Research on Speech Recognition Technology in English Pronunciation Similarity

Jiajun Hu
Nanyang Polytechnic
hujiajun19961024@gmail.com

Abstract. Authentic English pronunciation is very important in daily communication and mutual understanding. With the improvement of automatic speech recognition technology, the computer-aided pronunciation practice (CAPT) system can already provide limited interaction for second language learners. Pronunciation similarity comparison is a key step in building a system, designing a model that recognizes the similarity of English pronunciations. By calculating the similarity between the learner's pronunciation and the standard voice, and comparing with the expert's score on the learner's pronunciation, the correlation between the two is calculated to verify the feasibility and effectiveness of the model.

1. Introduction
This paper believes that the functions of a computer-assisted learning system (CALL, Computer Assisted Language Learning) based on automatic speech recognition technology (ASR, Automatic Speech Recognition) can be summarized in five stages: (1) Speech recognition stage: This is the first and most important link, because the implementation of subsequent stages depends on the accuracy of the results of this stage. (2) Scoring: At this stage, the quality of pronunciation can be evaluated in the form of scores. The ASR system can analyze the speech previously recognized by the speech recognition stage. By analyzing and comparing the speech features of the learner's speech and the related speech features of the standard speech, the closer the two are, the higher the score. (3) Error detection: The CASR system based on ASR can also detect the position of the wrong pronunciation in a section of speech, and tell the learner where the error occurred, helping them to pay attention to and strengthen the practice in a certain part. Error diagnosis: The ASR system can identify specific types of errors and then give suggestions on how to improve them. (4) Feedback: This function needs to consider more about the design issues of the graphical user interface, which can be included in the information displayed to learners at the stage (2) (3) (4). (5) The main work of this paper focuses on (1) (2), including the pre-processing that needs to be done before the voice signal feature parameters, the method of voice signal endpoint detection, the extraction process of learners and standard voice feature parameters, and proposes the calculation of speech similarity based on EMD and the dynamic planning method of the maximum matching of two conditions [1].

2. Preprocessing of voice signal data
2.1. Pre-emphasis of voice signal
The average power spectrum of the speech signal will be affected by glottal excitation and the speaker's mouth and nasal cavity radiation, and its high-frequency part will drop P1 by about 6dB / octave above 800Hz. Therefore, the frequency spectrum of the high frequency part is weaker than that of the low
frequency part. When seeking the frequency spectrum of the voice signal, the voice signal needs to be pre-emphasized. The purpose is to eliminate the influence of the vocal cords and lip radiation during the vocalization process, to supplement the inherent decline of the voice power spectrum and the pronunciation system. The suppressed high-frequency part increases the high-frequency part, so that in the entire frequency band from low frequency to high frequency, the same signal-to-noise ratio can be used to find the spectrum, which is convenient for spectrum analysis [2].

After the speech is pre-emphasized, the difference in amplitude between the main utterance segment and the part that may be noise becomes more obvious, and the waveform change is also clearer, thereby reducing the impact of noise on subsequent endpoint detection and feature parameter extraction modules.

2.2. Framing and windowing of voice signals

The speech signal is a typical non-stationary signal due to the movement of a person's own pronunciation organ, but compared to the speed of sound wave vibration, the movement of the pronunciation organ is very slow. Therefore, technicians usually think that the length is 10ms ~ 30ms In the period of time, that is, about 33 to 100 frames per second, the voice signal is a steady signal A. Although the method of continuous division can be used for frame division, the method of overlapping segments is generally adopted. This is to make the frame-to-frame smooth transition and maintain its continuity. Frame shift is the overlapping part of the previous frame and the next frame, and its length is usually 0 ~ 1/2 of the frame length.

The process of weighting the speech signal with a movable finite-length window to obtain a new signal waveform is called windowing, that is, the windowed waveform is obtained by multiplying the original speech waveform signal s (n) by a specific window function w (n) Signal Sw (n). The criterion for selecting the window function in the time domain is: because it is the voice waveform signal multiplied by the window function, so to reduce the truncation effect on both sides of the windowed signal frame waveform, the edge is smoothly transitioned to zero without sharp changes [3]. Therefore, compared with the rectangular window, the Hamming window is suitable for the speech framed windowing of this model.

