A HMM-Based System for Training of Second Language Aquisition

Lingyun Gu and John G. Harris

***Abstract*— We describe a system for the training of Second Language Acquisition Pronunciation (SLAP) for non-native speakers. This speech recognition-based system is designed to mimic the valuable interactions between second-language students and a fluent teacher. When a student speaks a word into SLAP's microphone, it is analyzed to determine the part of the word (if any) that is incorrectly pronounced. A fluent utterance of the word is then played back to the student with emphasis on the mispronounced part of the word. Just as a live teacher naturally does, the difficult part of the word is played back louder, extended in time and possibly with higher pitch. We demonstrate SLAP on a multisyllabic word to show typical performance.**

***Index Terms*— Second language acquisition, HMM, objective speech assessment.**

I. Introduction

SECOND Language Acquisition (SLA) for adults can be a difficult and frustrating process. It is our experience that foreign students can master the necessary English vocabulary, listening and reading comprehension skills but their spoken English leaves much to be desired. The students have trouble pronouncing English words even though they receive frequent exposure to fluent English through audio learning tapes, lectures and radio/TV programs [8], [11]. We believe that their deficit is due to the lack of live interaction with fluent English speakers. Merely listening to a fluent speaker is not sufficient; feedback is required [9], [10]. Many times the start of the students' problems can be traced to English teachers in their home countries who have poor English pronunciation themselves and therefore usually concentrate on the reading comprehension aspects of language learning [8].

SLAP provides many of the benefits of live interaction with a fluent native teacher. The basic interaction cycle requires the student to read a displayed word aloud into a microphone. The system compares the word to a database of native utterances of the same word. If the pronunciation differs significantly from the correct pronunciation, then the correct native utterance is played back for the student to hear. Furthermore, the part of the word that was mispronounced is precisely located within the word and used to modify the native utterance so that the mispronounced component is emphasized by being louder, longer and possibly with higher pitch. The student then says the word again and the system repeats. This interaction cycle mimics observed live interactions with skilled SLA teachers [5], [10].

Automatic speech recognition (ASR) has applications in many areas. From continuous recognition for dictation systems to isolated word spotting systems for information retrieval and command and control, speech recognition is being used in more and more ways [1], [3]. One interesting application is to use ASR to detect and correct foreign speakers' pronunciation. The use of an automatic recognition system to help a user improve his/her accent and pronunciation is appealing for at least two reasons: first, it affords the user more practice time than a human teacher can provide, and second, the user is not faced with the sometimes overwhelming problem of human judgment of his production of "foreign" sounds without feedback. Speech recognition is a natural choice for this type of application [5]. Though some attempts have already been made to market systems based on speech recognition [3], there has been very little work in the literature in combining SLA with ASR.

In this paper and in the initial versions of SLAP, we chose to concentrate on learning English as a second language though the general principles can be applied to learning any language. Furthermore, our initial target students are native Chinese speakers so that the most frequent English pronunciation problems of these speakers can be identified and categorized. We expect the final system to be valuable for any non-native speaker. The demand for such a system will be high since more and more Chinese students and scholars work and study in the US or other English speaking countries. Speaking fluent English and reducing their accent as much as possible is a big motivation to communicate well with other native English speakers, especially in a tight work group, in their classes or in some international conferences. Speech utterances can be decomposed into smaller linguistic units called phonemes [3]. The English language consists of 48 phonemes, certain permutation of which produces words and phrases. In Chinese and other languages, some missing English phonemes cause the non-native speakers much trouble. Another problem occurs when two distinct phonemes exist in a foreign language while there is only one counterpart in English. The classic example of these phoneme mismatch problems is the L/R confusion that many Chinese and Japanese speakers experience [9]. With SLAP, students can practice on their own for as long as they like.

This paper is organized as follows: In section II, the HMM-Based accent detection system design and algorithm are discussed in detail. In section III, we propose more features suitable for accent detection. Experimental results

Lingyun Gu and John G. Harris are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, U.S.A. 32611.

are given in section IV. Finally, conclusions are drawn in section V.

