ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

COMPUTER ASSISTED SPOKEN ENGLISH LEARNING FOR CHINESE IN TAIWAN

Jiang-Chun Chen, Jui-Lin Lo, Jyh-Shing Roger Jang

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

ABSTRACT

This paper proposes an approach to computer assisted spoken English learning for Mandarin Chinese speaking people in Taiwan. Various studies have suggested the importance of acoustic models for pronunciation assessment. For English and Chinese people, their mother tongues are different; therefore the corresponding spoken English, as well as their acoustic models, are also different due to subtle difference in pronunciation. The aim of this work is to have robust acoustic models and better phoneme segmentation in the recognition phase of the assessment. The proposed approach improves the speech recognition rate, leading to better reliability of HMM log-probability and the higher accuracy of phoneme segmentation. These two factors, in turn, contribute to the success of our pronunciation assessment system, as demonstrated in the experimental results.

1. INTRODUCTION

With the fast-growing computing power of personal computers and the advances in speech processing and recognition technologies, computer assisted language learning (CALL) has now become a useful tool to automatically assess a person's pronunciation via computer software, especially for the second language (L2) learning. With the integration of automatic speech recognition (ASR) technology, a computer-assisted pronunciation training (CAPT) system can even provide the feedback to the student and successful applications have been reported [6]. In general, a CAPT system requires the computer to evaluate the pronunciation quality using various speech features and derives a scoring function imitating human experts.

For text-dependent pronunciation assessment, we use four speech features, including magnitude, pitch contour, rhythm, and log-probability of Hidden Markov Model (HMM) [9]. A nonlinear regression method is also applied on the speech features to derive a parametric scoring function [4]. In particular, for the difference of pronunciation for native speaker and L2 learner, the design of acoustic model and the phoneme segmentation approach has been investigated and satisfied performance is achieved.

In this work, we divide the pronunciation assessment into three parts:

1. Preprocess phase: acoustic model training
2. Recognize phase: phonemes segmentation and speech features extraction.
3. Scoring phase: score tuning based on parametric nonlinear regression

The four speech features are evaluated based on the unit of phonemes, while the segments are isolated via forced alignment of Viterbi decoding. We then combine the score of feature-level, phone-level and word-level into a final score 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. In this study, we adopt a nonlinear scoring function optimized by downhill simplex method for pronunciation assessment. The experimental results demonstrate the feasibility of the proposed approach.

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 the adaptation of acoustic model and phoneme segmentation. Section 4 demonstrates the experimental results. Section 5 gives concluding remarks.

2. RELATED WORK

Recently, L2 learning has become a very popular research topic [6]. Catia et al defines the speech features for Dutch, including time segment duration, rate of speech, and log-probability of HMM [1]. For French, Franco et al combines several kinds of machine scores with linear/nonlinear regression and statistic method [2]. For tonal language, Chen et al propose a CAPT for Mandarin Chinese based on speech recognition of HMM and tone classification of Gaussian Mixture Model (GMM) [4]. Studies also show that, for L2 learning, the ASR system trained via native speaker often has lower recognition rate for non-native speaker. Several approaches have been addressed to enhance the ASR performance [5]. However, for Chinese people, the acoustic model of spoken English differs a lot from native speakers. Therefore, in the pronunciation assessment, properly speech processing and suitable acoustic models for target learner are essential. The proposed system considers this specific problem and tries to create a comprehensive English CAPT for Chinese people. Related research on spoken English learning for

Chinese speaking people is seldom reported in the literature previously.

3. THE PROPOSED APPROACH

Generally speaking, the phone-level acoustic model is an appropriate unit for acoustic model training and pronunciation assessment. Section 3.1 describes how the phone models are trained and Section 3.2 introduces the phoneme segmentation algorithm used in our system. Section 3.3 explains the extraction of speech features for each segment. Finally, Section 3.4 gives the scoring function based on nonlinear regression.

3.1. Acoustic Model Training

To train a phone-level acoustic model, a machine-readable dictionary is needed to translate the word into phonemic units. In TIMIT, there are 60 phones in its dictionary. However, for Chinese, some phones are not easily differentiated and are usually mis-pronounced in a similar way. Therefore, in this work, we use the corpus of TIMIT but the dictionary of CMU [7]..The phone set conversion from the TIMIT to CMU is listed in Table 1 and smaller 40 phones of CMU dictionary are derived. Not only the acoustic models are more consistent, a smaller phone set also increases the robustness of model training when the training data is not abundant.

