ACOUSTIC ANALYSIS OF L1 INFLUENCE ON L2 PRONUNCIATION ERRORS: A CASE STUDY OF ACCENTED ENGLISH SPEECH BY CHINESE LEARNERS

Shuju Shi, Chilin Shih

University of Illinois at Urbana-Champaign
shujus2@illinois.edu, cls@illinois.edu

ABSTRACT

This paper aims to quantify dissimilarity of sound pairs between language inventories and to realize automatic prediction of error patterns of L2 production based on the proposed quantifying measures. Language pairs used are Chinese and English and only their vowel inventories are considered. To study the two languages systematically, maximal phone combinations are desired to cover possible variations introduced by different contexts. The stimuli used are: 1) all possible syllables with 4 tonal variations in Chinese (n=1860 tokens), and 2) real monosyllabic English words (n=1667 words). Nine native Chinese speakers (F=5, M=4) and two native English speakers (F=1,M=1) were recorded. Mel-Frequency Cepstral Coefficients (MFCCs) are used as features, Principle Component Analysis (PCA) is applied and Euclidean distance is used to compute the acoustic distances. The results showed that error prediction using our proposed method are consistent with Perceptual Assimilation Model for L2 Learning (PAM-L2) and Speech Learning Model (SLM). Results also showed that our proposed quantitative measures based on phonetic features are able to incorporate certain phonological influence.

1. INTRODUCTION

The possible influence of first language (L1) on second language (L2) learning has long been studied and a lot of progress has been made regarding this issue during the last several decades. Among them, two influential models dealing with L1 influence on both production and perception of L2 are Speech Learning Model (SLM) by Flege et al. [6,7] and Perception Assimilation Model (PAM) by Best et al. [2,3]. SLM focused more on the production of speech and tried to account for the variation in the extent of individuals' learning phonetic segments in an L2 whereas PAM, specifically PAM-L2, focused more on the perception aspect of L2 acquisition and studies how L2 learners assimilate/dissimilate a new sound in L2 according to his/her L1 phonology categories. There have been intensive research effort following these two models suggesting that the reason behind L1 influence on L2 is diverse, and can come from phonology, phonetics and the combination of the two [1,5,9,10].

Researchers in Computer Aided Language Learning (CALL) work from a different angle to incorporate L1 knowledge in machine learning to explain L2 speaker's pronunciation,with a goal to improve error detection and to increase accuracy of pronunciation evaluation. Depending on the way they utilize the L1 knowledge, their approach can be classified as being implicit or explicit. In typical implicit approaches, researchers usually use L1 acoustic features to do adaptive training of model trained using native speaker's data of the target language and later use this hybrid model to do evaluation or pronunciation error detection. And results have shown that using L1 dependent models outperform L1 independent models [4, 15]. The explicit approach tend to be more laborious compared with its implicit counterpart. One way to use this knowledge is to pre-define error types and then manually annotated data to later train statistical models and do automatic evaluation or error detection [8]. A more automatic or semiautomatic approach researchers used is to first annotate L2 speech and then using the mapping of transcription and the canonical forms to automatically derive rules of error pattern [13].

One problem which is still seldom studied is automatic prediction of L1-related pronunciation errors based on L1 and L2 phonological and phonetic information with linguistically explainable results. In this study, we address the following research questions: 1) how to quantify the acoustic differences of speech sounds between two different languages, 2) how to capture phonological errors in a phonetically-based measurements, 3) how to automatically derive L1-dependent errors based on the quantified differences.

The rest of the paper is organized as follows: Section 2 introduces methodology. Sections 3-4 give the

experiments and results and Sections 5-6 talk about discussion, conclusion and future work.

