# English Pronunciation Error Detection Based on Multimedia Data

Lijuan Wu

Heilongjiang International University (English Department)

cerrity1982@sina.com

**Abstract.** Pronunciation is an interdisciplinary subject, and speech recognition is gradually becoming the key technology of man-machine interface in information technology. In recent years, the application of computer speech recognition has made great progress. Based on the special status of English, there have been many designs and productions of English as a second language speech database in the world. However, due to the increasing popularity of English, more and more people use English as a second language, so it is necessary to establish an English pronunciation error detection system for multimedia data. In this paper, the initial data … is designed, made and trained as a model, and the data of standard phonetic database (using the existing database of AVICAR) is tested and compared with the collected phonetic database. It is found that the recognition rate of collected voice data is much lower than that of standard voice data, and the conclusion is drawn that it is important to collect voice databases from different regions. The reason of low recognition rate is analyzed. Then, the data collected in the phonetic database are compared with each other according to different regions, and the reasons for the difference in recognition rate are summarized, which provides reference experience for the design and manufacture of phonetic database.

**Keywords:** media data; English; Pronunciation; detection

## 1 Introduction

With the progress of multimedia technology, the development of learning theories such as constructivism, the rise of college English reform and the emergence of a series of contradictions in the current situation of college English teaching, it will become a meaningful research to promote college students' autonomous English learning by using multimedia [1]. Through the literature review of "college students' English autonomous learning", it is found that there is a lack of research on teaching resources to ensure college students' English autonomous learning, and multimedia teaching software, as a common resource type, has more research value. The design of multimedia software determines its development, management, evaluation and other aspects, so the design of multimedia software plays a vital role in the quality of multimedia teaching software[2].


A. Rauf et al. (eds.), *Proceedings of the 3rd International Conference on Management Science and Software Engineering (ICMSSE 2023)*, Atlantis Highlights in Engineering 20,
https://doi.org/10.2991/978-94-6463-262-0_91

## 2 The basic structure of English pronunciation recognition system

Generally speaking, the process of English pronunciation recognition is a process of pattern recognition and matching. In this process, firstly, the speech model should be established according to the characteristics of human voice, and the input speech signal should be analyzed, and the required features should be extracted, on this basis, the template needed for speech recognition should be established. In the process of recognition, we should compare the characteristics of the input speech signal with the existing speech templates according to the overall model of speech recognition, and find out a series of optimal templates that match the input speech according to certain search and matching strategies[3-5]. Then, according to the definition of this template number, the recognition result of the computer can be given by looking up the table. Figure 1 shows the basic structure of English speech recognition.

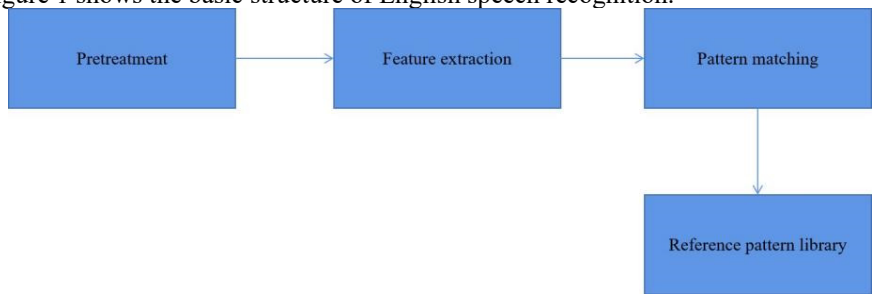


**Fig. 1.** Basic structure of speech recognition

### 2.1 Design and Collection of English Pronunciation Multimedia Database

The voice database contains three parts: personal information, numbers, English letters and short sentences. The language is English. The recording time of a single person is about 1. 5 hours.

(1) Personal information part This part includes five items: the name, gender, age, native place and oral English level of the recorder.

