Formant-Based English Vowel Assessment For Chinese in Taiwan

Jiang-Chun Chen, Wei-Tang Hsu, J.-S. Roger Jang
Department of Computer Science, National Tsing Hua University, Taiwan
{jtchen, deanton, jang}@cs.nthu.edu.tw

Ren-Yuan Lyu
Department of Electrical Engineering, Chang Gung University, Taiwan
rylyu@mail.cgu.edu.tw

Yuang-Chin Chiang
Institute of Statistics, National Tsing Hua University, Taiwan
chiang@stat.nthu.edu.tw

ABSTRACT

This paper proposes a formant-based approach for computer-assisted English vowel assessment. Various studies in formant-based speech synthesis have suggested the importance of formant coefficients; this motivates us to investigate pronunciation assessment using formant information instead of MFCC (Mel-frequency cepstral coefficients) alone. In particular, we explore the multi-stream HMM with the addition of formant information to improve the phoneme segmentation. We then propose the use of PCN (pronunciation confusion network) together with a formant-based confidence measure to improve error detection rates. Furthermore, the pros and cons of using cross-word phone model for both native speakers and L2 learners are discussed. Experimental results demonstrate the feasibility of the proposed approach for automatic vowel pronunciation assessment.

Index Terms: computer assisted pronunciation training, formant, assessment, pronunciation confusion network, speech recognition

1. INTRODUCTION

In computer assisted pronunciation training (CAPT), it is well known that the pronunciation of vowels is much more important than that of consonants. Successful CAPT applications have been reported 0, but few of them have considered the influence of formants for the pronunciation modeling of vowels.

Correct formation of oral cavity is the most important factor for generating correct vowel pronunciation. The relationship between formants and the oral cavity has been discussed in the literature [10]. In this paper, we propose a pronunciation assessment method based on HMM and formant coefficients, which is able to give reliable assessments about the articulator. Previously MFCC-based approach assesses the pronunciation based on the log-probability of acoustic models of the underlying utterances [7]. However, the information of exact formant coefficients, such as F1 and F2, are partially missed due to the feature reduction of triangular filter bank when computing MFCC. Therefore, we propose a formant-based pronunciation assessment system, which involves the following three phases:

1. Preprocessing phase: Acoustical model training using MFCC and normalized formant coefficients.
2. Recognition phase: Vowel pronunciation labeling and segmentation using PCN.
3. Confidence measure: Ranking-based formant-level assessment.

Methodologies of these three phases are described in the following sections. The rest of this paper is organized as follows. Section 2 explains various techniques used in our approach. Section 3 demonstrates the experimental results and Section 4 gives concluding remarks.

2. THE PROPOSED APPROACH

2.1. Automatic Phoneme Segmentation

is the flowchart of the proposed system. In the flowchart, PCN (pronunciation confusion network) is used to embed common error patterns for achieving better error detection. Another important block is formant-level assessment, which is responsible for computing confidence measure in the phone-level pronunciation. To achieve a reliable performance, an accurate phoneme segmentation is crucial. All the acoustic models are trained without manual labeling/segmentation. Comparable performance has been achieved with similar settings in previous work [1].

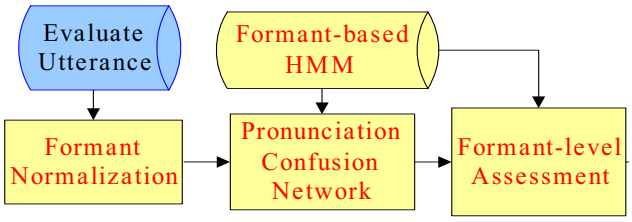

Figure 1. The system flowchart.