Action	Original phones and new phones
Delete	BCL, DCL, GCL, EPI, KCL, PAU, PCL, TCL
Substitute	AX→AH, AX-H→AH, AXR→ER, DX→D, H#→SIL, HV→HH, IX→IH, NX→N, Q→T, UX→UW
Split	ENG→IH NG, EL→AH L, EM→AH M, EN→AH N

Table 1. The conversion rule for the new phone set. The other phones, which do not list here, are reserved in the new set of phone model.

To improve the performance in large vocabulary continuous speech recognition (LVCSR) and phoneme segmentation, a simplified silence model is also trained [8], which is the so-called short-pause model (SP).

Some studies suggest that well-trained acoustic models for native speaker should adapt to the acoustic feature of L2 learner in the recognition phase of CAPT [5]. To be able to recognize spoken English from Chinese people in Taiwan, we have also collected a speech corpus (C-TIMIT) from Chinese speaking people in Taiwan. The C-TIMIT follows the same convention of TIMIT except that the subjects are non-native speakers of Chinese people in Taiwan. The experimental result is to be detailed in Section 4.

3.2. Phoneme Segmentation

Phoneme is the basic unit in the English pronunciation. For a test utterance, the phoneme segmentation is accomplished by forced alignment of Viterbi decoding. Traditionally, a SP model is inserted between words and the short inter-silence can be aligned into SP model. However, for utterances from non-native speakers, this approach does not perform satisfactorily. In particular, for a Chinese who is learning English, an unfamiliar vocabulary usually delays the speaking rate and causes longer silent duration between words. Traditionally, the gap between words is inserted with SP, but this action could cause problems for un-fluent utterance from non-native speakers, such as the waveform shown in Figure 1.

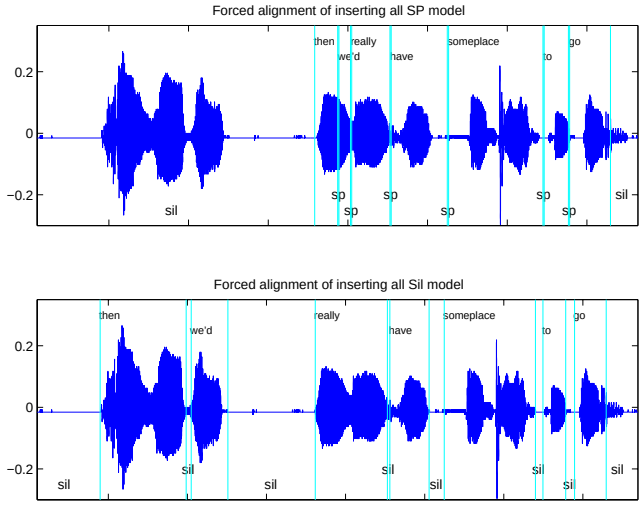

Figure 1. The forced alignment of the sentence "Then we'd really have some place to go" uttered by a Chinese. The top plot is aligned with a SP (short pause) model between words; and the bottom one is aligned with a silence model (Sil) between words. The prefix and suffix models of both cases are default to be Sil.

Basically the silence model (Sil) is a special kind of SP model with longer duration. Therefore, in Figure 1 the utterance can be segmented more correctly when all the gaps between words are filled with Sil. Note that the use of Sil model maybe aligned with longer duration that it actually is, especially for a fluent utterance. To solve this problem, a procedure called 'Dynamic Insertion' is suggested here. Assume an utterance has been forced aligned twice, one is with SP between all words, and the other is with Sil. The alignment results are called fa_{SP} and fa_{Sil} , respectively. A better insertion sequence for alignment can be determined by the next equation:

$$insert(i) = \arg\max_{model}(prob(fa_{SP}(i)), prob(fa_{Sil}(i))$$

i is the index of the gap, $prob$ is the function of time-normalized log probability in the i-th gap after forced alignment. For the case of Figure 1, a better insertion sequence may be like 'Sil SP Sil SP Sil Sil Sil Sil', but not all 'Sil' or all

'SP'. The dynamic insertion can be applied not only to the phoneme segmentation, but also to the bootstrap procedure of acoustic model training, which may lead to a better initialization of HMM training [8]. The experimental result will be shown in Section 4.