3. Endpoint detection method based on short-term energy and zero-crossing rate

This article combines two basic methods of endpoint detection to divide a pre-processed recording into independent words as much as possible. Ideally, in terms of short-term energy: the short-term energy of the silent segment is zero, and the short-term energy of the unvoiced segment is smaller than the voiced segment; from the short-term zero-crossing rate, the zero-crossing rate of the silent segment is zero, and the unvoiced segment's zero-crossing rate is greater than the voiced segment. Therefore, if a certain part of the speech has a small short-term energy and a low zero-crossing rate, it can be regarded as a silent segment; if its short-term energy is large but the zero-crossing rate is small, it can be regarded as a voiced segment; if its short-term energy is small but the If the zero rate is large, it can be regarded as an unvoiced segment.

Figure 1. Preprocessing of voice signal data
4. Feature parameter extraction

4.1. Extraction of MFCC

In the process of learners imitating standard speech, learners are required to resemble the standard speech in vocabulary and tone as much as possible, that is, try to imitate the pronunciation habits and utterances of native speakers as much as possible. Frequency Cepstral Coefficients (MFCC) can express the dynamic characteristics of vocal tract movement and can better simulate the auditory characteristics of the human ear, and the anti-noise ability is also high, so the model in this paper uses MFCCs as the characteristic parameters for calculating the similarity of speech [4].

The process of extracting MFCCs: first, pre-process the speech to be tested, and transform the speech from the time-domain waveform to the frequency-domain graph by fast Fourier transform (FFT) in each frame, according to the auditory characteristics of the human ear, through the Mel filter The generator group obtains part of the frequency characteristics of the frame of speech, and then through discrete cosine transform (DCT), the MFCC can be obtained. In order to more accurately represent the characteristics of speech, this paper also extracts the first-order difference coefficients of the MFCCs of the speech as its dynamic change characterization, thereby calculating the 24-dimensional feature parameters of the speech based on the MFCCs.

4.2. Extract speech intensity

When learners imitate standard speech, they will imagine the context and emotions of the speaker. The level of the voice can often express whether the speaker's emotions are happy or sad, excited or flat. Therefore, the change of the learner's speech intensity can reflect the quality of its imitated pronunciation to a certain extent. In this paper, the first-order difference coefficient of the short-term energy of each frame of speech is extracted as the characterization of its dynamic change, so that the 1-dimensional feature is obtained based on the short-term energy calculation. So far, the model feature parameter extraction process of this article ends, and the learner voice and standard voice use the 25-dimensional features (24-dimensional MFCCs and their dynamic change features and 1-dimensional short-term energy dynamic change features) obtained by the above method of extracting feature parameters to calculate the similarity.

5. Voice similarity calculation

Quasi-speech and learner's speech often have different lengths after endpoint detection and feature extraction, so you can't directly use the cosine distance or Euclidean distance to measure the similarity of two speech segments. By consulting relevant literature, this model uses EMD and conditional The maximum matching dynamic programming algorithm to solve the above problems [5].