II. Algorithm Description and System Design

The Hidden Markov Model (HMM) approach is the most well-known and widely used statistical method for characterizing the spectral properties of the frames of speech. The underlying assumption of the HMM is that the speech signal can be well characterized as a parametric random process, and that the parameters of the stochastic process can be determined in a precise, well-defined manner. It has been proved that the HMM method provides a natural and highly reliable way of recognizing speech for a wide range of applications and integrates well into systems incorporating both task syntax and semantics [7], [12]. In this paper, we will use the log-likelihood as the system's automatic score to evaluate the testing utterance. We apply an HMM-Based ASR system trained on native English speech. The correction of the pronunciation of each phoneme quantified by applying the Viterbi algorithm and then using the HMM to derive the log-likelihood values. This method has much better performance than our previously designed SLAP system based on the DTW algorithm. For each sentence the phone segmentation is obtained along with the corresponding log-likelihood for each segment [7], [12]. Then, for each phone segment we define the normalized log-likelihood $\hat{l_i}$ as:

$$\hat{l_i} = \frac{l_i}{d_i} \tag{1}$$

where l_i is the log-likelihood corresponding to the i-th phone and d_i is its duration in frames. The likelihood-based scores for a whole sentence L, is defined as the average of the individual normalized log-likelihood scores for each phone segment:

$$L = \frac{1}{N}\sum_{i=1}^{N} l_i \tag{2}$$

where the sum runs over the number of phones in the whole word.

The block diagram shown in Figure 1 illustrates our SLAP system and the details will be presented later. To best describe the utterance and improve the detection accuracy, a good endpoint detection algorithm is needed [1], [4], [13], [14]. For all experiments, we extend 30 ms in each direction (i.e., backward in time for the onset and forward in time for the offset) to allow for some inaccuracies in the automatic detection, and also to include a small amount of silence at the beginning and end of each utterance.

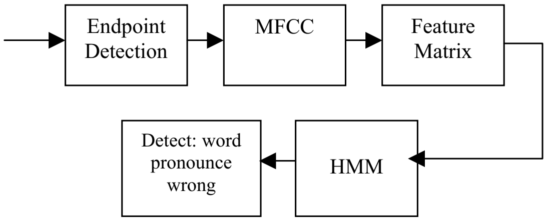

Fig. 1. The Block diagram for the whole SLAP system design.

Mel-frequency Cepstral Coefficients (MFCC) are used in systems with various languages and various tasks. This feature set stems from two ideas: vocal tract modeling and homomorphic filtering. The MFCC filter bank is composed of triangular filters spaced on a linear-logarithm scale. The spacing of filters follows the mel-frequency scale, which is inspired by critical band measurements of the human auditory system [2]. The MFCC features allow the comparison of the student and teacher's utterances to be invariant to loudness and pitch of the voices.

III. Additional SLAP Features

Other features have been found to be valuable for SLAP [15], including the duration of the word ($\triangle t$), the second formant (F2) and the third formant (F3). By analyzing the non-native English speech (Chinese) and native English speech (American), we find the statistics of the duration of the whole utterance produced by these two groups are very significant. This gives us a great opportunity to distinguish non-native and native English speakers. In addition, although the first formant (F1) is a very important feature in speech recognition systems, which represents the whole vocal tract information, it is not that important for the accent detection case. However, F2 and F3 play a very important role, since these two features represent the tongue movement information. As a matter of fact, many non-native English speakers bring their mother language pronunciation habits to English pronunciation causing their accent. One of these obvious habits is the improper tongue positions. In order to achieve a high accuracy accent detection by using the three features listed in this section, very accurate endpoint detection and formant detection algorithms are needed. Some moderate estimated errors may destroy the whole performance.

Here, we give the example of the usage of the first feature: $\triangle t$. From the training data, we can get the mean and stand deviation of the Dura from both non-native and native English speakers μ_{non}, μ_{nat}, δ_{non} and δ_{nat}. The following equations compute the posterior probability:

$$P(\omega_1| \triangle t) = \frac{P(\omega_1, \triangle t)}{P(\triangle t)} = \frac{P(\triangle t|\omega_1)P(\omega_1)}{P(\triangle t)} \tag{3}$$

$$P(\omega_2| \triangle t) = \frac{P(\omega_2, \triangle t)}{P(\triangle t)} = \frac{P(\triangle t|\omega_2)P(\omega_2)}{P(\triangle t)} \tag{4}$$

where, ω_1 and ω_2 represent the non-native and native English speakers classes, respectively. Since $P(\triangle t)$ is the same, and we suppose $P(\omega_1)$ and $P(\omega_2)$ are equal to 0.5. Thus, finally, we only need to compare the probability of $P(\triangle t|\omega_1)$ and $P(\triangle t|\omega_2)$. Assuming these two probability are Gaussian distributed with the means and variances derived from the training data listed above, we obtain:

$$P(\triangle t|\omega_1) = \frac{1}{\sqrt{2\pi}\delta_{non}} \exp\frac{(\triangle t - \mu_{non})^2}{2\delta_{non}^2} \tag{5}$$

$$P(\triangle t|\omega_2) = \frac{2}{\sqrt{2\pi}\delta_{nat}} \exp \frac{(\triangle t - \mu_{nat})^2}{2\delta_{nat}^2} \qquad (6)$$

IV. Experimental Results

In SLAP, utterances are digitized at a sampling rate of 11025 Hz. We take 256 samples as our MFCC window size, while the windows step size is 110 samples. A feature vector with 29 dimensions is used. The 29 dimensions consist of:

- Energy based features(E, △E, △△E);
- MFCC based features (10MFCC, 10△MFCC, 6△△MFCC);

Therefore, for each utterance, we finally get an N by M feature matrix, where N is 29 and M is the total window number for the given utterance.