2.2. Formant Normalization

We use the ESPS software [3] to extract formant coefficients. Formant is highly speaker-dependent, so numerous approaches to formant normalization are proposed in the literature [2][9], but few of them considered the factor of language transfer for L2 learner. Moreover, formant coefficients depend not only on the articulator of the speaker but also on his/her native language. Hence we need to normalize the formant coefficients by considering the influence of the L1 language. Our corpus contains both Mandarin and English sentences, thus the F1 and F2 of five basic vowels ("aa", "eh", "iy", "ow" and "uw") for each speaker can be extracted first. The maximum and

minimum of F1 and F2 of five vowels for both languages can then be used in computing the normalized formant coefficients, as shown in Figure 2. For a given L2 learner, for instance, the normalized F1 of the phone model "*er*" of English can be calculated as:

$$normalize(F1_{er}) = \frac{F1_{er} - Min_{F1}}{Max_{F1} - Min_{F1}},$$

where Max_{F1}, Min_{F1}can be derived as:

$$Max_{F1} = Max(L_1 Max_{F1}, L_2 Max_{F1})$$
$$Min_{F1} = Min(L_1 Min_{F1}, L_2 Min_{F1})$$

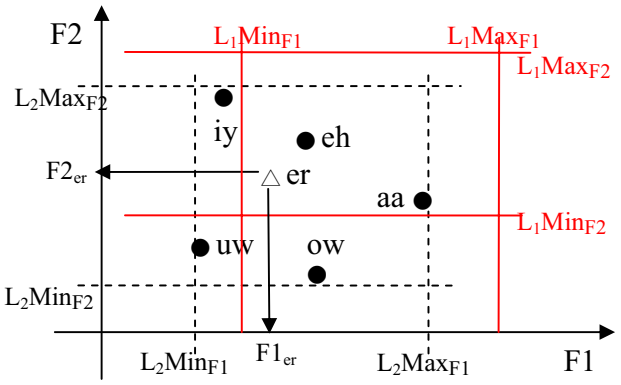

Figure 2. Speaker-dependent normalization of F1 and F2 for model "*er*".

In others words, formant coefficients can be normalized to the ragne [0, 1] for each speaker. To verify the usability of formant tracking and normalization, a HMM-based vowel classifier (comparable with the system of Schmid et al [9]) for TIMIT was designed to achieve 70% classification accuracy for 14 vowels and 4 semi-vowels.

2.3. Formant-based HMM

MFCC-based HMM is widely used in speech recognition and segmentation. However, MFCC cannot capture detailed formant information due to the feature reduction process of triangular filter bank. To improve the phoneme segmentation performance, we propose a multi-stream HMM employing both MFCC and formant coefficients as a new feature set. With a set of suitable stream weights, the new features set can effectively improve phoneme segmentation, in particular for vowels, as shown in the experimental results.

Context-dependent HMM is a typical technique in phoneme segmentation. In order to increase the robustness of speech recognition, the cross-word phone model is commonly adopted in the literature. However, our experiments demonstrate that cross-word phone models are not suitable for utterances from L2 learners since they are usually not fluent in coarticulation between words. The pros and cons of using cross-word phone models will be discussed in Section 3.

2.4. Pronunciation Confusion Network

A pronunciation confusion network (PCN) is recognition network that comprises the common pronunciation errors for L2 learner. Using Viterbi decoding on the utterance and a given PCN in the phoneme level proved to be an effective way to detect the pronunciation variation [12].

Most of the typical pronunciation errors of L2 learners come from the different phonological structure between the L2 language and native languages [10]. Therefore it is essential to consider the native language when designing a CAPT system. Table 1 lists some of the common English pronunciation errors (including both vowels and constants) for Chinese in Taiwan [4]. To detect these errors in the assessment, these patterns are embedded into a PCN. For example, Figure 3 shows the PCN of the word "husband", where the solid lines indicate the correct sequence of the pronunciation and the dotted lines provide alternative paths to detect the possible pronunciation errors by L2 learner. The "sil" nodes represent the start and the end of the utterance. To align the usually long and influent utterance of an L2 learner, a dynamic insertion approach is used here [6]. Note that in this paper, we use the pronunciation dictionary from CMU.

