INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Capturing L1 Influence on L2 Pronunciation by Simulating Perceptual Space Using Acoustic Features

Shuju Shi[1], *Chilin Shih*[1], *Jinsong Zhang*[2]

[1]University of Illinois at Urbana-Champaign, USA
[2]Beijing Language and Culture University, China

shujus2@illinois.edu, cls@illinois.edu, jinsong.zhang@blcu.edu.cn

Abstract

Theories of second language (L2) acquisition of phonology/phonetics/pronunciation/accent often resort to the similarity/dissimilarity between the first language (L1) and L2 sound inventories. Measuring the similarity of two speech sounds could involve many acoustic dimensions, e.g., fundamental frequency (F0), formants, duration, etc.. The measurement of the sound inventories of two languages can be further complicated by the distribution of sounds within each inventory as well as the interaction of phonology and phonetics between the two inventories. This paper attempts to propose a tentative approach to quantify similarity/dissimilarity of sound pairs between two language inventories and to incorporate phonological influence in the acoustic measures used. The language pairs studied are English and Mandarin Chinese and only their vowel inventories are considered. Mel-Frequency Cepstral Coefficients (MFCCs) are used as features, and Principle Component Analysis (PCA) is used and slightly adjusted to simulate the perceptual space. Similarity/dissimilarity of sound pairs between the language inventories are examined and potential L2 error patterns are predicted based on the proposed approach. Results showed that predicted results using the proposed approach can be well related to those by Speech Learning Model (SLM), Perceptual Assimilation Model for L2 (PAM-L2) and Native Language Magnet Model (NLM).

Index Terms: L2 acquisition, speech perception, speech production

1. Introduction

How phonology and/or phonetics of learners' first language (L1) influence the production and/or perception of their second language (L2) has long been studied in the area of L2 acquisition. Several models have been proposed to address L1 influence on L2 perception and/or production and among them three influential ones are: the Speech Learning Model (SLM) [1][2], the Perceptual Assimilation Model for L2 learners (PAM-L2) [3] [4] and the Native Language Magnet Model (NLM) [5] [6]. To summarize briefly: 1) SLM accounts for the variation in the extent of individuals learning phonetic segments in an L2; 2) PAM-L2 accounts for how L2 learners assimilate/dissimilate a new sound in L2 according to their L1 phonology categories; 3) NLM accounts for how L1 experience serves as language-specific filters to warp the acoustic dimensions and influence how sounds in L2 are perceived.

The three models share commonalities in some aspects and complement with each other in the other. SLM and PAM-L2 are more similar with each other in the sense that notions they both agree on are: 1) L1 and L2 phonology categories exist in a common space; and 2) the mechanisms and processes used in learning the L1 sound system can also be applied to L2 learning. What the two models complement with each other are two-fold. First, SLM focused more on the production of speech whereas PAM-L2 focused more on the perception aspect of L2 acquisition. Second, PAM and PAM-L2 claimed that the perceived invariants in the course of learner's perceptual learning are articulatory gestures whereas SLM emphasized acoustic-phonetic cues more. Compared with SLM, NLM is more similar to PAM-L2 in the sense that the basic unit in NLM is phonological and NLM also focus more on the perception perspective. What differs between NLM and PAM-L2 is that: 1) both being phonological, the basic unit in NLM is phonemic instead of articulatory gestures in PAM-L2; and 2) NLM accounts for L1 influence on L2 in terms of perceptual warping rather than sound assimilation/dissmilation. There have been intensive research efforts following paradigms proposed by these three models suggesting that the reason behind L1 influence on L2 is diverse and can come from phonology, phonetics, and the combination of the two [2][7][8][9].

