OPEN ACCESS
Computer Science & IT Research Journal
Volume 1, Issue 1, P.29-36, March, 2020
Fair East Publishers
Journal Homepage: www.fepbl.com/index.php/csitrj

FEP

COMPUTER BASED ENGLISH SPEAKING TEST BASED ON ARTIFICAL NEURAL NETWORK

Yu Min[1], Chao Li[2], Xin Wang[3]
[1,2,3] Department of Computer Science and Technology, Nanjing University, China

*Corresponding Author: Yu Min

Article Received: 04-10-19 **Accepted:** 22-02-20 **Published:** 30-03-20

ABSTRACT

English testing is a most common test conducted around the world for evaluating an individual's English capabilities in mostly reading, writing, speaking, and listening domain. With increased cost and higher subjective assessment attached in some tests, there is required to change the test from traditional method to computer based. In this study, a proposed method for conducting speaking test for English based on objective assessment method. The proposed system is able to identify different dialects based on unit analysis of syllable along with phonetic errors. The proposed system is based on pronunciation parameters and neural network for evaluation purpose. The PSO algorithm is used for training the artificial neural network. The experiment result conducted for validating the proposed system shows promising performance.

Keywords: English Test, Automated Test, Computer Based, Neural Network.

INTRODUCTION

English is a standard language around the world and important communication tools between various diverse groups. The popularity of English is increased severalfold after the increased globalization and resulting trade. The use of English is varying and range from education, research, and business to governance and so on. There are various tests which are used for assessment of a candidate's English proficiency. These test score significantly shape an individual's education and employment chances in some particular situations. For example, in many English-speaking countries, admission in higher education for foreigners is subject to a suitable test score. TOEFL and IELTS are the two main internationally recognized English language tests. The IELTS is more popular in countries including UK, Australia, New Zealand and Canada; while, for USA, mostly, TOEFL is famous. In this study, we focus on the IELTS since it has larger coverage. The IELTS consist of four parts including the reading, writing, listening, and speaking. The speaking test is face to face where an examiner tests the speaking power of an individual. However, face to face testing for this test is that due to the difficult dialects by people from different countries, the examiners may provide unequal score to the participants. The other challenges are arrangements of physical and human resource to conduct the test. Logistic problems for the candidates and the examiner are also a challenge. Because of different dialects, maintenance of same standard in testing and marking is also a greater challenge in the face to face test. The higher cost of administering the test and fee to the candidates is also a challenge. Therefore, an artificial test is a suitable option since it can overcome the challenges of the face to face test. With computer-based test, the test can be conducted in large population without incurring much additional cost. Furthermore, the computer-based test can also improve the quality of the testing and remove subjective assessments. The computer-based test can be developed based on the data mining technology through which we can obtain the different pronunciations defects attributes.

Subjective and objective assessment are two main methods of pronunciations assessments. The subjective assessment involves assessment made by an examiner based on predefined rules and award of grades accordingly. It reflects the evaluator's assessment of speech quality of the candidate. The main methods of subjective assessments include Diagnostic Acceptability Measure, Degradation Mean Opinion Score, Diagnostic Rhyme Test, and Mean Opinion Score. The Mean Opinion Score or MOS is the most frequently used method which test the speech quality based on the average opinion score of the examiner. The candidate speech quality is assessed based on set scale such as 5-point scale ranging from excellent, good, average, poor, and bad having score range from 5 to 1 respectively.

There are certain advantages of using the subjective assessment for testing the pronunciation such as easier and simple to understand. Because it is conducted in person, so its results can be considered highly reliable since they truly judge the speaking and communication power of a candidate. However, there are certain limitations. For example, one examiner may have subjective bias which cannot be reduced until the test is conducted by a panel. Furthermore, it is also resource heavy since it is inflexible, time consuming, and cost a lot of time and financial resources.