5.1. Conditional maximum matching dynamic programming algorithm

In isolated word speech recognition, the most simple and effective method is to use DTW (DynamicTimeWrapping, dynamic time bending) algorithm, which is based on the idea of dynamic programming and solves the problem of template matching with different pronunciation lengths in speech recognition. An earlier and more classic algorithm. This paper draws on the idea of DTW and combined with the actual situation of English pronunciation training model, using conditional maximum matching dynamic programming algorithm to solve the feature parameter matching and similarity calculation problem of different speech lengths. Before the English pronunciation training model extracts speech feature parameters, an endpoint detection algorithm is used to find the start and end of each word. Suppose that the standard speech template is represented by \{S(1), S(2),..., S(n),..., S(N)\}, n is the timing label of the standard speech frame, n = 1 is the starting speech frame, n = N is the end speech frame, S(n) is the speech feature vector of the nth frame; the learner speech template uses \{T(1), T(2),..., T(m),..., T(M)\} To indicate, m is the timing label of the learner's speech frame, m = 1 is the starting speech frame, m = M is the ending speech frame, and T(m) is the speech feature vector of the mth frame. Standard speech and learner speech templates use the same type of feature vectors.
(12-dimensional MFCCs, 12-dimensional MFCCs dynamic change features, 1-dimensional short-term energy dynamic change features), the same frame length (256 sampling points), the same Window function (Hamming window) and the same frame shift (80 sampling points). Assuming that standard speech and learner speech templates are denoted by S and T respectively, in order to compare their similarity, the matching value \( D[S, T] \) between them is calculated in the model. The larger the matching value, the higher the similarity. To calculate this matching value, first calculate the matching value between the corresponding frames in S and T, that is, \( d[S(n)] [T(m)] \), where \( n \) and \( m \) are arbitrarily selected frames in S and T, respectively number. Since \( N \) and \( M \) are often unequal in the actual two-segment speech, this paper uses conditional maximum matching dynamic programming algorithm to find the maximum matching degree, and makes the two frames in the speech aligned without matching items. If the frame numbers \( n = 1 \sim N \) of the standard template are marked on the horizontal axis in a two-dimensional rectangular coordinate system, the frame numbers \( m = 1 \sim M \) of the template to be tested are marked on the vertical axis. Integer coordinates representing the frame number draw some vertical and horizontal lines to form a network, and each intersection (\( n, m \)) in the grid represents the intersection of two frames in the test mode [6]. The conditional maximum matching dynamic programming algorithm can be summarized as finding a path through several grid points in the grid. The grid point through which the path passes is the frame number of the distance calculation in the standard speech and learner speech templates. The path is not arbitrarily chosen. Although the pronunciation speed of any kind of speech may change, the order of its parts cannot be changed. According to the above state transition equation, it is assumed that the path has passed the grid point (\( n_{i-1}, m_{j-1} \)) Then the next passing grid (\( n_i, m_j \)) may be the following three cases:

\[
(n_i, m_j) = (n_{i+1}, m_{j+1})
\]

\[
(n_i, m_j) = (n_{i+1}, m_{j+2})
\]

\[
(n_i, m_j) = (n_{i+2}, m_{j+1})
\]

This allows each frame number to be mapped at most once, and the cumulative distance along the path that can be found reaches a maximum. It is easy to prove that only one search path can pass through any grid point (\( n_i, m_j \)) within the limited range.

6. Experimental results

By calculating the correlation between the two sets of data and the expert's score, you can compare the pros and cons of the two methods of calculating the similarity of speech. Calculate 20 learner voices using MATLAB to obtain the pronunciation quality rating and expert score obtained by the model: the correlation between the rating based on the EMD algorithm and the scores of expert 1 and expert 2 are 0.5474 and 0.6715, respectively; the rating based on the dynamic programming algorithm and expert 1 The correlations with Expert 2 scores were 0.2064 and 0.2405, respectively. However, in terms of time complexity, the average time based on the EMD algorithm is 16.729s longer, while the average time based on the dynamic programming algorithm is 1.191s.

7. Conclusion

The scores obtained by the two experts in the experiment are highly correlated, indicating that the scores are of reference value. Because the rating scale given by the model is 1 highest and 4 lowest; the expert score is the highest 10 and the lowest 0, so The higher the correlation between model score and expert score is, the closer the absolute value of correlation coefficient is to 1, from the experimental results it can be seen that the rating result of the model's calculation of similarity under the EMD algorithm is more relevant to the expert's rating result, and it can also more realistically feedback the learner's pronunciation level when imitating standard speech; at the same time, the model based on the EMD algorithm takes more time , And may affect the user experience. Considering the needs of learners and the purpose of English pronunciation training, EMD algorithm is more suitable for calculating the similarity of English pronunciation in CAPT system.
References
[1] Li Shaomei, Chen Hongchang, Wang Kai. Design and implementation of DSP-based high-speed real-time speech recognition system. Modern Electronic Technology, 2007, 30 (15): 109-111.
[2] Wu Tianying, Dai Zaiping. Computer speech signal synthesis based on the inverse operation of Burg algorithm. Fujian Computer, 2005 (9): 84-85.
[3] Ma Li, Dang Youyun. Simulation and Analysis of Speech Recognition System for Isolated Words of Specific Persons. Journal of Xi'an University Engineering Science and Technology, 2007, 21 (3): 371-373.
[4] Zhao Bo, Tan Xiaohong. Application of Speech Recognition Technology in English Pronunciation Error Correction [J]. Computer Applications, 2009, 29 (3): 761-763.
[5] Chen Yongquan, Li Jianzhong, Zheng Rongwen, et al. Analysis of isolated word speech recognition technology and its application prospects [J]. Communication World, 2015, 000 (018): 250-251,252.