Automatic Pronunciation Assessment for Mandarin Chinese

Jiang-Chun Chen, Jyh-Shing Roger Jang, Jun-Yi Li and Ming-Chun Wu

Multimedia Information Retrieval Lab
CS Department, National Tsing Hua University, Taiwan
{jtchen, jang, owen, erison}@cs.nthu.edu.tw

Abstract

This paper describes the algorithms used in a prototypical software system for automatic pronunciation assessment of Mandarin Chinese. The system uses Viterbi decoding to isolate each syllable and find the log probability of a given utterance based on HMM (Hidden Markov models). The isolated syllables are then sent to a GMM (Gaussian mixture model) for tone recognition. Based on the log probability and the result from tone recognition, a parametric scoring function using a neural network is constructed to approximate the scoring results from human experts. The experimental results demonstrate the system can consistently gives scores that are close to those from human's subjective evaluation.

1. Introduction

Computer-aided language learning (CALL) involves testing procedures for both receptive and productive skills of a given subject. To evaluate receptive skills such as reading and listening, the procedure is relative simple since the score is usually based on exams of single or multiple choices. On the other hand, to evaluate productive skills such as speaking and writing is relatively difficult and time consuming since a human expert is usually required for evaluating the input signals (from speech or writing) in a highly subjective and time-consuming manner.

With the fast-growing speed of personal computers and the advances in speech processing and recognition technologies, it is now become possible to automatically assess a person's pronunciation via computer software. This paper proposed such algorithms for assessing pronunciation in Mandarin Chinese based on Viterbi decoding, tone recognition, and nonlinear regression via neural networks. In particular, we divide pronunciation assessment into three parts:

1. Similarity in speech contents
2. Similarity in speech intonation
3. Score tuning based on nonlinear regression

The assessment of speech contents is achieved by computing the normalized HMM log probability of the given utterance. The assessment of speech intonation is based on GMM-based tone recognition, where the input is the isolated syllables obtained via Viterbi decoding. We then combine these two results into a single score between 0 and 100 via a parametric scoring function that can be tuned to approximate the scores from human experts. The scoring function can be implemented either as a simple linear function or as an advanced nonlinear model such as a neural network. The experimental results demonstrate the feasibility of using a nonlinear scoring function optimized by downhill Simplex method for pronunciation assessment.

The rest of this paper is organized as follows. Section 2 gives a quick review of related previous work on automatic pronunciation assessment. Section 3 explains the speech-related techniques used in our approach, including Viterbi decoding and tone recognition. The method of combining weighted scores based on neural networks is also explained. Section 4 demonstrates the experimental results. Sections 5 gives concluding remarks and future work.

2. Related work

Recently, research about pronunciation assessment has been investigated a lot in the area of CALL (computer assisted language learning) [8][2][1][3] and successful applications have been reported. In general, pronunciation assessment within CALL requires the computer to evaluate the pronunciation quality using various speech features and derives a parametric scoring function that can minimize the discrepancies between the scores from computers and those from human experts. However, Mandarin Chinese is a tonal language and each character has a tone (out of five possibility candidates) associated

0-7803-8603-5/04/$20.00 ©2004 IEEE

with it. Moreover, the tone of a given character is also context-dependent to some extent. Hence the correct pronunciation of the tone of each character in a sentence is the most challenging problem to a non-native speaker who tries to speak Mandarin Chinese. The proposed system takes this specific problem into consideration and tries to create a comprehensive Mandarin Chinese pronunciation assessment system. Related research on similar systems is little reported in the literature previously.

3. The proposed approach

The proposed pronunciation assessment system uses two speech processing techniques to extract related features in a given utterance, that is, Viterbi decoding using HMM (Hidden Markov models) and tone classification using GMM (Gaussian mixture model), as explained next. The two scores are then combined through a scoring function that aims to mimic the scoring mechanisms of human experts.

3.1. Viterbi Decoding Using HMM

HMM (Hidden Markov models) has been used for speech recognition with satisfactory results for the past few decades [10][5]. In our system, the HMM training is based on a balanced corpus of Mandarin Chinese recorded by 70 subjects in Taiwan. Each speech feature vector contains 39 dimensions, including 12 MFCC (Mel-frequency cepstral coefficients) and 1 log energy, and their delta and double delta values. After the training, the parameters of 526 RCD (right context dependent) models were extracted using HTK (Hidden Markov Model Toolkit) [4]. We have also implemented an efficient speaker-independent continuous-word speech recognizer [6] based on the extracted parameters and tree lexicon.