3.3. Speech Features Extraction

Like in prevous work [9], four speech features are used in our system, including magnitude (intensity or volume), pitch contour, rhythm and log-probability. Due to the differences in length, volume or pitch level, interpolation and linear shifting are applied here to normalize some of the features, as shown in Table 2.

Speech Feature	Normalization	Distance Measurement
Magnitude	Interpolation, Linear Shifting	Euclidean Distance
Pitch Contour	Interpolation, Linear Shifting	Euclidean Distance
Rhythm	None	Euclidean Distance
Log-Probability	None	Ranking Function

Table 2. The normalization method of each speech feature.

Instead of using the absolute value of log-probability directly, we define a ranking function and use a relative measure obtained from each phonemic segment [4]. In other words, we align each phonemic segment against all 40 phone models to obtain 40 log-probabilities. After sorting these probabilities based on descending order, the position of the correct syllable is then used as distance. This kind of distance measure for log-probability is commonly used in the research of utterance verification [10].

3.4. Score tuning based on parametric nonlinear regression

After measuring the distance of four features, the utterance is scored in the feature-level, phoneme-level and word-level. Finally the score is tuned by a parametric scoring function.

3.4.1 Feature-level scoring

We define the score of *i*-th feature as the function of distance:

$$score_{fea_i} = \frac{100}{1 + a_i \cdot (distance_i)^{b_i}}, \text{for } i = 1 \sim 4$$

$distance_i$ is the distance of *i*-th speech feature.

The $score_{fea_i}$ will range from 0 to 100 and is parameterized by a_i and b_i. In this part there are totally eight free parameters to be determined by nonlinear regression method later.

3.4.2 Phoneme-level scoring

Define the weighting w_i for *i*-th speech feature and combine all feature-level scores as the score of *j*-th phoneme in the equation:

$$score_{pho_j} = \sum_{i=1}^{4} w_i \cdot score_{fea_i}$$

w_1, w_2, w_3, w_4 represent the weighting of four speech features respectively and will be determined later.

3.4.3 Word-level scoring

Also, define the weighting of *j*-th phoneme as a function of time duration in a word. The score of *k*-th word is defined as:

$$score_{word_k} = \sum_{j=1}^{N} \frac{len(pho_j)}{len(word)} \cdot score_{pho_j}$$

N is the number of phoneme in a word, *len()* is a function of the time duration for each phoneme.

3.4.4 Parametric scoring function

Finally, define the weighting of *k*-th word as a function of time duration in a sentence. The overall score is defined as:

$$score_{overall} = \sum_{k=1}^{W} \frac{len(word_k)}{len(sentence)} \cdot score_{word_k}$$

W is the number of word in a sentence, *len()* is a function of the time duration for each word. The overall scoring function is composed of the feature-level scoring, phoneme-level scoring and word-level scoring functions. Apparently this function is parameterized with several parameters, including a_1, b_1, a_2, b_2, a_3, b_3, a_4, b_4, w_1, w_2, w_3, w_4. To tune these parameters to approximate the scores from human experts, we employ the downhill simplex method to find the optimal values of these parameters [3]. The experimental results are covered in the next section.

4. EXPERIMENTAL RESULTS

TIMIT is a corpus recorded by 630 persons, 438 males and 192 females, 10 sentences for each person and totally 6300 sentences. 4620 sentences are designed to be training data and 1680 files are test data. To adapt the acoustic model for Chinese, we collected the C-TIMIT corpus which is recorded by 33 persons, including 23 males and 10 females; 213 sentences are uttered by each person and totally 7029 sentences are recorded. We take the 4684 sentences as training data and 2345 as test data according to the original category of TIMIT. A tree net consisted of all vocubularies in TIMIT is also applied in speech recognition.

Each spectral feature vector contains 39 dimensions, including 12 MFCC (Mel-frequency cepstral coefficients) and 1 log energy, and their delta and double delta values. For parameters of HMM, we use right-context dependent bi-phone model, three states in each phone model, and four Gaussian mixtures in each state.