The number part includes three parts: single number, common simple number and number string. Individual figures: 0, ..... 9. Among them, the pronunciation of 0 is divided into oh and zero, and it is distinguished by 0 and z when recording. Commonly used simple numbers: including emergency numbers commonly used in China, such as 110, 119, 120, 122, 10000 and 12315. Read it once with 0 and z respectively. Number string: It comes from two parts. One part is a number string containing 3, 4 and 7 in TIDIGIT, with a total of 110 sentences. The other part comes from AVICAR database, with a total of 200 sentences[6].

(2) English letters and short sentences

The content includes three parts: English 24 letters, five vowels A, E, I, O, U and English short sentences (due to the average spoken English level of the collected peo-

ple, the effect of some collected English short sentences is not ideal, so the short sentences are not collected in this database for the time being). The vowel part requires each recorder to read it five times. There are 200 English short sentences in total, which come from AVICAR database. The selection of these English short sentences takes into account the balance of phonetic phonemes. According to the actual needs of the database, when selecting the recorded content, more numbers and letters are covered, such as numbers 3, 4, 7 and vowels, which are quite different from those used as a second language.

### 2.2 Construction of HTK speech pronunciation recognition system

HTK (HMM Toolkit) is a tool for constructing HMM (Hidden Markov Model), which is an open source speech recognition toolkit developed by the machine intelligence laboratory of the engineering department of Cambridge University. The tool includes a series of library modules and tools written in C language. HTK is mainly used to construct speech recognition system based on HMM, and it can also be used to construct HMM of other time series signals. HTK is a training and recognition tool adopted by most speech recognition systems at present. Because it is an open source tool, users can add code to it if necessary to get other special functions.

Hidden Markov Model (HMM) is a statistical model commonly used to model sequence data with potentially unobserved states. It is composed of three basic components, namely state set, observation set, and state transition probability and observation probability.

State set: The system in HMM is assumed to be in one of a series of discrete states. These states may be hidden (unknown) or unobservable. The set of states is usually represented by S, where S={s1, s2,..., sn}, where n represents the total number of states.

Observation set: The system in HMM can be observed to generate a series of observation values, which may be indirect observations of the internal state of the system. The set of observations is usually represented by O, where O={o1, o2,..., om}, where m represents the number of types of observations.

State transition probability and observation probability: HMM assumes that there is a transition probability between states, that is, the probability of the system transitioning from one state to another at a certain moment. These state transition probabilities can be represented by State-transition matrix A. In addition, HMM also assumes the probability of generating a certain observation value for each state, which can be represented by the observation probability matrix B.

Firstly, the pronunciation dictionary, HMM and binary grammar language model are generated, and then the binary language model is converted into word net, and the tool HBuild is used here. Then, the dictionary, HMM and word net are input into the HNET module of HTK, and the functions in this module can generate the corresponding HMM recognition network. Words in the pronunciation dictionary can have different pronunciations. In this case, each pronunciation in the network corresponds to a branch. After the HMM network is built, it can be input into HRec (HTK module for processing recognition in HTK) to recognize speech. It should be noted that the construction of

HMM network is carried out at the time of identification, which is part of the initialization process[7].

After completing the training of the model, it is the recognition of voice data. HTK provides a recognition tool HVite, HVite can be used to recognize sound files stored on disk and handle timely voice input. After the test is completed, it is the analysis of the results, and the analysis and recognition results are realized by HTKI with Hresults. Hresults compares the recognition result with the standard text, and obtains some statistical results. It is the optimal string matching based on Dynamic Programming(DF). Let N be the total number of symbols recognized, S be the number of substitution errors, D be the number of deletion errors, and I be the number of insertion errors. The Percent Correct rate can be calculated by the following formula (1):

$$percentcorrect = \frac{N - D - S}{N} * 100 \tag{1}$$

## 3 Test and analysis of experimental parameters

In this part, the characteristic parameters and the number of states are tested: firstly, various characteristic parameters are tested, and the characteristic parameters with the highest recognition rate are selected to continue the later experiments; Then, the recognition effect of different state numbers is tested. From 5 to 12, the change of state numbers also affects the recognition rate. In order to achieve the best effect, we take the state number with the highest recognition rate for the following speech recognition test[8].

