Intelligent Computing Technology and Automation
Z. Hou (Ed.)

doi:10.3233/ATDE231288

An GMM-HMM Based Intelligent Scoring Mechanism for Oral English Pronunciation

Zhiwei YANG[a,1], Yiwei LIU[a] and Qinghai ZHANG[a]
[a] *Army Logistics Academy of PLA, Chongqing, 401331, China*

Abstract. To improve the ability of intelligent detection of oral English pronunciation errors, this paper proposes an intelligent scoring model for pronunciation based on related speech recognition principles. The scheme is based on automatic speech recognition technology. First, MFCC is used to process the speech signal from the oral test system. Second, the corpus is standardized by the posterior probability scoring method based on the Hidden Markov Model(HMM), to fit the distribution of speech feature observation vectors. Finally, the pronunciation error detection is recognized according to the spectrum difference. The simulation results show that the mechanism presents high accuracy in the automatic evaluation of oral pronunciation quality, which helps to improve the self-learning ability of English learners.

Keywords. GM-HMM; pronunciation; intelligent scoring; MFCC; oral English

1. Introduction

The computer-aided pronunciation learning system can accurately evaluate users' pronunciation and provide practical pronunciation guidance for users. With the development of the global economy, mastering one or more foreign languages has become a social skill demand. In traditional English teaching, teachers focus on reading, writing, and listening comprehension but pay little attention to the training of students' oral competence, which leads to students' inefficient oral performance. Using computers to assist with oral pronunciation is one of the fundamental ways to improve oral English level. Speech recognition is an indispensable part of computer translation software. It mainly identifies and processes different languages to help students quickly understand the meaning of English knowledge. Speech recognition technology mainly includes three major technologies: feature extraction technology, pattern matching technology, and model training technology, which are also key considerations in existing system design. An oral English learning system based on speech recognition technology applies automatic speech recognition technology. It automatically compares the user's phonetic phoneme series with the standard model's phonetic phoneme sequence, and provide the comparison results visually and intuitively to the user as a judgment basis through the view. This paper involves speech recognition, second language teaching, phonetics, and many other disciplines. Effective feedback is of great help to learners in learning single word pronunciation [1]. The emergence of the computer-aided language learning (CALL) system provides a suitable learning environment for oral English learners. In a call system, the key part that effectively guides learners to learn oral English efficiently is the scoring mechanism focusing on

[1] Corresponding Author: Zhiwei Yang; Army Logistics Academy of PLA, Chongqing, 401331, China; zhiwei_yangyang@163.com

evaluating learners' pronunciation [2].

In order to provide effective feedback and guide learners to correct their wrong pronunciations and improve their ability to speak English, this paper mainly studies spoken English training and pronunciation correction technology based on speech recognition technology. The advantages and disadvantages of these algorithms are summarized through the in-depth discussion of the English pronunciation and its existing scoring schemes, including endpoint detection, feature extraction, and HMM scoring model. According to the practical needs of the college English oral test, this paper designs an evaluation mechanism based on GMM-HMM logarithmic posterior probability. According to the characteristics of English pronunciation, we use linear mapping to standardize the computer scoring and use the machine learning method to calculate the final score of learners, and adopt multidimensional mixed Gaussian distribution density function to recognize feature vectors of speech. The simulation results show that the system has good intelligent detection performance, high detection accuracy, high anti-interference and error correction ability.

2. Scoring Mechanism and Related Techniques of The Oral Pronunciation

2.1 Principle Idea

The scoring mechanism is the core technology of the oral English learning system. Its primary purpose is to match the recorded voice data with the preset standard value. Through machine learning and recognition technology, this paper analyzes the interference factors and problems of the expert language practitioners so as to give recommended guidance and suggestions to help them gradually improve their pronunciation. Therefore, it is of practical significance to evaluate pronunciation to guide learners' oral English learning. Voice scoring technology is to define the accuracy of the speaker's pronunciation. Most existing technologies are implemented by DTW or HMM [3]. These two techniques need to compare the feature parameters of the speech to be evaluated with those of the standard template or reference model. The former is mainly applied to speech speed variation and other speech feature parameters of different speakers at different times. The latter decodes the observation state queue of each layer, obtains the hidden state queue as the media quality score, and then obtains the IP network media quality analysis result queue by a weighted average of the media quality scores, as depicted in figure 1.