The models offered easy-to-follow paradigms in studying acquisition of single L2 phones (SLM), sound contrasts (PAM-L2) and phoneme pairs (NLM). But there are potential issues in these approaches when studying two language inventories systematically is desired. For SLM and PAM-L2, the potential problem lies in that features used in both models are more descriptive than quantitative, such as the articulatory gestures in PAM-L2 and "perceived similarity" in SLM. The potential problem for NLM approach is that it might be too laborious when applied to study language inventories systematically, considering all the possible number of phoneme pairs from both languages. In the case of [9], to study Japanese learners' perceptual space of /r/ and /l/ and how it differs from native English speakers, 36 /ra/ and /la/ stimuli are synthesized and different combinations of the stimuli are generated for the identification and discrimination tasks. Multiple native English speakers and L1-Japanese L2-English learners participated in both the identification and discrimination task to get data needed to generate the perceptual space using Multidimensional scaling (MDS). This approach is valuable and essential in understanding L2 perception but it is less feasible to be extended to study two language inventories systematically.

In this study, we would like to explore the possibility of proposing an approach that is able to quantify sound similarity/dissmilarity between language inventories and in the meantime capturing L1 influence without conducting perceptual experiments, i.e., simulating perceptual space using acoustic features.

The rest of the paper is organized as follows: Section 2 introduces methodology. Section 3 gives the experiments and results and Sections 4-5 talk about discussion, conclusion and future work.

http://dx.doi.org/10.21437/Interspeech.2019-3183

2. Methodology

This section introduces and justifies stimuli used, participants recruited and the statistical methods adopted. Stimuli used in this study is similar to what we used in a previous study [10].

2.1. Vowel Inventories and Stimuli Design

This study focuses on language pairs of English and Mandarin Chinese and only their vowels are considered. The reason why only vowels are considered is because compared with consonants, features used to differentiate vowels are more uniform and less scattered. English and Mandarin Chinese are chosen for they have different vowel inventories and they are probably the most studied languages. Having a well-established vowel inventory is important in our approach. According to Ladefoged and Disner [11], 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 [12], has six monophthongs (/ɑ, o, ɤ, i, u, y/) and four diphthongs (/ai, ao, ei, ou/). In order to address possible influence of orthographic-mapping on phonological awareness as stated in Burnham [13], rhotic /aɻ/ (as er in Pinyin) and /ie, uo, ve/ (as ie, uo and ue in Pinyin) are also included in Chinese vowel inventory.

To simplify the situation, only monosyllabic words are considered in both languages. Since acoustics of vowels are heavily influenced by their contexts, maximal coverage of phone combinations are desired to cover all possible variations. With those being said, the stimuli we used 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 native Mandarin speakers (5F, 4M) and two English speakers (1F, 1M) were recruited and each participant got paid 15 US dollars per hour for their participation. All Chinese participants were born and grew up in Beijing (ages between 19-34; mean: 27; std.: 4.2) and both English participants were born and grew up in the suburbs of the Chicago area (F: 22, M: 21). The Mandarin speakers are classified into two groups based on their English proficiency, 3 (2F, 1M) are advanced and the remaining 6 are intermediate level.

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. English speakers only did recording for the English stimuli and Mandarin speakers did recording for both English and Chinese stimuli.

2.3. Forced 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 [14] and that for Mandarin is trained based on the recorded speech (around 15 hours) using Kaldi [15]. 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 accommodate 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) *being able to derive weighted features and warping acoustic space.* 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. NLM asserted that learners' perception of L2 sounds/phonemes is filtered by their L1 perceptual space.

Principle Component Analysis (PCA) is adopted in this study to explore the possibility of addressing the above conditions. 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, given dataset X of n samples, principal components are calculated by:

$$C = \frac{1}{n-1} X^* \otimes X \tag{1}$$

$$V^{-1} C V = D \tag{2}$$

$$Y = X \cdot \mathbf{W} \tag{3}$$

where C is the covariance matrix, V and D are the eigenvectors and eigenvalues respectively, W is the selected eigenvectors and Y is the resulted principle components. In this study we proposed to use PCA in two slightly different ways regarding how we get the principal components (hereafter refereed to as PCA1 and PCA2 respectively):

- PCA1($W = W_{NC}$): computing principal directions based only on native Chinese (NC) data.
- PCA2 ($W = W_{NC+NE}$): computing principal directions based on both NC and native English (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 assumes that learners would use a combination of the features from the two languages and weight the features accordingly.

2.5. Quantifying Sound Similarity

Gaussian Mixture Models (GMMs) are trained based on either raw MFCCs or the principle components achieved for each corpus and then the classification results are used to quantify similarity of sounds.