2. METHODOLOGY

2.1. Vowel Inventories and Stimuli Design

In this study, we focused on Chinese learners' acquisition of English vowels. According to Ladefoged and Disner (2012), the American English as spoken by national newscasters has 15 distinct vowels, which include 10 monophthongs (/ɑ, ɛ, æ, ʌ, i, ɪ, ɚ, ɔ, u, ʊ/) and 5 diphthongs (/eɪ, oʊ, aɪ, aʊ, ɔɪ/). Mandarin, according to Lee and Zee (2003), has six monophthongs (/ɑ, o, ɤ, i, u, y/) and four diphthongs (/ai, ao, ei, ou/). Burnham et al. (2002) found that orthographic-mapping could contribute to greater phonological awareness. To address possible influence of the official romanization system, Pinyin, on the pronunciation of speakers, we also include rhotic /aɻ/ (as er in Pinyin) and /ie, uo/ (as ie and uo in Pinyin). .

To maximize coverage of phone combination in monosyllabic words in both languages, the stimuli we use are:1) all possible syllables with 4 tonal variations in Chinese with necessary repetitions (1860 words), and 2) monosyllabic words of English covering as many phone combinations as possible and of equivalent size with the Chinese counterpart (1667 words).

2.2. Participants and Recording Procedure

Nine standard Chinese (Mandarin) speakers (5F, 4M) and two English speakers (1F, 1M) and each participant was paid 15 US dollars per hour. All of the Chinese participants were born and grew up in Beijing, speaking the Beijing dialect, ages between 19-34 (mean: 27; std.: 4.2) and they started learning English at ages 6-7. The English participants are from the suburbs of the Chicago area.The female participant is 22 years old and the the male participant is 21.

The recording took place in a soundproof booth. The stimuli were presented to the participants one by one using a program written in Matlab and the participants were free to have a break after each 100 words. The recording of the English and Mandarin stimuli for each speaker is around 1.5 hours and 2 hours respectively.

2.3. Force Alignment and Acoustic Features

Forced alignment is then applied to the recorded data. Acoustic model used for English is trained based on 100 hours of clean speech from LibriSpeech (Panayotov et al., 2016) and that for Mandarin is trained based on the recorded speech (around 15 hours) using Kaldi [14]. The alignment results are then manually checked and adjusted.

Acoustic features we used in this study are 39 dimensional Mel Frequency Ceptural Coefficients(MFCCs) which are able to capture the spectral information and meanwhile accomodate to human perception of the frequency components. MFCCs were extracted at five equally distributed intervals for each vowel segment (10%, 30%, 50%, 70% and 90%), which gives us a feature dimension of 195.

2.4. Statistical Analysis

The statistical method we need should be able to address the following conditions:

a) *whether phonological categories of L1 and L2 exist in a common space.* PAM-L2 and SLM both agreed that L1 and L2 phonological categories share a common space.

b) *whether phonological influence can be incorporated into the acoustic-phonetic measures.* PAM-L2 emphasized that the perceived invariants for learners were at higher phonological and phonetic levels rather than phonetic detail. SLM emphasized more on acoustic cues of phonetic contrasts.

c) *being able to derive weighted features.* SLM hypothesized that the phonetic category learners established for L2 sounds might differ from L1 speakers' and learners' representations of the sounds might be based on different features or different weights of the same features from L1 speakers.

With all those conditions considered, the quantifying method chosen in this study is Principle Component Analysis (PCA). PCA is a statistical procedure to convert observations of variables into a set of vectors, named principle components, where each component is the linear combination of the variables and any two of them are uncorrelated orthogonal. In a traditional PCA approach, principal components are calculated by first getting the eigenvectors, i.e., principal component directions, of the covariance matrix of the observation data and then projecting each training example onto the principle component directions (as the procedure above). In this study we proposed to use PCA in three slightly different ways regarding how we get the principal components (hereafter refereed to as PCA1, PCA2, and PCA3, respectively):

- PCA1: computing principal directions based

only on native Chinese (NC) data.

- PCA2: computing principal directions based only on native English (NE) data.
- PCA3: computing principal directions based on both NC and NE data.