The alternative to the subjective assessment is the objective assessment which can be done with the help of computer. The speech assessment conducted in the objective assessment overcome some of the difficulties related to the subjective assessment. The objective assessment of speech moves from time-domain analysis to frequency domain analysis and from frequency domain analysis to the perception analysis. Under the frequency domain analysis, the perceptual domain analysis is mostly converting in to speech signals to internal acoustic feature which reflects the psychological characteristics. Up to some extent, the transformation in the objective assessment is resembles with the psycho-acoustic characteristics of human and the speech processing in peripheral auditory system and cochlea of human. The objective assessment method relies on feature parameters of speech perceptual analysis. Two main categories of objective assessment include non-reference and reference methods. The reference methods evaluate the quality of speech based on distortion between reference speech signal and the output speech signals. Thus, we can say that in reference-based method, there is a reference against which the data input is compared. On the other hand, the non-reference method, only use output speech signals for evaluating the quality of the speech. Objective speech quality metric is thus a type of feature space. Different types of metrics include Bark Spectral Distortion, Mel Frequency Cepstrum Distance, and Linear Prediction Cepstrum Coefficients Distance (Kader & Deb, 2012).

The idea of Linear Prediction Cepstrum Coefficients Distance is to make use of a linear combination of the previous sample speech signals to an approximate speech signal sampling (Gray & Markel, 1976). Predictive parameter group is used for optimization of difference between liner predictive under some criterion and the speech sample. Thus, it can be argued that LPCCD is a speech distortion metric. However, due to the not considering of human ear characteristics in LPCCD, it can be classed as a subjective method of assessment.

Human auditory sensitivity is different based on changing frequencies of sound waves. There is logarithmic relationship between acoustic frequencies and the human auditory which has making effects. The frequency of acoustic signal is divided non-uniformly and some metrics are proposed accordingly.

Mel-CD is proposed by Kublchek which is based on cepstral coefficients in the distortion measure and the non-linear frequency characteristics of human auditory perception. According to this metric, for Mel-Cepstral, the frequency axis is transformed into cepstral coefficients which are obtained from cepstral domain. For evaluating the distortion metrics, weighted sum of square is used. In speech recognitions system, the Mel-CD is widely used now because of its higher reliability and similarity to the human auditory characteristics.

Another subjective assessment method is the Bark Spetral Distortion which is based on human auditory perception. BSD works on the basis of constructing transformation models which are similar to the human perception mechanism related to the identification of speech signals (Yang, Benbouchta, & Yantorno, 1998). The speech spectrum is converted in to auditory perception spectrum in the BSD in the 20 HZ-16 KHZ audible region. Different center frequencies are constructed which makes the Barker frequency groups. The BSD metric is based on the Barker spectrum Euclidean distance between the speech signal and the original signal. A wide range of human auditory can be simulated in the BSD metric.

The Mel-CD and BSD are two common spectrum distortion assessment methods. The human auditory characteristics are taken into account while processing the speech signals in these two methods. The frequencies of the speech signals are divided by non-uniform hence the results of the assessment made based on these two methods is more likely to be classed as subjective assessment.

Subjective Grading for English Pronunciation

The English-speaking test can be done based on the pronunciation of syllables in the English language which are made by 26 characteristics in total. Some major errors in English pronunciation include different errors. For example, the related compounds vowels defects mean not allowed to place the tongue and not enough movement. The tone errors refer to making use of wrong tone volume (Kader & Deb, 2012). The consonant defect means oral part of the defect is not correct. Phonetic error means reading another syllable instead of the right syllable. The syllable defect is when the error in the consonant vowel or tone. Thus, there can be different type of defects which can affect the score of a candidate in the speech testing (Dash & Nayak, 2013).

Syllable is the smallest unit in English pronunciation and basic unit of speech structure. A syllable is analyzed into vowels and consonants in the traditional phonology. Consonant refers to the beginning of a syllable. These consonant letters are used for indicating the sound. The pronunciation length of consonant is relatively stable and is becomes very handy in the objective assessment. Vowel refers to the part of the back consonant in a syllable. The assessment of the English syllable pronunciation accuracy is based on two features. First is the composition of a syllable phoneme which can be evaluated based on acoustic channel information. Second is the tone of a syllable which is included in the information of acoustic source. A pitch curve of a syllable contains the English tone information.

The process of objective grading of English starts with the input speech where consonants and vowels segmentation are separated and assessed which leads to the binominal fitting. Finally, the results of the objective assessment are converted into the subjective assessment and result is displayed. Thus, it can be seen that syllable information in these two aspects are separate from each other and hence have separate processing.