3. Experiments and Results

3.1. Phoneme Recognition using raw MFCCs

Speaker-dependent GMMs are trained using the recorded stimuli based on raw MFCC features and 10-fold cross validation

was performed. Table 1 showed phoneme recognition results for each speech type. In Table 1, results for native English, native Chinese and L2 English(1) are average phoneme recognition rates based on GMMs trained on the corresponding speech data, whereas the result for L2 English(2) are average classification rates by GMMs trained on native English data.

Table 1: *Phoneme Recognition Rate using GMM*

Speech Type	**Correctness Rate**
Native English	80.5%
Native Chinese	86.7%
L2 English(1)	72.1%
L2 English(2)	30.0%

As shown in Table 1, the correctness rate of phoneme recognition is the highest for native Chinese, following by native English and then L2 English data. The fact that the recognition rate is relatively high for L2 English when using GMMs trained based on L2 English data but significantly lower when using GMMs trained based on native English data suggests that: 1) for intermediate and advanced level L2 learners, they have developed their own pattern of pronouncing phonemes in an L2, or they have established the so-called "prototypes" of phonemes in an L2 as suggested in NLM; 2) the "prototypes" L2 learners established are different from native speakers as indicated by the low correctness rate for L2 English(2). We would then further examine the difference and explore the possibility of addressing the underlying mechanisms that could explain this difference.

3.2. Classification Results using PCA features

As stated in Section 2, besides the traditional approach of PCA, here we slightly adjusted how PCA is applied to the data regarding ways of calculation the principle directions. We applied all three ways, traditional PCA, PCA1 and PCA2 to the data and only the first 30 principle components were kept (keeping over 90% of variance of the data under each condition). Then GMMs are trained and Figures 1-3 showed the plots of confusion matrices of native English data for each approach, traditional PCA, PCA1 and PCA2, respectively.

As can be seen, when changing from the traditional PCA to PCA2 then PCA1, the overall tendency among them are consistent with each other, suggesting that generally speaking, the acoustic features used to differentiate vowels are similar in native English and native Chinese. On the other hand, the correctness rate for most phonemes went down when using PCA1 and PCA2 approach, suggesting that there is feature and/or feature weights difference in differentiating vowels in these two languages. The correctness rate for /ʊ/ decreased to 0. Correctness rates for phonemes /ɑ, æ, ʌ/ all decreased and the misclassification rate among them increased. The correctness rate for /aʊ/ increased.

Figures 4-5 showed the plots of confusion matrices between native English and native Chinese phonemes based on features achieved by the PCA1 and PCA2 approaches, respectively. Again, we can see that the overall tendency between the approaches are similar with each other with slightly difference in certain sounds. The results showed that English vowels /ɑ, ɛ, æ, ʌ, ɔ, aʊ, aɪ/ are all very likely to be assimilated to Chinese /ɑ/, and English vowels /i, ɪ, eɪ/ are very likely to be assimilated to Chinese vowel /ei/. The tendency is stronger using PCA1 than using PCA2.

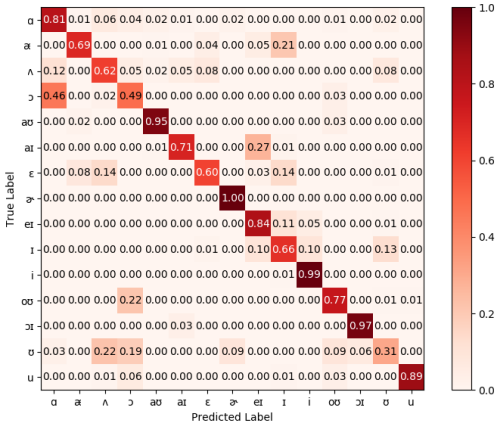

Figure 1: *Confusion Matrix for Native English data based on Features by traditional PCA Approach, overall correctness rate: 76.1%*