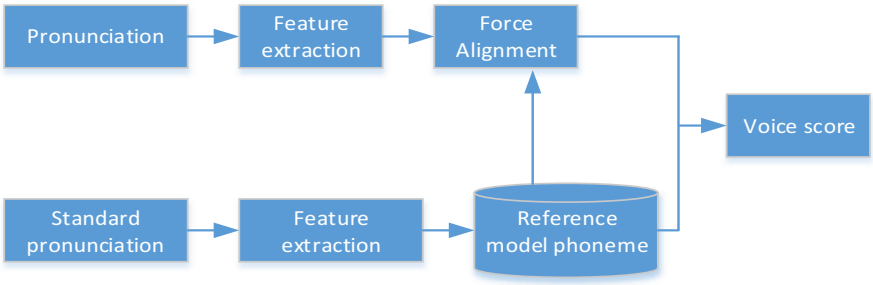

Figure 1. Traditional pronunciation scoring flow.

2.2 Related Technology Research

(1) Selection of speech feature parameters

Speech recognition has the following requirements for feature parameters:

- It can transform speech signals into speech feature vectors which a computer can process
- It can match or be similar to the auditory perception characteristics of the human ear
- To a certain extent, it can enhance speech signals and suppress non-speech signal

The main characteristic parameters include LPCC, LPCC, and MFCC. People produce sound through the vocal tract, and the shape of the vocal tract determines what kind of sound they make, so we can accurately describe the phoneme produced. The shape of the channel is shown in the envelope of the short-term power spectrum. MFCC is an accurate description of the envelope of a feature [4]. This paper mainly evaluates pronunciation, which is closely related to the characteristics of human pronunciation and acoustic perception.

(2) GMM- HMM

HMM is a kind of Markov chain. Its state cannot be observed directly, but it can be observed through the sequence of observation vectors. Each observation vector is expressed as various states through some probability density distribution, and is generated by a state sequence with corresponding probability density distribution. The model has been widely used in speech recognition and applied in speech recognition, machine translation, partial discharge analysis, cryptanalysis, gene prediction, and other related fields [5].

GMM-HMM is a classic acoustic model, and speech recognition technology based on deep neural networks actually replaces GMM with neural networks to model the observation probability of HMM. The lattice modules of recognition processes such as modeling and decoding still use classic speech recognition technology GMM will learn all the training samples once. Train the emission matrix similar to that in discrete Markov networks. That is, state ->observation value. In this way, HMM can train its own state transition matrix, etc., so that HMM can be used. Train one HMM type for each word. As long as the signal is extracted with feature values through MFCC, MFCC can obtain a matrix of feature dimension * frame number. That is to say, the observation sequence is a matrix of feature dimension * frame number.

(3) Sphinx speech recognition

Sphinx is an extensive vocabulary speech recognition system written in java language, which adopts continuous hidden Markov model modeling. Compared with previous versions, sphinx has improved modularity, flexibility, and algorithm, and at the same time, adopts a new search strategy, and supports various grammar and language models, auditory models, and feature flows [6]. The innovative algorithm allows multiple information sources to be merged into an elegant knowledge rule which is more consistent with the actual semantics. Due to the full development of Java language use, a highly flexible multi-threaded interface with a high degree of portability and multi-threaded technology is allowed. We try to use Sphinx 4 to realize speech recognition clients. The goal is to let someone read written sentences in the text and have a confidence score for each word. The main modules are front-end, decoder, and knowledge base.

3. GMM-HMM Based Intelligent Scoring of College Oral ENGLISH

3.1 Overall Framework of System

Figure 2 depicts the basic principle of the oral English learning system. The standard speech database is composed of the authoritative knowledge database downloaded from the Internet and the temporary database of pronunciation experts selected by us, which can meet the different test needs of students at the same time. The model is classified according to the difficulty of pronunciation and word structure. These words are further labeled in the expert knowledge base for training in a high-intensity environment. The system first collects the phoneme data of all the speakers and stores them in the DBS, and then forms small units for calculation through noise filtering, feature extraction, classification training, and other operations. Next, we use GMM-HMM to evaluate the pronunciation quality, intelligently give the corresponding level and summarize the reasons for this result. We then compare the results with the standard value, combined with the expert evaluation score, and give some suggestions on correcting pronunciation errors.