Type	Pair-wise Confusion Phones
Vowel Substitution	z/s, ah/aa, d/t, ih/iy, ai/aa, eh/ey/ae, uh/uw, ow/ao, aa/ao
Consonant Deletion	hh/[], r/[], ks/[]s
Consonant Substitution	th/s, th/l, th/d

Table 1. Some of the common English pronunciation errors for Chinese in Taiwan. ("[]" indicates deletion.)

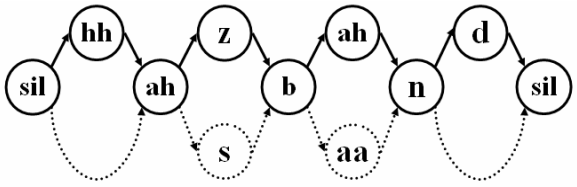

Figure 3. The PCN for the word "husband" in the phone-level.

2.5. Formant-level Assessment

The PCN approach can help us to detect typical pronunciation errors that are known in advance. However, to deal with error patterns that are not known in advance, a more general and robust error detection method is called for. To this end, we propose a ranking-based confidence measure (RCM), as explained next.

A formant-based multi-stream HMM can be used to better align the phoneme boundaries of a context-dependent triphone model (CDTM). In this study, for each CDTM, we can compute its time-normalized log-probability. Instead of using the log-probability directly, we define a ranking-based

confidence measure (RCM) that takes both the ranking as well as the gap in log-probabilities into consideration. For instance, for a given CDTM "*f-ao+r*", we need to compute the time-normalized log-probabilities of the competing CDTMs, defined as the CDTMs of the form "f-*+r", where * is a wildcard. After sorting these log-probabilities based on descending order, the confidence measure of a CDTM *c*, denoted by $Conf_C$, is defined as:

$$Conf_C = \frac{2}{1+\exp\left(\alpha \cdot \left(Rank_c - 1\right) \cdot \left(\mathrm{Prob}_{\max} - \mathrm{Prob}_c\right)\right)},$$

where $Prob_c$ is the log-probability of c, $Rank_c$ is the corresponding rank among all competing CDTMs, and $Prob_{max}$ is the max log-probability of all competing CDTMs. The constant α is set to 0.09 empirically. The value of $Conf_C$ is always between 0 and 1.

The above rank-based confidence measure is commonly used in utterance verification [11]. The threshold of the confidence measure for rejection is set to a value that minimizes the total error counts of false positive and false negative of all CDTMs. The performance evaluation will be covered in Section 3.

3. EXPERIMENTAL RESULTS

Our experiment is based on EAT (English Across Taiwan) corpus recorded by 1200 subjects [5]. This corpus contains 96000 sentences, with 80 sentences for each person, in which 13 sentences contain both Mandarin and English while the other 67 are purely English. For microphone recordings, half of the subjects are Foreign Language Department students and others are Non-English Department student, denoted as EAT_Eng and EAT_NonEng respectively. One-forth sentences are used as test data while the others are training data for both EAT_Eng and EAT_NonEng. A set of acoustic models of native speakers is also obtained using TIMIT corpus. According to the default setting in [8], we take the 3696 sentences as training data and the others as test data.

The acoustic analysis is performed at 10 ms frame rate using 20 ms hamming window. Each spectral feature vector contains 39 dimensions, including 12 MFCC and 1 log energy, and their delta and double delta values. In particular, the delta and double delta operators are also applied to the F1 and F2 formant coefficients, resulting in six formant features denoted as $Formant_6$. F3, F4 and F5 are skipped due to their instability.

For the multi-stream HMM, we use context-dependent tri-phone model, with three states in each phone model. Within each state, six Gaussian mixtures are used in $MFCC_{39}$ stream and two in $Formant_6$ stream.