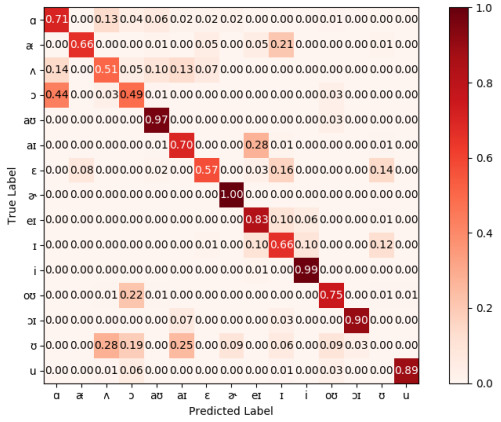

Figure 2: *Confusion Matrix for Native English data based on Features by traditional PCA1 Approach, overall correctness rate: 73.2%*

3.3. L2 English Production Results

Figure 6 showed the confusion matrix between L2 English data and native English data based on classification results of GMMs trained using raw MFCCs. Both confusion clusters predicted in the previous section turned out to be true in the actual production of L2 English data.

4. Discussion

As to the predictions SLM, PAM-L2 and NLM would make based on the relationship of vowel inventories between NE and NC, some possibilities are: 1) /i, ɪ/, /u, ʊ/ could be assimilated to Chinese counterparts /i/ and /u/ respectively; 2) 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

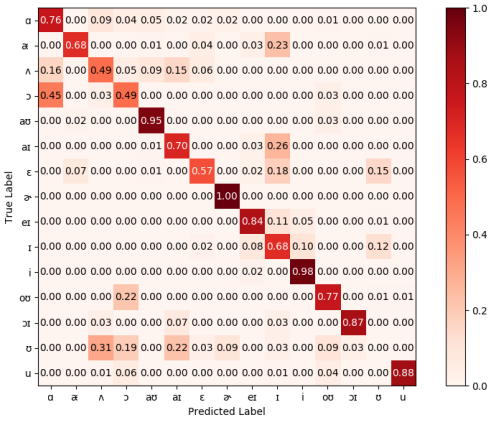

Figure 3: *Confusion Matrix for Native English data based on Features by traditional PCA Approach, overall correctness rate: 73.5%*

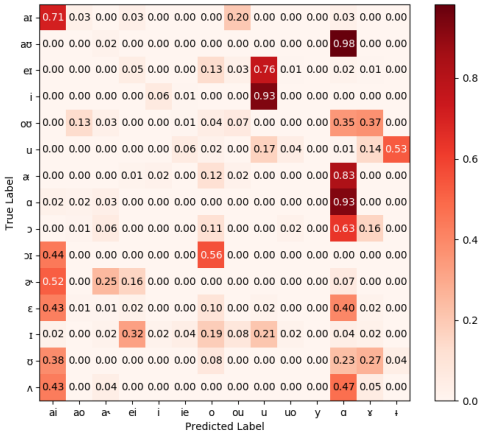

Figure 5: *Confusion Matrix between Native English and Native Chinese using PCA2 Approach.*

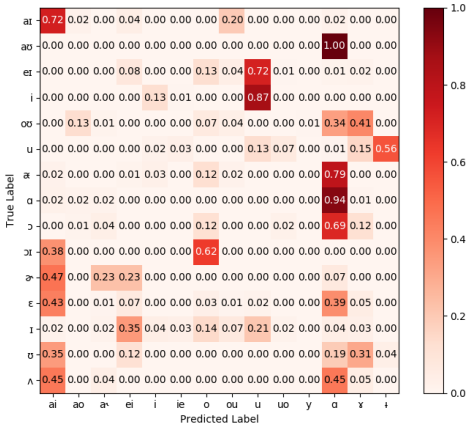

Figure 4: *Confusion Matrix between Native English and Native Chinese using PCA1 Approach*