The PCA1 approach assumes that learners would apply the same features as well as feature weights they used in their L1. The PCA2 approach, on the other hand, assumes that learners would use the same features as well as feature weights as native speakers of the target language. The PCA3 approach assumes that learners would use a combination of the features from the two languages and weight the features accordingly.

3. EXPERIMENTS AND RESULTS

In order to examine the acoustic distributions of vowel inventories in each language, first and second formants (F1 and F2) of NE, NC and English speech by Chinese learners (L2E) are extracted using FormantPro [16].

Figures 1-2 showed the F1-F2 plots of monophthongs of native Chinese and native English respectively. As can be seen, the overall tendency of vowel distributions for both NE and NC is consistent with previous findings. What are problematic is F2 values gotten for high/mid-high back vowels (such as /u/, /o/, /ʊ/) are not accurate enough. This is due to the fact that high back vowels tend to have low F1s and F2s to the extent that it is difficult to separate them and sometimes the algorithm takes F3 as F2 for these vowels, which happens to be the case in data presented here. That is also one of the reasons why MFCCs other than formants were used as features in this study.

As to the predictions SLM/PAM would make based on the relationship of vowel inventories between NE and NC, some possibilities are given:

a) /i, ɪ/ and /u, ʊ/ would probably fall into either single-category or category-goodness assimilation with their Chinese counterpart /i/ and /u/ respectively.
b) /ɚ/ and /ɔ/ also have Chinese counterparts rhotic /aˠ/ and /o/.
c) The situations for /ɑ, ɛ, æ, ʌ/ might be more complicated. Each of them could be assimilated to Chinese /ɑ/ or it could also be the case that the difference between them and Chinese /ɑ/ is salient enough thus new category/categories would be established for them.

As stated in Section 2, PCA was applied in three different ways to the data and only the first 30 principle components (accounting for over 90% of data variance for all three subsets of of the corpus) were kept and were then used to calculate the Euclidean distance as the acoustic distance between each pair of sounds.

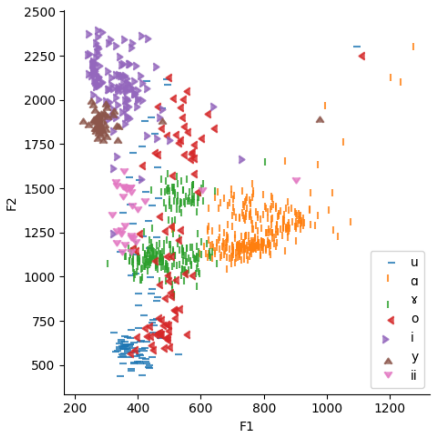

Figure 1: F1-F2 plot for NC monophthongs, where ii represents ɨ, and points on the upper left could be outliers resulting from extraction errors.

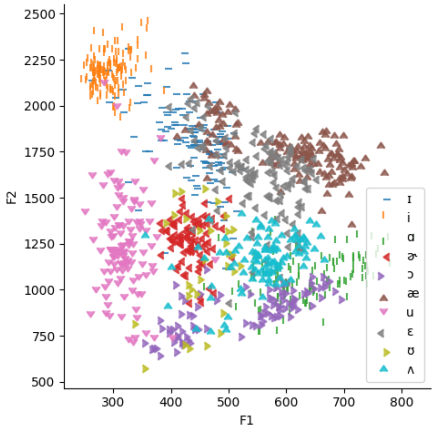

Figure 2: F1-F2 plot for NE monophthongs.

Table 1 showed if assimilation were ever to happen, the top 2 candidates of Chinese for each English vowels using three different PCA approaches. The predicated assimilation candidates are largely consistent with each other except for /i/ and /ɪ/, suggesting strong patterns of error tendency.

Figure 3 showed the confusion matrix between NE and L2E vowels using PCA3, under which condition it is assumed that NC and NE share the same feature sets and feature weights in differentiating vowels. Results showed that vowels within each cluster /ɑ, ɛ, æ, ʌ, ɔ, aʊ, aɪ/, /i, ɪ, eɪ/ are likely to be confused with each other.