To compare the accuracy of automatic phoneme segmentation, the manual transcription of TIMIT and the same 512 sentences were used as the test data, as described in [8]. The performance of correctly positioned boundaries within 20 ms is 78.3% using $MFCC_{39}$ only, which is about 5% improvement over those achieved in Brugnara et al [1]. The primary reason of improvement is due to the use of cross-word tri-phone model (as compared with bi-phone model) and 39-dimensional MFCC (as compared with 26-dimensional MFCC).

Stream Weighting for $MFCC_{39}$:$Formant_6$		HMM without Cross-Word
Vowels Acoustic Models	Non-Vowels Acoustic Models	
2 : 0 for both		78.3%
1 : 1 for both		73.7%
1.4 : 0.6	1.8 : 0.2	74.4%
1.8 : 0.2	1.99 : 0.01	79.1%
1.9 : 0.1	1.999 : 0.001	79.6%

Table 2. Percentages within 20 ms tolerance for correctly positioned boundaries of TIMIT. (The "Non-Vowels" category includes semivowels-glides, stops, nasals, fricatives and affricatives).

Table 2 lists segmentation accuracy with the introduction of formants. From the table, it is obvious that different stream weighting leads to different performance. Using equal weights (1:1) for both vowels and consonants generates the worst performance, which is reasonable since most consonants do not have a stable formant structure. The best performance of 79.6% is at the last row of Table 2, which indicates an improvement of 1.3% (over the first row of the $MFCC_{39}$-only case) when two suitable set of weights (one for vowels, the other for non-vowels) are applied to the $Formant_6$ stream.

Corpus	Cross-Word HMM	Cross-Word HMM
TIMIT	68.58%	70.25%
EAT_Eng	56.48%	55.29%
EAT_NonEng	54.86%	53.47%

Table 3. Continuous phone recognition with no language model. The feature set is $MFCC_{39}$ and $Formant_6$, with the best stream weighting obtained from Table 2.

A continuous phone recognition test without phone-level language model is performed to verify our acoustic model for EAT_Eng and EAT_NonEng, as shown in Table 3. The test set of TIMIT is the same as the 160 sentences used by Lee [8]. For each EAT_ENG and EAT_NonEng corpus, we manually labeled 400 utterances at the phone-level and used them as the test set. From the table, it is obvious that the cross-word approach increases the recognition rates by 1.6% for TIMIT (native speakers) but decreases by 1.19% and 1.39% for EAT_Eng and EAT_NonEng (representing corpra from L2 learners), respectively, indicating that the cross-word model is not suitable for EAT (both EAT_Eng and EAT_NonEng) since these non-native utterances are not fluent in nature. Moreover, the EAT_Eng category has a better recognition than EAT_NonEng, meaning that the identified acoustic models are more consistent for the students at the Foreign Language Department student than those in other departments. Note that

EAT corpus does not have phone-level manual transcription, therefore the performance may degrade due to the miss of phone boundary information during training

To determine the threshold of RCM, the training portion of TIMIT (with manual segmentation) were used to plot the receiver operating characteristic (ROC) curves, as shown in Figure 4. The circle represents the true positive (false accept) while the square represents the true negative (false reject). The equal-error-rate threshold of 0.67 is selected for our system.

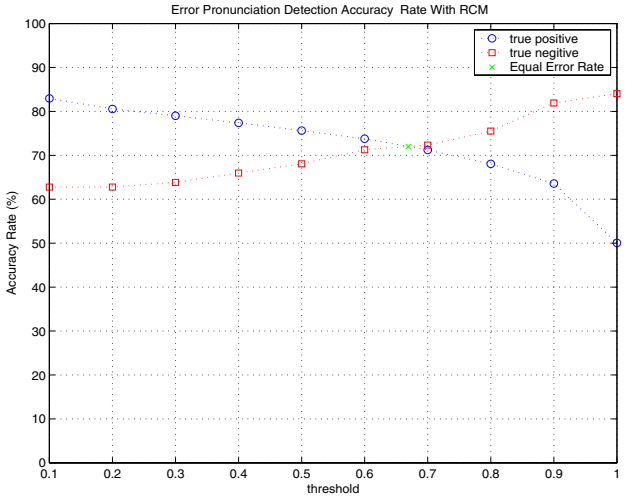

Figure 4. The ROC curve of RCM using different thresholds.

