pronunciation evaluation are the development of computer technology which is not advanced enough and its application in automatic speech recognition is not of enough experience. Update and optimization are needed in computer programming and algorithm so as to generate a satisfying effect of English pronunciation rating. Jared Bernstein et al. have designed a study rating system through English speech recognition technology in Stanford Research Institute. The system, applying the speech recognition system of Markov model, processes the voice made by the speaker and compares it with the data in the database. At last, it will give ratings based on the differences in between. The experiment has finally made it clear that the coefficients of experts' ratings and readers' ratings can be above 0.8. The system effect is obvious.

Currently, the domestic rating products for pronunciation of English are mainly studied by Tsinghua University, Chinese Academy of Sciences (CAS), University of Science and Technology of China, and other research institutes and universities. One of the famous products is Versant which covers various languages and can give ratings for pronunciation of English by reading and short-answer questions. So far, Versant has been widely used in many areas. Hong Kong Polytechnic University and CAS have jointly

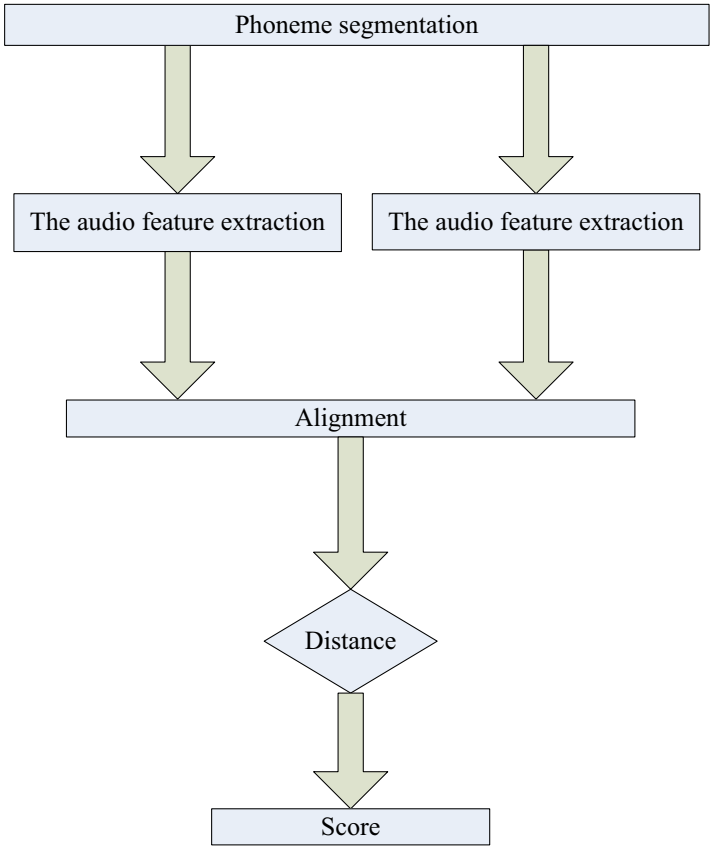

Figure 1. Ratings frame structure chart

developed an evaluating system of English phonetic test which can automatically identify wrong pronunciation. Qingcai Chen et al. have proposed an automatic evaluating method about grammatical verbal pronunciation, directing at addressing the instability of current English pronunciation phoneme system. This method can detect continuous mistakes made in English pronunciation and mispronunciations which are confusing. Weiqian Liang has established a pronunciation quality evaluating system model of English learning system based on the combination of pronunciation quality, pronouncing network generation, evaluating model training and extraction of pronunciation parameters. In the learning process of embedded mode, this system can help realize the goal of using resources with less expense and more stability.

With the research experience and results created by former researchers, this paper will establish a rating algorithm model of English pronunciation which can match the audio feature pattern and obtain the evaluation result that there's high relevance between the operation result of rating algorithm and that of manual rating. Summary and outlook about the English audio rating algorithm will be provided at the end of this paper.

2 FRAME OF ENGLISH PRONUNCIATION EVALUATING ALGORITHM MATCHING AUDIO FEATURE PATTERN

As a processing result of audio information made by brain, people tend to hear what others say more clearly in complex situation. Audio recognition technology can be fully applied in our communication process, so as to help us understand the exact content of what others plan to express and enhance the applicability of the system in noisy environment. In recent years, how to apply the audio feature pattern matching technology in evaluating English pronunciation has become a heated research theme.