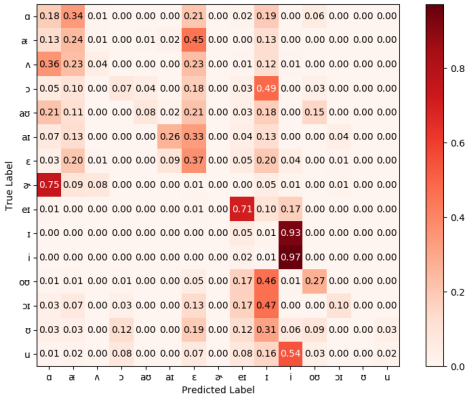

Figure 6: *Confusion Matrix between Native English and L2E English using Raw MFCCs*

new category/categories would be established. In this study, confusing clusters /i, ɪ/ and /ɑ, ɛ, æ, ʌ/ are correctly predicted. In addition, diphthongs /eɪ/ and /aɪ/ are also predicted to be involved in each of the corresponding cluster. As suggested in [16], Mandarin compound finals tend to be pronounced with quicker and smaller lip and tongue movements and thus Chinese learners tend to shorten diphthongs or replace diphthongs with monophthongs in their English pronunciation. However, the mis-classification rate is very high among high vowels /u, i/ in both the prediction result and the actual production of L2 English data. The reason why this is the case is so far less explainable.

5. Conclusions

In this study, we proposed an approach to simulate perceptual space using acoustic features and then it is further applied to quantify similarity of sounds between language pairs. Similarity/dissimilarity of sound pairs between the language inventories of Mandarin Chinese and English are examined and potential L2 error patterns are predicted based on the proposed approach. Results showed that predicted results using the proposed approach can be well related to those by Speech Learning Model (SLM) and Perceptual Assimilation Model for L2 (PAM-L2) and Native Language Magnetic Model (NLM). In the future, we would like to first examine in more detail results about the high vowels and also test the performance of the approach in the counter learning direction, that is, Chinese Speech by English learners.

6. Acknowledgements

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

7. References

[1] J. E. Flege, "Age of learning affects the authenticity of voice onset time (vot) in stop consonants produced in a second language," *The Journal of the Acoustical Society of America*, vol. 89(1), pp. 395–411, 1991.

[2] J. Flege, "Native italian speakers' perception and production of english vowels," *The Journal of the Acoustical Society of America*, vol. 106(5), pp. 2973–2987, 1999.

[3] C. T. Best, "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*, vol. 167(224), pp. 233–277, 1994.

[4] C. Best and M. Tyler, "Nonnative and second-language speech perception: Commonalities and complementarities," *Language experience in second language speech learning: In honor of James Emil Flege*, vol. 1334, pp. 1–47, 2007.

[5] P. Kuhl, "Learning and representation in speech and language," *Current Opinion in Neurobiology*, vol. 4, pp. 812–822, 1994.

[6] P. K. Kuhl, "A new view of language acquisition," *Proceedings of the National Academy of Sciences USA*, vol. 97, pp. 11 850–11 857, 2000.

[7] P. Almbark, N. Bouchhioua, and S. Hellmuth, "Acquiring the phonetics and phonology of english word stress: comparing learners from different l1 backgrounds." in *In Proceedings of the international symposium on the acquisition of second language speech*, 2014, pp. 19–35.

[8] P. Escudero and P. Boersma, "Bridging the gap between l2 speech perception research and phonological theory," *Studies in Second Language Acquisition*, vol. 26(4), pp. 551–585, 2004.

[9] P. Iverson and P. K. Kuhl, "A perceptual interference account of acquisition difficulties for non-native phonemes," *Cognition*, vol. 87(1), pp. B47–B57, 2003.

[10] S. Shi and C. Shih, "Acoustic analysis of l1 influence on l2 pronunciation errors: A case study of accented english speech by chinese learners," in *ICPhS2019*, Melbourne, Australia, 2019.

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

[12] W. S. Lee and E. Zee, "Standard chinese (beijing)," *Journal of the International Phonetic Association*, vol. 33(1), pp. 109–112, 2003.

[13] D. Burnham, M. Tyler, and S. Horlyck, "Periods of speech perception development and their vestiges in adulthood," in *In P. Burmeister, T. Piske & A. Rohde (Eds.)*, Trier, Germany, 2002, pp. 281–300.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *In P. Burmeister, T. Piske & A. Rohde (Eds.)*, ICASSP 2015, 2015.

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

[16] S. Michael and B. Smith, *Learner English: A teacher's guide to interference and other problems.* Cambridge University Press, 2001.