To verify the performance of the proposed RCM, in additional to the test data of TIMIT, we also asked an English expert to label 400 sentences of EAT_NonEng corpus for further test. The test was performed using the acoustic models from EAT_Eng and the RCM threshold from TIMIT. Table 4 lists the detection rates (equal to the sum of true positives and true negatives divided by all test data) of various setups. Note that the method "PCN + RCM" listed in Table 4 is a 2-pass scoring process, in which the RCM is applied after PCN. It is obvious that RCM can reduce the error rate and refine the result of PCN. Note that the threshold may not work well in Case 2 because of the mismatch between the acoustic models of TIMIT and EAT. It is believed that if we can obtain the RCM threshold from EAT (which is impossible at this stage since EAT does not have phone-level manual transcription), the performance of Case 2 will be better.

Method / Cases	PCN	PCN + RCM
1. Test data: TIMIT Acoustic model: TIMIT RCM threshold: TIMIT	85.15%	87.95%
2. Test data: EAT_NonEng Acoustic Model: EAT_Eng RCM threshold: TIMIT	79.27%	80.79%

Table 4. Error detection rates of the vowel pronunciation.

4. CONCLUSIONS

In this paper, we have proposed a CAPT system of English vowel learning for Chinese people in Taiwan. The flexibility of the proposed approach provides a more delicate way to assessing the vowel pronunciation for L2 learner. To detect unforeseen pronunciation errors, a ranking-based confidence measure (RCM) using formant information is proposed. Experimental results demonstrate the feasibility of the proposed approach.

Immediate future work of this study will focus on the accuracy of the phoneme segmentation and the robustness of multi-stream HMM. Additional acoustic features that are not embedded in MFCC will be introduced for HMM in order to achieve better performance in phoneme segmentation, particularly for utterances from L2 learners.

5. REFERENCES

[1] Brugnara, F., Falavigna, D., and Omologo, M., "Automatic Segmentation and Labeling of Speech Based on Hidden MarkovModels," Speech Communication, 12(4): 357-370, 1993.

[2] Chiba, S., New classification method of place of articulation of consonants in connected speech using formants, ICASSP '86.

[3] David Talkin and John Shore. The ESPS formant tracker. Entropic Research Laboratory, Inc., 1997

[4] http://ccms.ntu.edu.tw/~karchung/intro%20page%2029.htm

[5] http://www.aclclp.org.tw/doc/eat_brief.pdf

[6] Jiang-Chun Chen, Jui-Lin Lo, Jyh-Shing Roger Jang, "Computer Assisted Spoken English Learning for Chinese in Taiwan", ISCSLP 2004, HongKong.

[7] 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.

[8] Lee, K.-F.; Hon, H.-W., "Speaker-independent phone recognition using hidden Markov models", Acoustics, Speech, and Signal Processing, IEEE Transactions on Volume 37, Issue 11, Nov. 1989 Page(s):1641 – 1648

[9] P. Schmid, E. Barnard, "Explicit, N-Best Formant Features for Vowel Classification," ICASSP '97

[10] Peter Ladefoged, *A Course in Phonetics*, Harcourt Brace Johanovich, 2001.

[11] Rafid A. Sukkar and Chin-Hui Lee, "Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword based Speech Recognition", ICASSP 1996

[12] Strik, Helmer; Cucchiarini, Catia, Special Issue of Speech Communication on 'Modeling Pronunciation Variation for Automatic Speech Recognition', Vol. 29, No. 2-4, pp. 225-246

Yasushi Tsubota, Tatsuya Kawahara, and Masatake Dantsuji. "Computer-assisted English vowel learning system for Japanese speakers using cross language formant structures". Proc. ICSLP 2000.