As the information processing realized through the frame of English pronunciation evaluating algorithm that matches audio feature pattern is individually completed in two channels, the final results shall be obtained by weighting the result of each channel. According to the evaluating situation of the actual English pronunciation, the detailed progress is as follows: first, set phoneme segmentation; and then, make comparison between the to-be-evaluated pronunciation and the standard pronunciation according to phoneme levels, and calculate the similarity in between. Take the distance between standard template and the

feature of to-be-tested audio factor as the measurement for similarity. At last, use related formulas to make calculation and get rating scores. Detailed steps are as follows:

① make phoneme segmentation of English pronunciation as the input;

② process the above segmentation by specified methods and extract the audio features;

③ apply specified methods to calculate the corresponding European distance;

④ obtain the final ratings through formula calculation.

The figure for the corresponding structure chart is as figure1.

2.1 *Audio Feature Extraction*

At first, process the audio signals of English pronunciation; and then, complete the audio feature extraction. During the process of audio feature extraction, windowing and pre-emphasis are two important parts.

2.2 *Windowing*

Within the range of 10~20ms, audio signals are considered to be comparatively stable. The processed audio segmentation signals can be divided into several sections and each section can be regarded as an analysis frame extracted from a continuous audio section. After processing, the original audio sequence can generate a new audio signal feature. In general, the analysis frames formed by the audio signals extracted from the functions of the same length can be approached through several functions in accordance with different applications, so as to reach an ideal frequency response.

1) The square function formula of an N point:

$$\omega(\mathrm{n})=\begin{cases}1 & (0\le \mathrm{n}\le \mathrm{N}-1)\\ 0 & \text{Others}\end{cases} \quad (1)$$

2) The Hamming function formula of a N point:

$$\omega(\mathrm{n})=\begin{cases}0.54-0.64\cos\left(\dfrac{2\pi}{N-1}\right) & (0\le \mathrm{n}\le \mathrm{N}-1)\\ 0 & \text{Others}\end{cases} \quad (2)$$

3) The Hanin function formula of a N point:

$$\omega(\mathrm{n})=\begin{cases}0.42-0.5\cos[2\pi\,\mathrm{n}/(\mathrm{N}-1)]+0.08\cos[2\pi\,\mathrm{n}/(\mathrm{N}-1)] & (0\le \mathrm{n}\le \mathrm{N}-1)\\ 0 & \text{Others}\end{cases} \quad (3)$$

2.3 *Pre-emphasis*

In order to complete the pre-emphasis on the high-frequency part of the voice, this paper applies pre-emphasis to improve the high-frequency resolution ratio of the pronunciation. Oro-nasal radiation effect and glottis stimulation mainly affect the signal rating power spectrum of the voice while the high frequency will have rapid attenuation when it is above 800Hz. Filter the high frequency and the transfer function corresponding to the filter is as follows:

$$H(\mathrm{z})=1-\alpha Z^{-1} \quad (4)$$

Among which, the pre-emphasis coefficient is α with a range of $0.9<\alpha<1.0$. In general, take α as 0.97.

2.4 *Mel Frequency Cepstrum Coefficient*

Research finding has shown that there's paring relationship between frequency and perceptual skill when the frequency is above 1000Hz. However, when the frequency is lower than 1000 Hz, there's linear relationship between perceptual skill and it. Consequently, *Mel* frequency was proposed to have more accurate audio feature pattern matching. The special perception performance obtained by human ear is called *M*FCC. The corresponding process is as figure 2:

Introduction to the extraction process:

$x(\mathrm{n})$ refers to the time-domain signal of each frame level; and $s(\mathrm{n})$ refers to the original audio signal after processing. According to the feature of the audio signal, windowing can be used for processing. The widely-used windowing is done through Hamming window as it can avoid leaking and can be low-pass.

An N sequence will be formed after processing the time-domain signals. And then, after the discrete Fourier transform, the N sequence will be turned into a linear frequency spectrum. The related transform formula is as follows:

$$\mathrm{X(k)}=\sum_{n=0}^{N-1}x(\mathrm{n})\,\mathrm{e}^{-j2\pi nk/N} \quad (0\le \mathrm{n,k}\le \mathrm{N}-1) \quad (5)$$

Among which, N refers to FFT or the window width of DFT.