Table 1: The First Two Closest Chinese Candidates for Each English Vowel Calculated Using Each PCA Approach

	PCA1		PCA2		PCA3	
	1st	2nd	1st	2nd	1st	2nd
/ɑ/	/ɑ/	/aˇ/	/ɑ/	/aˇ/	/ɑ/	/aˇ/
/ɛ/	/ai/	/ɑ/	/ai/	/ɑ/	/ai/	/ɑ/
/æ/	/ai/	/ɑ/	/ai/	/ɑ/	/ai/	/ɑ/
/ʌ/	/ɑ/	/ai/	/ɑ/	/ai/	/ɑ/	/ai/
/i/	/i/	/u/	/i/	/u/	/u/	/i/
/ɪ/	/u/	/ei/	/ei/	/ie/	/ei/	/u/
/ɚ/	/ai/	/ei/	/ai/	/ei/	/ai/	/ei/
/ɔ/	/ɑ/	/ao/	/ɑ/	/ao/	/ɑ/	/ao/
/u/	/ɨ/	/u/	/ɨ/	/u/	/ɨ/	/u/
/ʊ/	/ɤ/	/ɑ/	/ɤ/	/ɑ/	/ɤ/	/ɑ/
/aɪ/	/ai/	/ɑ/	/ai/	/ɑ/	/ai/	/ɑ/
/aʊ/	/ɑ/	/aˇ/	/ɑ/	/aˇ/	/ɑ/	/aˇ/
/eɪ/	/u/	/ei/	/u/	/ei/	/u/	/ei/
/ɔɪ/	/ai/	/ɤ/	/ai/	/ɤ/	/ai/	/ɤ/
/oʊ/	/ɤ/	/ɑ/	/ɤ/	/ɑ/	/ɤ/	/ɑ/

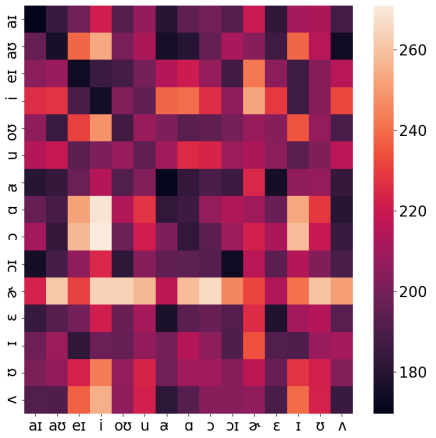

Figure 3: Heatmap plot of confusion matrix of vowels between NE and L2E using PCA3, where darker colors indicate smaller acoustic difference.

4. DISCUSSION

Due to the limitation of space, a lot more results might not be able to be explained in more details but what need to be notified is that in our case, the language pairs we use have different distribution of vowel inventories. Figure 4 showed that hierarchical clustering results based on native English and native Chinese data using raw MFCCs, which suggests that English vowels are less separable even in native data. In addition, in this study we followed traditions in PAM-L2 where only up to two assimilation candidates is given for each vowel. In the future we would like to examine possible threshold values to determine the best number of candidates for each vowel based on their distribution.