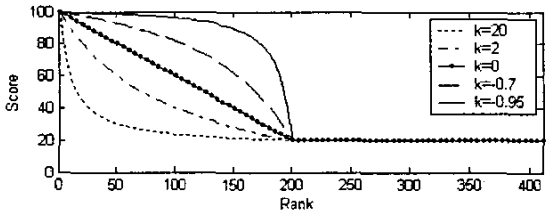
For pronunciation assessment, full-functional speech recognition is not required for our system. Instead, we need to build a linear net structure that consists of the models of the utterance text. Then Viterbi decoding is used to do force alignment between the speech frames and the models in the net. The final results include frame indices of each isolated syllable and the log probability of the given utterance. The log probability is an absolute measure of how the utterance is close to the acoustic models identified from the speech corpus. Consequently the log probability varies a lot and cannot be used for pronunciation assessment directly.

Instead of using the log probability directly, we use a relative measure obtained from each syllable segment. In other words, we align each syllable segment against all 411 syllables to obtain 411 log probabilities. After sorting these probabilities based on descending order, the position of the correct syllable is then used for scoring, as follows

1. If the rank is more than 200, then the score is 20.
2. If the rank is less of equal to 200, then the score is computed according to the following formula:

$$s = \frac{1 - \frac{r-1}{199}}{1 + k \cdot \frac{r-1}{199}} \cdot 80 + 20$$

In the above formula, r is the rank of a syllable segment, s is the score, and k is a parameter that controls the curvature of the scoring function. When r is 1, the formula gives a perfect score of 100. The over all scoring function can be plotted as the following curves for various values of k:



The actual value of k will be determined based on downhill simplex search, to be detailed later.

Therefore the overall score of an utterance with m characters can be expressed as a weighted average of all syllable segments' scores, as follows:

$$Score_{mfcc} = \sum_{i=1}^{m} \frac{d_i^p}{\sum_{j=1}^{m} d_j^p} s_i$$

where d_i is the duration of the i-th syllable segment and the weighting factor $\frac{d_i^p}{\sum_{j=1}^{m} d_j^p}$ is parameterized by p. In other words, the longer a syllable segment is, the larger the weighting factor is for the corresponding score. (For a more clear notation definition, we shall use k_{mfcc} and p_{mfcc} to denote the free parameters k and p, respectively, discussed in this sub-section.)

3.2. Tone Recognition Using GMM

Note that the speech feature vector for computing the log probability does not contain pitch information of the

utterance. However, Mandarin Chinese is a tonal language and each character is associated with a syllable (out of 411 possible syllables) and a tone (out of 5 possible tones). The tone/pitch information plays an important part in assessing a given utterance. As a result, we need to use the isolated segment of a single syllable for tone recognition.

To evaluate the pitch of a given utterance, we use the pitch vector of each syllable segment identified via the above forced alignment based on Viterbi decoding. Each pitch vector of a syllable segment is identified via the common autocorrelation method [5], and then transformed into a Legendre polynomial of order 3. The polynomial coefficients are then used a pitch feature vector for further tone classification via GMM (Gaussian mixture model). Our baseline test indicates that when the number of Gaussian density functions for each tone is 128, we can obtain a recognition rate of 94.5% on the test set of 2400 syllable segments, whereas the training set has different 2408 syllable segments. (In tone recognition, we only consider the most common four tones. The fifth tone seldom appears and is thus not considered here.)

Once a syllable segment is classified, we can find the position of the correct tone in the ranking list. Then a similar function is used to convert the rank into the score, as follows:

$$s = \frac{1 - \frac{r-1}{3}}{1 + k \cdot \frac{r-1}{3}} \cdot 100$$

When r=1 (the correct answer appears at the top of the output ranking list), we have a perfect score of 100. On the other hand, if r=4 (the correct answer appears at the last of the ranking list), we have a score of 0. The overall score is then again computed as a weighted average:

$$Score_{tone} = \sum_{i=1}^{m} \frac{d_i^p}{\sum_{j=1}^{m} d_j^p} s_i$$

where d_i is the duration of the i-th syllable segment and the weighting factor $\frac{d_i^p}{\sum_{j=1}^{m} d_j^p}$ is parameterized by p. In other words, the longer a syllable segment is, the larger the weighting factor is for the corresponding score. (For a more clear notation definition, we shall use k_{tone} and p_{tone} to denote the free parameters k and p, respectively, discussed in this sub-section.)

3.3. Parametric Scoring Function

As mentioned in the previous two subsection, we have obtained two scores based on MFCC and tones. The overall scoring function can be designed as a weighted average of the two scores:

$$Score = w \cdot Score_{mfcc} + (1 - w) \cdot Score_{tone}$$

Apparently the overall scoring function is parameterized with several parameters, including w, k_{mfcc}, p_{mfcc}, k_{tone}, and p_{tone}. To tune these parameters to approximate the scores from human experts, we employ the downhill simplex method [7] to find the optimal values of these parameters. The experimental results are covered in the next section.