After the linear frequency spectrum mentioned above passes through the filter bank, we can get the *M*el frequency spectrum. Within the matching range of audio feature pattern, set several band-pass filters $H_m(\mathrm{k})$, in which the range for m is $0\le m<M$. M refers to the number of filters while the filter center frequency can be expressed as $f(\mathrm{m})$. The expression is as follows:

$$\mathrm{f(m)}=\left(\frac{N}{F_s}\right)B^{-1}\left(B(\mathrm{f}_l)+\frac{B(\mathrm{f}_h)-B(\mathrm{f}_l)}{M+1}\right) \quad (6)$$

The function expressions of filter are as follows:

$$H_m(\mathrm{k})=\begin{cases}0 & (\mathrm{k}<f(\mathrm{m}-1))\\ \dfrac{k-f(\mathrm{m}-1)}{f(\mathrm{m})-f(\mathrm{m}-1)} & (f(\mathrm{m}-1)\le k\le f(\mathrm{m}))\\ \dfrac{f(\mathrm{m}+1)-k}{f(\mathrm{m}+1)-f(\mathrm{m})} & (f(\mathrm{m})\le \mathrm{k}\le f(\mathrm{m}+1))\\ 0 & (k>f(\mathrm{m}+1))\end{cases} \quad (7)$$

Among which, the high frequency and the low frequency of filter can be expressed as f_l and f_h. B^{-1} refers to the B inverse function and the related equation is $B^{-1}(\mathrm{b})=700(\mathrm{e}^{b/1125}-1)$. F_s refers to the

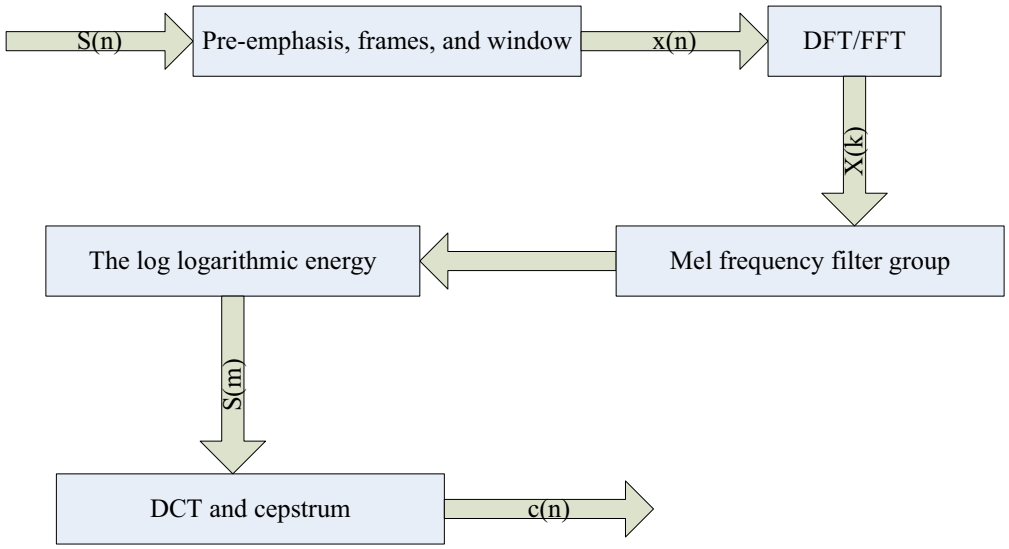

Figure 2. MFCC extraction process

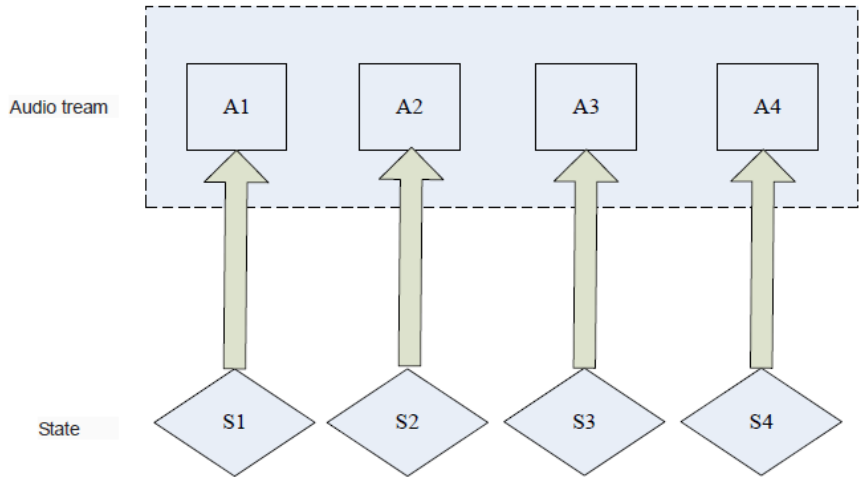

Figure 3. Two strams HMM

frequency of inverse function sampling. The logarithmic power of the *Mel* frequency can bring better robustness to the resulting error.