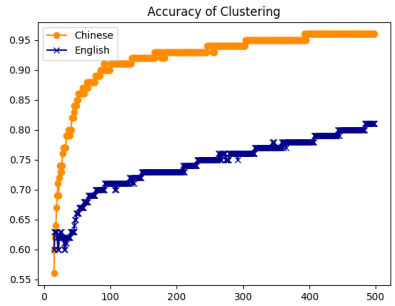

Figure 4: Accuracy of Hierarchical Clustering

5. CONCLUSIONS

In this study, we explored the possibility of quantifying acoustic difference of sounds between two different languages as well the possibility to automatically derive L1-dependent errors based on the quantified differences. The language pairs we used were English and Chinese and the error prediction results were tested on L2E speech. PCA was adopted and slightly adjusted to simulate feature and feature weights speakers and learners use in differentiating vowels within each language inventory and between language inventories. The results showed that our predicted results are consistent with predictions by PAM-L2 and SLM and our predicted error patterns are largely consistent with the actual production results of L2E learners. Results are still needed to be further examined in more detail and be tested in counter learning directions. By using PCA approaches we were able to simulate features and weights of features used in each vowel inventory, but the resulted measures were largely production-based and are acoustic measures. In the future, one of the research interests would be how to capture more perceptual effect and phonological influence in the quantified measures.

6. ACKNOWLEDGEMENT

We would like to thank all the participants who did the recording. This work also benefits from feedback from Prof. Ryan Shosted and Prof. Mark Hasegawa-Johnson.

7. REFERENCES

[1] Almbark, R., Bouchhioua, N., and Hellmuth, S. (2014, March). Acquiring the phonetics and phonology of English word Stress: comparing learners from different L1 backgrounds. *In Proceedings of the international symposium on the acquisition of second language speech*, Concordia Working Papers in Applied Linguistics (Vol. 5, pp. 19-35).

[2] Best, C. T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. *The development of speech perception: The transition from speech sounds to spoken words*, 167(224), 233-277.

[3] Best, C. T., and Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. *Language experience in second language speech learning: In honor of James Emil Flege*, 1334, 1-47.

[4] Duan, R., Kawahara, T., Dantsuji, M., and Nanjo, H. (2017). Transfer Learning based Non-native Acoustic Modeling for Pronunciation Error Detection. *In SLaTE* (pp. 42-46).

[5] Escudero, P., and Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition*, 26(4), 551-585.

[6] Flege, J. E. (1991). Age of learning affects the aughenticity of voice onset time (VOT) in stop consonants produced in a second language.*The Journal of the Acoustical Society of America*, 89(1), 395-411.

[7] Flege, J. E., MacKay, I. R., and Meador, D. (1999). Native Italian speakers' perception and production of English vowels.*The Journal of the Acoustical Society of America*, 106(5), 2973-2987.

[8] Gao, Y., Xie, Y., Cao, W., and Zhang, J. (2015). A study on robust detection of pronunciation erroneous tendency based on deep neural network. *In Sixteenth Annual Conference of the International Speech Communication Association*.

[9] Guion, S. G., Flege, J. E., Akahane-Yamada, R., and Pruitt, J. C. (2000). An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants. *The Journal of the Acoustical Society of America*, 107(5), 2711-2724.

[10] Keidel, J. L., Zevin, J. D., Kluender, K. R., and Seidenberg, M. S. (2003). Modeling the role of native language knowledge in perceiving nonnative speech contrasts. *In Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 2221-2224).

[11] Ladefoged, P., and Disner, S. F. (2012). Vowels and consonants. *John Wiley and Sons.*

[12] Lee, W. S., and Zee, E. (2003). Standard Chinese (Beijing). *Journal of the International Phonetic Association*, 33(1), 109-112.

[13] Lo, W. K., Zhang, S., and Meng, H. (2010). Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system. *In Eleventh Annual Conference of the International Speech Communication Association*.

[14] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Silovsky, J. (2011). The Kaldi speech recognition toolkit. *In IEEE 2011 workshop on automatic speech recognition and understanding* (No. EPFL-CONF-192584). IEEE Signal Processing Society.

[15] Shi, S., Kashiwagi, Y., Toyama, S., Yue, J., Yamauchi, Y., Saito, D., and Minematsu, N. (2016, September). Automatic Assessment and Error Detection of Shadowing Speech: Case of English Spoken by Japanese Learners. *In INTERSPEECH* (pp. 3142-3146).

[16] Xu, Y. and Gao, H. (2018). FormantPro as a tool for speech analysis and segmentation. *Revista de Estudos da Linguagem*.