A Framework of Automatic Pronunciation Grading System

Dr. Kennedy and Dr. Eberhart developed the Particle swarm optimization (PSO) which is considered as a population based stochastic optimization. The PSO has several characteristics which resembles the modern computation techniques such as Genetic Algorithms (GA). However, the PSO has edge over the GA since it has shorter computation time and requires less parameters definition. The successful application of PSO in different fields such as fuzzy system control, artificial neutral network training, and function optimization shows that it is a superior technique compared to many other contemporary techniques available (Kader & Deb, 2012).

The PSO function by initiating a group of random particles (solutions) and start searching for optima by updating generations. Every iteration involves updating of two particles by following two ‘best’ values. Pbest is the outcome which is the first best solution the iteration process has achieved so far. Gbest refers to the global best and is the second-best value tracked by the particle swarm optimizer. In situation, where a particle takes part of the population as its topological neighbors, it is referred as Ibest (Dash & Nayak, 2013).

The flying experience of own and companion is used for adjustment made by the particles in PSO. The particle is treated as a point in a D-dimensional space. The ith particle is represented as

Xi= (X_{i1}, X_{i2},…., X_{iD}). For any ith particle, the best previous position is recorded as PI= (P_{i1}, P_{i2}, … P_{iD}). Symbol g is used for representing the index of the best particle among all the particles in population. The rate of position change or velocity for particle is represented as VI= (vi1, vi2, …viD). Particles can also be manipulated based on a modification equation. The new equation can be used for calculating the particle's based on its past velocity and distance from the best experience and the group's best experience. The particles this way can be flied towards new position. A predefined fitness function is used for measuring the performance of each particle. To control the influence of previous history of velocities on the current velocity, the inertia weight w is used. Thus, it controls the trade-off between local and global exploration abilities of the flying points. Generally, a smaller inertia weight supports the local exploration; whereas, a larger inertia weight supports the global exploration. Thus, if a suitable inertia weight is selected, it can provide a good balance between the global and local exploration by objects.

Artificial Neutral Network

Artificial neural network is that mathematical model which attempts to mimic the structure or function of a biological neural networks. It utilizes the connectionist approach for computation and consists of interconnected group of artificial neurons (Beale, Demuth, & Hagan, 1996). The artificial neural network is flexible and change its structure based on changing information retrieved from external or internal environment. The application of artificial neural network is establishing pattern in data or large data modelling.

The Framework of the System

A framework for computerized grading of English-speaking test is as follows.

The framework starts with sample speech and test speech which are used for feature extraction by the system. The extraction goes to the speech parameter database which leads to the development of PSO-based neural network model. The model matching is used for calculating the performance of the candidates and preparing the score. The steps are as follows;

In first step, the pronunciation parameters are established which are based on making comparison of standard pronunciation and the pronunciation characteristics of the examinee. Generally, development of a database related to the pronunciation characteristics is the prime importance here.

The next step is to develop a neural network model for evaluation purpose. The objective prediction model is developed using the neural network model by calculating the difference between the sample speech and the test speech.

In next step, the examinee's pronunciation data is collected which is divided into smaller units for result calculation purpose. The small unit can be a phoneme which is compared against the standard phoneme based on speech recognition principle. The result is prepared accordingly.

OBJECTIVE EVALUATION MODEL BASED ON PSO-ANN

Feature Parameter Extraction

For measuring the pronunciation quality objectively, the feature parameter extraction is the most important step. In human, the process is based on psychoacoustics knowledge. Me1 is a measurement unit for measuring the pitch in psychoacoustics for describing the subjective sensation of human ear to sound frequency. Me1 cepstrum distortion measure is a bending frequency spectrum. The Me1 associates with spectrum synthesis feature and frequency analysis of human ear in the event of hearing complex sound. In speech signal processing, the Me1 Cepstrum is a famous tool.