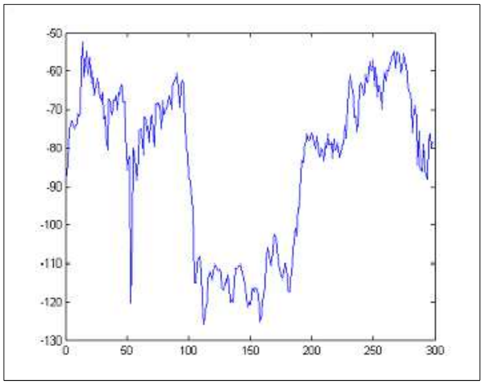

Fig. 2. The log-likelihood value of the testing utterance "arbitrary" along time windows.

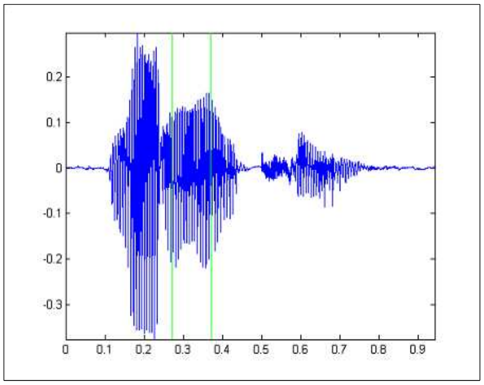

Fig. 3. Word "arbitrary" by native English speaker with highlighted portion to be emphasized.

The described algorithm has been tested with a database, which was collected by the Computational Neuro-Engineering Laboratory in the Electrical and Computer Engineering Department at University of Florida. From the database, we chose six non-native English speakers, six native English speakers to test 10 words that Chinese natives might have difficulties to pronounce. This leads to a training database of 720 sound files (each word was repeated six times by each speaker). All recordings were made in a typical computer lab environment.

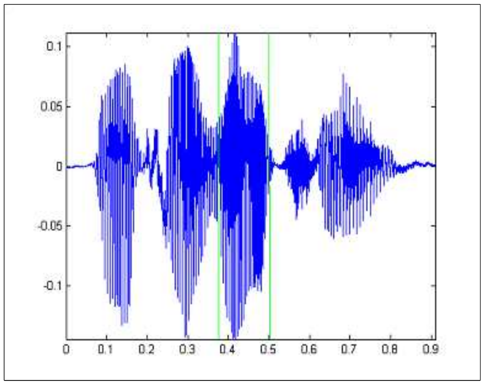

Fig. 4. Word "arbitrary" by non-native English speaker with highlighted portion to be corrected.

Figure 2 shows the log-likelihood value of the testing utterance compared with the statistical information obtained from the HMM training procedure. Figure 3 shows one utterance produced by a native English speaker, where the highlighted portion is the emphasized area corresponding to the error made by the non-native English speaker. Figure 4 shows the detected mismatched pronunciation portion made by the non-native English speakers. Here we use the word "arbitrary" for illustration. In this case, the error resulted from the incorrect inclusion of an addition phoneme in the word . Using normalized log-likelihood values, the SLAP system can classify native vs. non-native speech with an accuracy of 90.33%±4.23%.

Table I shows the details of feature $\triangle t$ of three words from our database ("anticipation", "ridiculous" and "literature"). Here we list 3 non-native and 3 native speakers. Each of them pronounces 4 utterances of 3 given words. These three words are picked up by the linguistic experts. And all of them are multi-syllabic words. Chinese usually have difficulties to correctly pronounce them. The means and standard deviation are also given. By using this feature only, we can roughly achieve a classification accuracy of non-native English speakers and Native English speakers around 85.32%±2.29%. We hope that inclusion of these features will further improve SLAP performance.

V. Conclusion

We designed SLAP, a system to automatically separate the non-native and native English speakers and detect foreign speakers' mispronunciation. The experimental results show our method can detect non-native English speakers' mispronunciation very robustly, especially for complicated, multi-syllabic words. Much work is needed in testing and fine-tuning the user interface of SLAP, and ultimately in quantifying pronunciation performance vs. a neutral control group. Additional SLAP features are also proposed. We also expect that future variation of SLAP will target children with learning disabilities.