Total transfer function formula is as follows:

$$S(m) = \ln\left(\sum_{k=0}^{N-1}|X(k)|^2 H_m(k)\right) \qquad (0 \le m < M) \quad (8)$$

The s(m) mentioned above can be turned into cepstrum domain through *DCT* transform, and the corresponding parameter c(n) can be obtained herein:

$$c(n)\sum_{m=1}^{M-1} s(m)\cos\left(\frac{\pi n(m+1/2)}{M}\right) \qquad (0 \le m < M) \quad (9)$$

In the test, take N as 512 and take window length as 20ms. The number of filters is 16 while the sampling frequency is 16K. The window shift is 10ms.

3 OPTIMIZATION MODEL OF ENGLISH PRONUNCIATION RATING ALGORITHM

As English pronunciation can be greatly affected by human impact and environmental influence, *CHMM* (Coupled Hidden Markov Model) can be applied to solve the problems existing in non-isomorphic audio feature matching and asynchronous audio feature matching, so as to complete the reasonable evaluation on English pronunciation rating.

In traditional audio recognition system, audio fea-

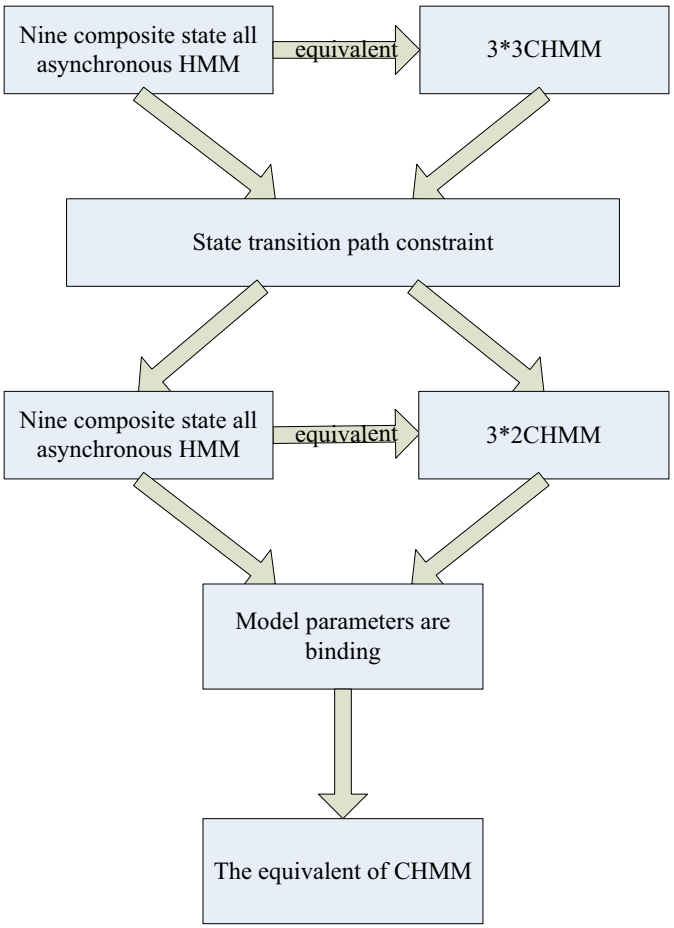

Figure 4.The flow chart of equivalent conversion

tures are integrated into one tream for modeling. In this method, asynchronization may appear during the generating process of language while recognition ability may be hindered due to different frequencies among people. This paper applies coupled *HMM* method for improvement and can overcome the deficiency existing in traditional system through the asynchronous information matching strategy of *CHMM* . See the following figure for the structure:

3.1 *CHMM Equivalent Conversion*

In the parallel chain of *CHMM* , if one state is divided into two discrete values, two states can be generated thereupon. There're two arithmetic products corresponding to the output probability of each state. Double-stram *HMM* can be conversed with coupled *HMM*. Before equivalent conversion, assume that there're three audio states. If the number of audio channels is m, take m as 3 in this case. The basic flow chart of conversion is as figure 4:

3.2 *Training of Observed Value Sequence*

In order to avoid the training without enough theoretical property in traditional system, this paper applies *HMM* training theory. Set observation sequence O^l in which $l = 1, 2, ..., L$ and $O^l = O_1^l, O_2^l, ..., O_{T_l}^l$. The observed values in the sequence are mutually independent and the observed frequency $p(O/\lambda)$ of the model is shown below:

$$P(\mathrm{O}/\lambda) = \prod_{l=1}^{L} P(\mathrm{O}^l/\lambda) \tag{10}$$

Revise *HMM* according to the theory mentioned above. Take several audio samples as the observed values and optimize them.