4. Experimental results

To construct the scoring function, we used a dataset containing 400 utterances from 20 speakers, 10 males and 10 females, each with various levels of proficiency in Mandarin Chinese. Each speaker is asked to utter 20 sentences chosen from the most famous 300 poems from Tang dynasty. Some of the utterances are purposely pronounced incorrectly in either contents or tones to give "low-score" examples of the training data. These utterances are evaluated by a human expert who gives a score between 0 and 100 to each utterance, according to the "correctness" subjectively determined by the human expert. We then used downhill Simplex method to fine tune the parameters w, k_{mfcc}, p_{mfcc}, k_{tone}, and p_{tone}. The resulting value of w is 0.71, indicating that the contents of the utterance are more important than the tones of characters in the utterance. This is also consistent with the phenomenon that an utterance with wrong tones (this is very common for a non-native speaker to have wrong tones) are easily recognized than an utterance with wrong contents.

To test the performance of the system, we carried out an outside test in which another set of 400 utterances recorded from 10 subjects were evaluated and given scores by the same human expert. According to the scores, each sentence is assigned a category out of three candidates: good (between 80 and 100), medium (between 60 and 80), and bad (below 60). The following table lists the test result in the form of a confusion matrix, in which each row corresponds to a category assigned by our system, and each column corresponds to a category assigned by the human expert.

Human / Machine	Unit: Number of sentences		
	Good	Medium	Bad
Good	121	44	28
Medium	6	67	10
Bad	5	7	112

Table 1: Confusion matrix in terms of three categories.

In the above table, it is obvious that our system can match the categories assigned by a human expert in a satisfactory manner. The overall recognition rate in terms of these three categories is (121+67+112)/400 = 75%. And the average of the absolute difference between scores from the computer and the human is 5.42.

5. Conclusions and future work

In this paper we have developed the algorithms to automatically evaluate the pronunciation of Mandarin Chinese. The proposed system uses several techniques from speech signal processing and recognition, including Viterbi decoding using HMM (Hidden Markov models), pitch determination using autocorrelation, and tone recognition using GMM (Gaussian mixture models). By using downhill Simplex method, we successfully derived a parametric scoring function in which the parameters are fine tuned to match the scores from human experts. The experimental results demonstrate the feasibility of the proposed system.

From this study, it is obvious that pronunciation assessment within CALL (computer assisted language learning) is becoming a practical application; it is multi-disciplinary in nature and successful application of such systems requires various techniques from different areas, including speech signal processing, speech recognition, computational linguistics, and natural language processing. Immediate future work that can extend this study includes the following items:

1. Develop fix-point version of the proposed algorithms such that the system can be ported to mobile device such as PDA or PDA-like electronic dictionaries.
2. Extend sentence-level pronunciation assessment to paragraph-level, and take other factors into consideration, including the speaker's moods, continuation, and prosodies.
3. Extend similar work to computer-assisted language learning for English. (In particular, we need to find the most common pronunciation mistakes by people in Taiwan who want to learn English.)

6. References

[1] Catia Cucchiarini, Helmer Strik, Lou Boves, "Automatic Evaluation Of Dutch Pronunciation By Using Speech ", *Proc. Eurospeech*, 1997.

[2] Chanwoo Kim and Wonong Sung, "Implementation of an intonational quality assessment system", *Proc. Inc. Conf. Spoken Language Processing*, Denver, Col, 2002.

[3] H. Franco, L. Neumeyer, V. Digalakis, O. Ronen, "Combination of machine scores for automatic grading of pronunciation quality", *Speech Communication*, vol. 30, pp.121-130, 2000.

[4] Hidden Markov Model Toolkit V3.1. Speech Vision and Robotics Group of the Cambridge University Engineering Department, 2002. (http://htk.eng.cam.ac.uk/)

[5] Huang, X., Acero A., and Hon, H. -W., Chapter 12 of *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, Upper Saddle River, New Jersey, 2001.

[6] Jang, J. -S. Roger and Lin, Shiuan-Sung, "Optimization of Viterbi Beam Search in Speech Recognition", *International Symposium on Chinese Spoken Language Processing*, Taiwan, August 2002.

[7] Jang, J. -S. Roger, Sun, C. -T. and Mizutani, E. *Neural-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence* Prentice Hall PTR, Upper Saddle River, New Jersey, 1997.

[8] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality", *Speech Communication*, vol. 30, no. 2-3, pp. 83-93, Feb.2000.

[9] Proakis, J. R. J. G. and Hansen, J. H. L. *Discrete-time processing of speech signals*, New York, Macmillan Pub. Co., 1993.

[10] Rabiner, L. and Juang, B.-W., *Fundamentals of Speech Recognition*. Prentice Hall PTR, Upper Saddle River, New Jersey, 1993.