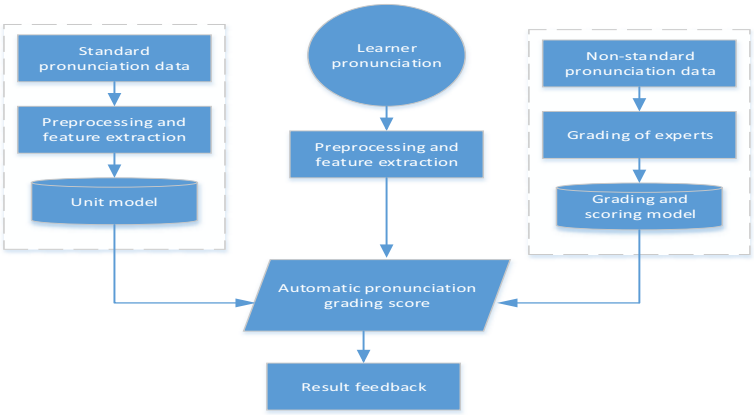

Figure 2. Overall functional structure diagram of the system.

3.2 Data Preprocessing

Sound signal processing is an essential technology of digital signal processing. This paper uses the method of digital signal processing and MATLAB function to distinguish the non-speech part and speech part of the signal [7]. Audio endpoint detection is to detect effective speech segments from continuous speech streams. It includes two aspects: detecting the starting point of effective speech, that is, the front endpoint, and detecting the ending point of effective speech, that is, the back endpoint. The first simple point is to separate the effective voice from the continuous voice stream in the scene of storing or transmitting voice, which can reduce the amount of data stored or transmitted. Secondly, in some application scenarios, endpoint detection can simplify human-computer interaction. For example, in the recording scene, endpoint detection after voice can omit the operation of ending the recording. Many amplitude segments of the speech signal are generated in the acquisition process. In order to transmit the acquired sequence on the digital communication channel or store the processing signal

in the digital memory, it is necessary to use the limited symbol set to represent the amplitude value quantized into the limited amplitude set in the computer. The actual collected audio usually has a certain intensity of background sound, normally background noise. When the background noise intensity is high, it will significantly impact the effect of speech application, such as the reduction of speech recognition rate and endpoint detection sensitivity. Therefore, noise suppression is vital in the front-end processing of speech.

We adopt MFCC to compare the speech and atlas for oral English. Then the acquisition sensor of the spoken English pronunciation signal is a uniformly distributed sequence, and the frequency domain of the characteristic of the spoken English signal of the reference array is $\mathrm{v}(\mathrm{t},\theta)$:

$$\mathrm{v}(\mathrm{t},\theta)=\sum_{m=1}^{M}\omega_i^*(\theta)=\sum_{m=1}^{M}x_i^*(t)\omega_i(\theta) \tag{1}$$

$*$ is the complex conjugate operator. The adaptive beamforming method is to perform time-domain matching and adaptive filtering, acquiring the frequency domain characteristics of the output signal as:

$$\mathrm{v}(\mathrm{t},\theta)=\sum_{m=1}^{M}\omega^H(\theta)x(t)=x^H(t)\omega(\theta) \tag{2}$$

H is complex conjugate transpose; $x(t)$ and $\omega(\theta)$ are instantaneous time-domain signal component and weighted component output by oral speech, which can also be denoted as

$$x(t)=[x_1(t),x_2(t),\ldots x_M(t)]^T \tag{3}$$

$$\omega(\theta)=[\omega_1(\theta),\omega_2(\theta),\ldots,\omega_M(\theta)]^T \tag{4}$$

3.3 Feature Extraction

Based on the feature decomposition of English pronunciation signal by using a multi-layer wavelet feature scale transform, the automatic evaluation algorithm of English pronunciation quality is optimized. The specific process of one-dimensional multi-level wavelet transform is as follows: the original signal is decomposed into low-frequency and high-frequency parts through two complementary symmetric filters. Then, the low-frequency part of the signal is decomposed with the same processing process. After multi-level decomposition, the original signal is decomposed into multiple signals. The low-frequency signal reflects the overall characteristics, while the high-frequency signal reflects the detailed characteristics. Finally, there is a one-to-one mapping relationship between the output independent phases R "and X"：

$$p(R^N=r_i)=p\begin{pmatrix}X^N=x_i\mid x_i\mid=\mid r_i\mid,angle(x_i)\\=(angle(r_i)-\Phi g)\bmod(2\pi)\end{pmatrix} \tag{5}$$