4 EXAMPLE VERIFICATION

Before conducting the experiment, use the above theories to pre-treat the data and apply the above model to complete training. Set three audio channel statuses

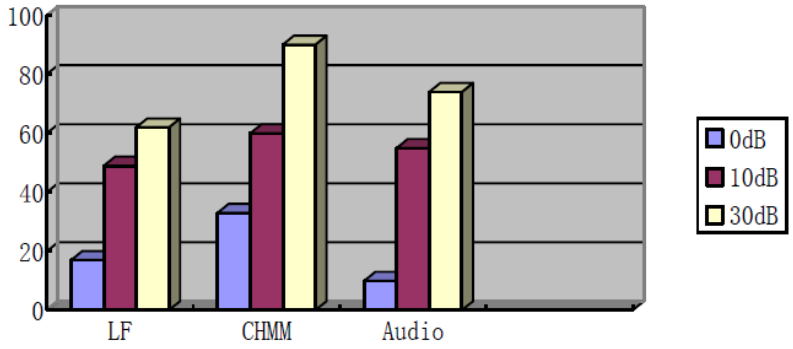

Figure 5.The result of the test

and 8 observed values. Use discrete values to express the observed values. In the experiment, take 10 girls and 10 boys as samples. Train the samples by the seven factors of "YESTERDAY". Take three samples as the test values and regard the others as the training values. The results of "Y" are as figure 5.

In the figure shown above, LF refers to the ratings of audio feature patterns; *CHMM* refers to the ratings of Hidden Markov Model training; and Audio refers to the ratings of audio channel information.

5 CONCLUSION

①This paper applies the rating algorithm matching audio feature pattern. This algorithm has considered the relevance between English pronunciation process and audio channel. It utilizes the matching method of interlayer and has accomplished the pattern matching of audio feature.

② *CHMM* algorithm can solve the synchronous and asynchronous decision problems in phoneme level. *CHMM* algorithm can be realized through *HMM* and can offer ratings based on the final English pronunciation. From the experiment, we have found that the result obtained in *CHMM* model is the best among those three rating methods.

③Due to the limitation of resource and time, this paper takes less "Y" from "YESTERDAY" as research object. A more improved English pronunciation word test database needs to be established in the future, so as to improve the accuracy and precision of the test.

REFERENCES

[1] Chen C.H. & Long W.B. 2011. Evaluation System of Mandarin Pronunciation Quality Based on ARM-Linux. *Computer System Application*, 20(9): 92-96.

[2] Rebeeea H. 2005.*Computer Support for Learners of Spoken English*. Stockholm: The Royal Institute of Technology.

[3] Huang J. 2011. Analysis of Chinglish--from Chinglish to English. *Contemporary Communication*, (8).

[4] Zhao L. 2011. *Voice Signal Processing*. Beijing: China Machine Press.

[5] Sun F. 2008. *Strategy Research of the Spoken English Learning with Computer-assisted Instruction*. Jinan: Shandong University.

[6] Sun X. 2009. *Analysis of Vowel Acoustic Feature from International Phonetic System*. Nankai University.

[7] Dong B. 2007. Objective Test method research of the syllable rime pronunciation level in mandarin based on formant pattern. *Acoustics Study*, 32(2): 111-119.

[8] Liang W.Q., Wang G.L. & Liu J. et al. 2005. Evaluating Algorithm of the Pronunciation Quality Based on Phoneme. *Journal of Tsinghua University Natural Science Edition*, 45(1): 9-12.

[9] Yan K. 2009. *Automatic Evaluating of English Reading Questions and Retelling Questions Technical Research Field*. Hefei: University of Science and Technology of China.

[10] Nie X.P. & Liu J. 2010. Evaluating System of the English Pronunciation Quality Based on ARM9.Audio Engineering, 34(8): 14-24.

[11] Warschauer M. 1996. *Computer Assisted Language Learning*: an Introduction: Multimedia Language Teching, Tokyo: Logos International, pp: 4-19.

[12] Adriana Garcia Kunzel. 2010. An Android Approach to the Web Services Resource Framework. Florida Atlantic University.