### 3.1 Characteristic parameter test

By setting different characteristic parameters, the model is trained on the constructed HTK speech recognition system, and the numbers and letters are tested and compared with the speech data in AVICAR. In all the experimental tests, we take the hidden Markov model with the state number of 10, and the Gaussian density is 4. The results show that the number of feature parameters affects the recognition rate: with the increase of feature parameters, the recognition rate also increases. The characteristic parameters that achieve the best recognition effect are: MFCC+OMFCC+OOMFCC+OE+AOE as shown in the following table 1:

**Table 1.** Comparison Table of Characteristic Parameters Test

| Characteristic Parameters | numbers | letters |
|---|---|---|
| MFCC+ O MFCC | 37.21 | 47.67 |
| MFCC+ O MFCC+O E | 38.52 | 49.24 |
| MFCC+△MFCC+O E+OO E | 40.31 | 52.25 |
| MFCC+ O MFCC+ 0o MFCC | 43.87 | 56.68 |

The sampling frequency of the experimental speech is 16KHz, the accuracy is 16bit, and the frame length and frame shift are 25ms and 10ms, respectively. The feature uses the 39 Mel-frequencycepstrumcoefficient (MFCC) parameter. The HMM model used to enforce alignment was trained with 120 hours of data. The Gaussian mixture number of GMM model is 16. For GLDS-SVM, its polynomial variation is of order 3, and the ratio of positive and negative samples in model training is controlled at about 1:1. SVM model training uses Libsvm. Falseacceptancerate (FAR) and Falserejectionrate (FRR) are used to measure the error detection performance of the system. Detecterrortradeoff (DET) and Equalerrorrate (EER) are also provided.

### 3.2 State number test

In the model of speech recognition, we introduce the hidden Markov model, train the model on the constructed HTK speech recognition system by setting different state numbers, and compare the numbers and letters with the speech data in AVICAR. The characteristic parameters we use here are: MFCC+△MFCC+△O MFCC+O E+OO E, the Gaussian density is 4, and the number of states is 5, 7, 10 and 12 respectively. The following table 2 can be obtained:

**Table 2.** State Number Test Table

| number of states | numbers | letters |
|---|---|---|
| 5 | 48.32 | 62.4 |
| 7 | 49.76 | 63.65 |
| 11 | 52.74 | 68.03 |
| 12 | 50.1 | 65.4 |

It can be compared that the accuracy rate is the highest when the number of states is 10. In later experiments, we set the number of states to 10.

### 3.3 Model improvement test

Through the experimental results, it is not difficult to see that when using the model trained by the data (digital part) in TIDIGIT to test the voice, the standard voice recognition result in TIDIGIT is more than 40 percentage points higher than the collected voice data of China people speaking English. This is because the trained model is aimed at the recognition of people whose first language is English, but it is not suitable for China people whose second language is English. So we will make some improvements to the model and then test and compare the results. Add the voice data of China people in AVICAR to the TIDIGIT data used in training the model (there are more than 1000 files). Select the voice data of China people in AVICAR (when the car is stationary), and add them to TIDIGIT once (TIDIGIT+1*AVICAR_ CHINESE_ IDLE), twice (TIDigit+2 * Avicar _ Chinese _ Idle) and thrice (TIDIGIT+3*AVICAR_ CHINESE_ IDLE) respectively[9]. Experimental results are as shown in Table 3:

**Table 3.** Model test comparison table

| | TIDIGIT (Acc%) | TIDIGIT+1*AVICAR_CHINESE_IDLE (Acc%) | TIDIGIT+2*AVICAR_ CHINESE_IDLE (Acc %) | TIDIGIT+3*AVICAR_ CHINESE_IDLE (Acc %) |
|---|---|---|---|---|
| Guizhou people | 45.36 | 49.4 | 51.11 | 51.85 |
| Can-tonese | 39.45 | 42.61 | 43.24 | 44.04 |
| Hunanese | 46.52 | 50.7 | 52.07 | 52.92 |