When the phase distribution $angle(X^N)$ of oral English signal has a uniform distribution on $[0,2\pi)$, R^N and Φg are independent. Then the phase information $I(R^N;\Phi g\mid Z^N)=0$ of the speech signal of the energy set $\{P_1,P_2,\ldots,P_j\}$ can be acquired as

$$H(X^N\mid Z^N)=H(R^N\mid Z^N)+(\Phi g\mid Z^N) \tag{6}$$

The wavelet entropy feature of the speech signal is extracted as:

$$H(R^N) = -\sum_{i=1}^{M} p(r_i)\log(p(r_i)) = H(X^N) \tag{7}$$

M is the number of elements in the symbol set. The adaptive filter parameters of oral English pronunciation are: $N^{(1)} = N, N^{(j)} = N_0^{(j-1)}\ 2 \le j \le J$,. According to the above analysis, intelligent speech recognition can be performed to improve the auto evaluation ability of the English pronunciation.

3.4 Voice Scoring

It can be concluded that the whole GMM+HMM network is mainly for the HMM network service. First, we discuss the problems that HMM needs to solve for speech recognition, such as correctly identifying a series of MFCC features into corresponding HMM state series. This process involves two probabilities that need to be learned [8]. One is to identify the characteristics of the current frame as the probability of this state, that is, the likelihood in HMM -- here refers to the computing level, that is, the GMM network is used to obtain the current state probability, and the other is the probability of the last state to be converted into this state, namely the transformation probability. In theory, there are exponential transformation methods for one sequence to another, so each frame takes only the state with the highest probability. Such a route selection method is called the Viterbi method.

We make S_a a score based on HMM logarithm posterior probability and l is the pronunciation duration of the primitive, then the actual score obtained by the spoken language learning system is:

$$S_r = S_a / l \tag{8}$$

4. Simulations

Considering the short-term stationarity of speech, the front-end signal processing of speech signals requires adding windows and frames, and the recognition features are extracted based on frames, as shown in Figure 2. Extracting speech features frame by frame from segmented speech signals for acoustic model modeling. TIMIT speech database is used as the standard speech to train the system. TIMIT speech database has accurate phoneme labeling to be applied to speech segmentation performance evaluation. At the same time, the database contains a large amount of speakers' speeches, so it is also the authoritative speech database widely used to evaluate speakes' recognition. More than 500 speakers record the speech database, and each person reads more than ten sentences with comprehensive phonemes. The acoustic model is mainly used to calculate the likelihood between speech features and each pronunciation template. The goal is to establish a set of model parameters for each acoustic unit. The training process has a large amount of data and complex operations, but the training process for each sentence is basically the same.

The experiment takes an example sentence in the standard speech database to describe the process of model training. The basic information is shown in table 1. The speech database should be able to better represent the speech tobe recognized, and it is better to use the recording data of multiple people, including multiple recording situations and all possible linguistic sentences. The speech database consists of two parts:

a training set and a test set. Generally speaking, the test set accounts for 1/10 of the entire database, but it is better not to exceed 4 hours of recording time. In addition, it needs to be mapped into a single phoneme according to the dictionary database, which becomes a sentence composed of phonemes, that is, the content of "phoneme information" in the table.

Table 1. Basic information of example sentences

Sampling frequency	**16KHx**	**Length**	**3.665s**
Quantization Bits	16bits	Vocal tract	Single tract
Bit rate	256Kbps	Format	MP3
Content	The language difference in pronunciation has puzzled data scientists for many years, where the training model needs a lot of data.		
Phoneme	Pau ow l ae s z iy s b t w er pau a f th ts k ax t pu da …		

After obtaining the basic information in Table 1, we parameterize the audio data and use the MFCC parameters in this system. First, we pre-apply the voice data, divide it into different windows, and then output the parameters of each window. The result of the form extraction is shown in figure 3, where the information about the voice parameters shows that the sampling point is the frame length. When the voice frame length is set to 20 meters, the number of sampling points in each frame is 400. In the above figure, the size of MFCC set for configuration parameters is 10 in this system.