Words List		Anticipation	Ridiculous	Literature
Non-native speaker1	Utterance1	1.043	1.119	0.873
	Utterance2	1.006	1.047	0.865
	Utterance3	1.035	1.083	0.950
	Utterance4	1.087	1.026	0.828
Non-native speaker2	Utterance1	1.099	1.067	0.917
	Utterance2	1.051	1.075	1.028
	Utterance3	1.103	1.164	0.942
	Utterance4	1.055	1.140	1.035
Non-native speaker3	Utterance1	1.083	0.869	0.942
	Utterance2	1.010	1.062	0.861
	Utterance3	1.075	0.921	0.929
	Utterance4	1.022	0.950	0.780
Mean for non-native speakers		1.0558	1.0436	0.9125
Std for non-native speakers		0.0337	0.0892	0.0756
Native speaker1	Utterance1	1.031	0.687	0.659
	Utterance2	0.881	0.731	0.667
	Utterance3	0.909	0.618	0.727
	Utterance4	0.800	0.650	0.679
Native speaker2	Utterance1	0.885	0.537	0.752
	Utterance2	0.929	0.687	0.707
	Utterance3	0.929	0.679	0.654
	Utterance4	0.921	0.735	0.671
Native speaker3	Utterance1	0.990	0.921	0.760
	Utterance2	0.962	0.626	0.699
	Utterance3	0.950	0.840	0.743
	Utterance4	0.966	0.764	0.683
Mean for native speakers		0.9291	0.7063	0.7001
Std for native speakers		0.0591	0.1028	0.0374

TABLE I

Word duration details of 24 utterances produced by 6 speakers for 3 different words.

References

[1] M. Karnjanadecha and S. A. Zahorian, *Signal modeling for high-performance robust isolated word recognition.* Speech and Audio Processing, IEEE Transactions on, Vol. 9 No. 6 , Sept. 2001, pp. 647 - 654

[2] M. D. Skowronski and J. G. Harris, *Increased MFCC filter bandwidth for noise-robust phoneme recognition.* Acoustics, Speech, and Signal Processing, 2002 IEEE International Conference on, Vol. 1, 2002 , pp. 801 - 804

[3] E. A. Yfantis, T. Lazarakis, A. Angelopoulos, J. D. Elison and Y. Zhang, *On time alignment and metric algorithms for speech recognition.* Information Intelligence and Systems, 1999. Proceedings. 1999 International Conference on, 1999, pp. 423 - 428

[4] L. Gu and S. A. Zahorian, *A new robust algorithm for isolated word endpoint detection.* Acoustics, Speech, and Signal Processing, 2002 IEEE International Conference on, Volume: 4 , 2002, pp. 4161 - 4164

[5] M. Eskenazi, *Detection of foreign speakers' pronunciation errors for second language training - preliminary results.* Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on , Volume: 3, pp. 1465 - 1468

[6] L. Rabiner and B. Juang, *Fundamental of speech recognition.* Prentice Hall, New Jersey, 1984

[7] H. Franco, L. Neumeyer and Y. Kim, O.Ronen, *Automatic Pronunciation Scoring for Language Instruction.* Acoustics, Speech, and Signal Processing, 1997 IEEE International Conference on, Volume: 2, 1997, pp. 1471 - 1474

[8] B. Beardsmore, H. Bilingualism, *Basic principles.* on-line version

[9] A. Brown, *Approaches to pronunciation teaching.* on-line version

[10] H. Seliger and E. Shohamy, *Second Language Research Methods.* Oxford, Oxford University Press, 1988.

[11] F. Grosjean, Life with two languages, *An introduction to bilingualism.* Cambridge, Harvard University Press

[12] L. Neumeyer, H. Franco, M. Weintraub and P. Price, *Automatic text-independent pronunciation scoring of foreign language student speech.* Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on , Volume: 3, pp. 1457 - 1460

[13] J. Gao, Y. Cao, L. Gu, J. Harris and J. Principe, *Detection of Weak Transitions in Signal Dynamics Using Recurrence Time Statistics.* Phys. Lett. A., Vol. 317, Issue. 1-2, pp. 64 - 72, 2003

[14] L. Gu, J. Gao and J. Harris, *Endpoint detection in noisy environment using a Poincare recurrence metric.* Acoustics, Speech, and Signal Processing, 2002 IEEE International Conference on, Volume: 1 , April 2003, pp. I-428 - I-431

[15] L. Arslan, *Foreign Accent Classification in American English.* Ph.D thesis, Duke University, 1996