According to the above table, we can see that after adding the data from AVICAR to the TIDIGIT training model for the first time, the test Acc results are generally 3 to 4 percentage points higher than without adding time. With the subsequent repetition, the Acc results continue to increase by more than 2 percentage points. When adding again, although the improvement of Acc is not as significant as the previous two times, it still shows a slow increasing trend. It can be seen that the trained speech model should focus on the importance of recognizing the object.

Establish a model on the collected voice database and match it with the tested voice. If the model matches the tested voice, the recognition rate will be higher. We can call the model a good model, and the voice database is a relatively successful database. On the contrary, if the model does not match the tested speech, the recognition rate will be greatly reduced[10].

## 4 Conclusion

Different languages have different phonetic databases, including English phonetic database, Chinese phonetic database and Korean phonetic database. The quality of speech database directly affects the speech recognition rate. Generally speaking, a database should include enough phonetic data in different languages and regions. Establish a model on the collected voice database and match it with the tested voice. If the model matches the tested voice, the recognition rate will be higher. We can call the model a good model, and the voice database is a relatively successful database. On the contrary, the model does not match the tested speech, and the recognition rate will be greatly reduced. Obviously, this model is not a good model, and this speech database is also a failed database. Obviously, the phonetic database of a certain language is highly targeted, and the model based on the phonetic database of a certain language can only test that language. In order to achieve a high recognition rate, even the phonetic database of the same language has different classifications. For example, in China, because of its vast territory and abundant resources, there are many dialects, so it is necessary to establish a phonetic database for each dialect. This paper proves this point through the experimental phonetic test.

## References


1. Pennington, M. C. . (2022). Review of kirkova-naskova, henderson & fouz-gonzáles (2021): english pronunciation instruction: research-based insights:. Journal of Second Language Pronunciation, 8(2), 304-308.
2. Tsunemoto, A. , & Mcdonough, K. . (2021). Exploring japanese efl learners' attitudes toward english pronunciation and its relationship to perceived accentedness:. Language and Speech, 64(1), 24-34.
3. Yan, C. . (2022). A research proposal on applying chinese phonetic system in teaching pronunciation of english words to older chinese efl adult learners. Journal of Higher Education Research, 3(1), 21-25.
4. Suzukida, Y. , & Saito, K. . (2022). What is second language pronunciation proficiency? an empirical study. System, 106, 1027（4）12-.
5. Nostrand, P. , & Horslund, C. S. . (2022). Segmental error patterns in finnish-accented english. European Journal of Applied Linguistics, 10(1), 109-141.
6. Suciati, S. , & Diyanti, Y. . (2021). Suprasegmental features of indonesian students' english pronunciation and the pedagogical implication. SAGA Journal of English Language Teaching and Applied Linguistics, 2(1), 9-18.
7. Martens, W. L. , & Wang, R. . (2021). Applying adaptive recognition of the learner's vowel space to english pronunciation training of native speakers of japanese. SHS Web of Conferences, 102(2), 01004.
8. Hong, Y. , & Nam, H. . (2021). Evaluating score reliability of automatic english pronunciation assessment system for education. Studies in Foreign Language Education, 35(1), 91-104.
9. Piotrowska, M. , Czyewski, A. , Ciszewski, T. , Korvel, G. , Kurowski, A. , & Kostek, B. . (2021). Evaluation of aspiration problems in l2 english pronunciation employing machine learning. The Journal of the Acoustical Society of America, 150(1), 120-132.
10. Cao, Q. , & Hao, H. . (2021). Optimization of intelligent english pronunciation training system based on android platform. Complexity, 2021(4), 1-11.