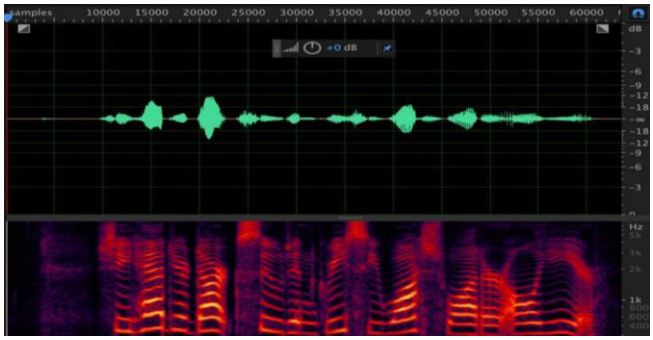

Figure 3. Parameterization of the audio frame.

The evaluation criteria of pronunciation quality can be divided into five grades: excellent, good, average, poor, and inferior. Several testers are invited to complete the test on the training set according to the strategy proposed in this paper. Compared with the expected score given by experts, the evaluation result list, as shown in Table 2, is obtained. From the analysis of the test results, it can be concluded that the scoring mechanism can help to improve the users' pronunciation learning to a certain extent because of the adaptive method of machine learning. However, in the case of high load, the actual value of some scoring deviates from the expected value. Overall, the efficiency of intelligent scoring is high, and the performance is good.

Table 2. Analysis of test results of specific samples

Person	Results	Experts	Available	Effectiveness
A	81.5	65.7	76.5	76.2
B	84.6	70.1	92.0	84.6
C	72.0	88.9	83.5	87.2
D	66.6	61.2	80.9	73.4
E	52.3	74.5	91.2	86.1

F	58.4	87.2	98.2	85.4
G	80.1	76.3	79.1	85.3
H	61.2	65.4	90.5	89.2
I	61.5	87.3	71.8	85.1
J	70.4	78.5	79.9	76.4
Total	73.5	72.1	82.4	82.2

5. Conclusion

In order to improve the artificial intelligence and accuracy of students' oral English tests, this paper uses speech signal processing technology to evaluate the quality of College English oral intelligence. We mainly discuss the common speech pronunciation principle and special diagnosis extraction method. We select the speech signal matching technology for filtering and use GMM-HMM technology to do further spectrum tests for the extracted specific features. The probability density distribution function of speech feature vectors is fit by using a mixed Gaussian model so as to output the quality score in line with the expert standard. The research shows that this method has high accuracy and exemplary performance in pronunciation error detection of the English oral test system.

References

[1] SONG Fangfang, SONG Xiaoli, MA Qingyu. The Research of the Scoring Mechanism of Spoken English Learning System Based on the Technology of Speech Recognition. Computer Knowledge and Technology, 2009, 5(7): 1726-1728

[2] XIE Xuemei. Research on the intelligent detection technology of pronunciation errors in spoken English test system. Automation & Instrumentation;, 2018, 230(12):64-67

[3] Weon, Hee, Yun. The Objectives of English Pronunciation Evaluations and the Usability of Machine Scoring. The Journal of Linguistics Science, 2012, 61:167-184

[4] Atia M M, Hilal A R, Stellings C, et al. A Low-Cost Lane-Determination System Using GNSS/IMU Fusion and HMM-Based Multistage Map Matching. IEEE Transactions on Intelligent Transportation Systems, 2017, 18(11):3027-3037

[5] Song Y, Liang W. Application of discriminative HMM in automatic English pronunciation assessment. Journal of Tsinghua University(Science and Technology), 2010, 50(4):503-506.070

[6] Nakamura N, Nakagawa S. English pronunciation evaluation for Japanese students. IEICE technical report. Speech, 2002, 102:13-18

[7] Wijekumar K, Meyer B J, Lei P, et al. Supplementing teacher knowledge using web‐based Intelligent Tutoring System for the Text Structure Strategy to improve content area reading comprehension with fourth‐ and fifth‐grade struggling readers. Dyslexia, 2020, 26(2):120-136

[8] Deng Jiangyun, Li Cheng. Speech Recognition Garbage Classification System Based on GMM-HMM. Modern computers,2020,26: 27-31