The Me1 extraction is based on pre-emphasis which leads to windowed interpolation leading to FFT followed by frequency bending and filtering, followed by Log and finally the DCT (Muda, Begam, & Elamvazauthi, 2010). Cepstrum analysis is used for developing the Me1 Cepstrum coefficient speech analysis. Followed by image deconvolution, the linear frequency scale is compared against the Me1 frequency scale filtering through group of triangle bandpass filter. Followed by frequency bend processing, power spectrum is passed through Me1 measure triangle bandpass filter and obtains energy weighted sums. For MFCC, the test pronunciation and sample pronunciation, the distortion of each test pronunciation is calculated and obtained the square sum of distortion measure. Obtained Me1 cepstrum coefficient is the distorted distance of each frame. For each frame, the arithmetic mean of Me1 cepstrum coefficient distortion distance is calculated which is the Me1 cepsturm coefficient distortion of the test pronunciation. For objective test speech quality evaluation, two features parameter are the input to the neural network. One is the Me1 cepstrum feature parameter and the Me1 cepstrum difference feature parameter.

Objective Evaluation Model based on Artificial Network

The neural network can be used for evaluating the speech quality objectively. For this, the input vector of the automatic pronunciation system is [x_{n1}, x_{n2}, …., x_{nm}] which refers to the output speech signal's m-dimensional Mel cepstrum feature parameter of the nth frame. The normalized output is obtained using the neural network. Through the linear relations, the negative correlation between the outlet of the neural network and the subjective score is transformed in to objective evaluation with value range from 0 to 100. The two stages of establishing the neural network for objective evaluation of the test include the learning and the network scoring stage. Sufficient sample is developed in the learning stage. Once the test pronunciation signal pass through the neural network, the network is trained and have certain generalization performance match feature using the extrapolation and interpolation.

Euclidean distance is used for measuring the distortion measure between MFCC and spectrum. A weighted coefficient is obtained for improving the anastomosis between Euclidean distance and subjective sensation judgment. The output is similar to the real hearing characteristics to a certain extent.

PSO Algorithm Training Artificial Neutral Network

The three main qualities of neural network are optimized using advanced computing methodologies. These three qualities are network architecture, network connection weights, and network learning algorithms. For training artificial neural network, PSO is used since it is considered a more successful option. Group of weights are particles which are encoded (Ghomsheh, Shoorehdeli, & Teshnehlab, 2007). All pattern to the network whose weight is calculated using the particle is compared against the standard output for overcoming the classification problem. Next, PSO is applied for training the artificial neural network which can have the lowest possible number of misclassified patterns.

EXPERIMENT RESULTS

Experiment was conducted to test the performance of the objective assessment system which was measured by correlation between the ideal value and the objective evaluation results. A pearson coefficient can be used to represent the correlation between subjective and objective assessment results. The output is obtained comparing the training samples and test samples.

Table 1: Experiment Results

Sample	The ANN Method	The Ideal Output	PSO-based ANN Method
1	0.756	0.812	0.893
2	0.756	0.819	0.791
3	0.654	0.745	0.731
4	0.545	0.593	0.645
5	0.659	0.711	0.739
6	0.788	0.812	0.843
7	0.794	0.805	0.832
8	0.743	0.872	0.810

The comparison of experiment results based on 8 samples indicate that difference of results between the traditional ANN, PSO based method, and the ideal output is small.

CONCLUSION

The paper presents a PSO-based ANN algorithm for automatically grading the English-speaking test as part of the IELTS. The proposed method utilizes the PSO algorithm connection weights of the ANN topology.

References

Beale, H. D., Demuth, H. B., & Hagan, M. T. (1996). Neural network design. *Pws, Boston*.

Dash, T., & Nayak, T. (2013). English Character Recognition using Artificial Neural Network. *arXiv preprint arXiv:1306.4621*.

Ghomsheh, V. S., Shoorehdeli, M. A., & Teshnehlab, M. (2007, June). Training ANFIS structure with modified PSO algorithm. In *2007 Mediterranean Conference on Control & Automation* (pp. 1-6). IEEE.

Gray, A., & Markel, J. (1976). Distance measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *24*(5), 380-391.

Kader, M. F., & Deb, K. (2012). Neural network-based English Alphanumeric character recognition. *International Journal of Computer Science, Engineering and Applications*, *2*(4), 1.

Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *arXiv preprint arXiv:1003.4083*.

Yang, W., Benbouchta, M., & Yantorno, R. (1998, May). Performance of the modified bark spectral distortion as an objective speech quality measure. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)* (Vol. 1, pp. 541-544). IEEE.