For simplicity of notation, the HMM trained by TIMIT is denoted as E_{HMM}; the testing sentences of TIMIT is denoted as E_{TST}; the HMM trained by both the training sentences of C-TIMIT and TIMIT is denoted as EC_{HMM}; and the testing sentences of C-TIMIT is C_{TST}. The dynamic insertion

approach is performed on the four combinations above and the result is shown in Table 3.

Testing / Insert seq.	C_{TST} / E_{HMM}	C_{TST} / EC_{HMM}	E_{TST} / E_{HMM}	E_{TST} / EC_{HMM}
All Sil	59.08%	87.30%	97.06%	93.55%
All SP	59.52%	87.74%	97.53%	94.04%
Dynamic	60.13%	88.37%	98.02%	94.44%

Table 3. The tree-net based LVCSR words recognition rate of different insertion sequence.

In Table 3, when compared with the column C_{TST}/E_{HMM}, the enhancement in column C_{TST}/EC_{HMM} is obvious, indicating the success of the use of C-TIMIT. The recognition rate falls a little down in E_{TST}/EC_{HMM} but still higher than C_{TST}/EC_{HMM}, which implies that the Chinese always try to learn the characteristics of native English but the pronunciation varies more. The dynamic insertion method also works pretty well and consistently in every case.

To construct the overall scoring function, we use a dataset containing 200 utterances from 20 speakers, 10 males and 10 females, each with various levels of proficiency in English. Each speaker is asked to utter 10 sentences chosen from the TIMIT. These utterances are evaluated by a human expert who gives a score between 1 and 100 to each utterance, according to the 'fluency' subjectively determined by the human expert. We then used downhill simplex method to fine-tune the parameters a_1, b_1, a_2, b_2, a_3, b_3, a_4, b_4, w_1, w_2, w_3, and w_4. The resulting value of w_1 is 0.07, w_2 is 0.22, w_3 is 0.17 and w_4 is 0.54, indicating that the contents and the pitch contour of the utterance are more important than the other two features in the utterance.

To verify the performance of the system, we evaluated an outside test in which another set of 200 utterances recorded from 10 subjects 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	63	20	7
Medium	11	27	14
Bad	10	20	28

Table 4: Confusion matrix in terms of three categories.

In table 4, 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 (63+27+28)/200 = 59%.

5. CONCLUSIONS

In this paper, we have proposed an English CAPT system for Chinese people in Taiwan. The improvement of using C-TIMIT (an English speech corpus recorded by students in Taiwan) for better recognition rates is obvious, especially for the combination of C_{TST}/EC_{HMM}. To deal with the characteristics of spoken English from Chinese in Taiwan, we have defined a suitable phone set for better results in forced alignment. A ranking-based distance measurement for log-probability is also applied in the feature-level scoring. Experiments demonstrate the feasibility of the proposed approach over conventional ones.

6. REFERENCES

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

[2] 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.

[3] 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.

[4] Jiang-Chun Chen, Jyh-Shing Roger Jang, Jun-Yi Li and Ming-Chun Wu, "Automatic Pronunciation Assessment for Mandarin Chinese", *Proc. Int. Conf. on Multimedia And Expo*, Taiwan, 2004

[5] V. Digalakis, L. Neumeyer, "Speaker Adapatioin Using Combined Transformation and Bayesian Methods", *IEEE Transactions on Speech and Audio Processing*, 94-300, 1996.

[6] A. Neri, C. Cucchiarini, H. Strik, "Automatic Speech Recognition for second language learning: How and why it actually works", *Proceedings of 15th ICPhS*, Barcelona, 1157-1160, 2003

[7] http://www.speech.cs.cmu.edu/sphinx/doc/phoneset_s2.html

[8] Hidden Markov Model Toolkit V3.2. Speech Vision and Robotics Group of the Cambridge University Engineering Department, 2002. (http://htk.eng.cam.ac.uk/)

[9] Jiang-Chun Chen, Jui-Lin Lo, Jyh-Shing Roger Jang, "以語音辨識與評分輔助口說英文學習", *ROCLING XVI*, Taiwan, 2004

[10] Rafid A. Sukkar and Chin-Hui Lee, "Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword based Speech Recognition", *IEEE Transactions on Speech and Audio Processing*, VOL. 4, No. 6, November 1996