SYLLABLE-BASED ACOUSTIC SIMULATION PROCESS IN SECOND LANGUAGE PRODUCTION

LI Ya

Abstract: This paper discusses how Chinese speakers produce English sonorant consonants embedded in syllable structures novel to them. It shows that speakers may simulate target second language (L2) sounds based on acoustically similar first language (L1) sounds and it is a rather complex process in which sonorant type, vowel context, and articulatory constraints all come into play. Particularly, simulation of L2 speech based on acoustic cues is found to occur not only at the segmental but also at the syllabic level in the English production of Chinese speakers. To explain this finding, the present study proposes an acoustic-articulatory model of L2 syllable production and assumes that the acoustic simulation process is syllable-based and the output form is guided by salient acoustic cues and modulated by both language general and specific coarticulatory mechanisms.

Keywords: Acoustic simulation, Coarticulatory mechanisms, L2 syllable production

1. INTRODUCTION

L1 transfer is most often manifested as segmental substitution. To understand how speakers substitute an L2 sound with an L1 sound is at the center of L2 speech research ([1], [2], [5], [6], [7], [19]). These studies generally agree that L2 learners approach a target L2 sound from three levels: segmental, syllabic/moraic, and prosodic. L1 transfer at the first level has been studied the most, and numerous models have been proposed to explain how L1 influences perception and production of L2 segments. The Speech Learning Model (SLM) proposed by [11], for example, assumes that the mechanism of equivalence classification is an obstacle for L2 segmental acquisition. To acquire an L2 sound, learners need to discern its difference from an L1 sound and establish a new category for the L2 sound, or else accurate acquisition is difficult. However, the SLM does not specify what metrics that L2 learners use to gauge the perceived difference between L1 and L2 sounds, whether or not the metrics are acoustically, articulatorily, perceptually, or phonologically based ([2], [6], [7]). As a result, exactly what constitutes a perceptual basis for segmental substitution is yet to be clarified.

In two earlier studies of L2 production of sonorant consonants conducted by [17] and [18], it is found that acoustic similarities between L1 and L2 sounds serve as a basis for segmental substitution, and L1 coarticulatory effects between rhotic/nasal sounds and adjacent vowels play an important role in shaping the output form of L2 syllables.

In addition, [9] found that not all L2 production patterns were related to L1 transfer: some could be attributed to general articulatory constraints such as aerodynamic and elasto-inertial effects. In order to tease apart language-specific factors from universal articulatory factors in L2 production, they modified the SLM by proposing the Perception-Production Model (PPM) of L2 segmental production. The PPM inherits the part of the SLM involving equivalence classification in the formation of old, similar, and new categories of sounds in L2 production; while in the meantime, it adds a new part involving production constraints that shape learner's output during the articulatory planning stage.

Note that neither SLM nor PPM has gone beyond L1 substitution at the segmental level. However, recent psycholinguistic research suggests the possibility for L1 substitution to occur at the syllabic level. Specifically, [21] and [22] proposed the proximate units principle based on their finding that languages differ in units of phonological encoding in speech production. This principle claims that the speech processing unit is not universally based on the phonemic template: some languages use the syllable template as the basic processing unit. Therefore, any theories of word production necessarily take linguistic diversity into consideration ([21]).

Specific to the Chinese language, [22] found that syllables were the primary units of processing at the early stage of phonological encoding in Chinese word production. Also, [27] found that L1 production of Chinese-English bilinguals showed both syllable and segment effects in processing units, so did their L2 production. These findings suggest that the syllable, just like the segment, can be a basic unit of speech production, depending on which language is involved.

The above research has prompted the present study to re-analyze the production data in [17] from a new perspective. Unlike the analysis in [17], which focuses on segmental substitution, the new analysis focuses on whole syllable substitution and reveals that the presumed segmental substitution is actually a result of L1 transfer at the syllabic rather than segmental level in Chinese speakers' English production. In support of this new analysis, the present study has collected new data and compared them with the 2008 data. Also, it has proposed the Acoustic and Articulatory Model of L2 syllable production (Hereafter AAM for short) to explain the findings.

By introducing the AAM, the present study is intended to answer two questions that previous models of segmental production have left open: what role acoustic cues play in helping L2 learners to produce an L2 syllable and what role coarticulation mechanisms play in determining the output form of the L2 syllable.

To do so, this paper first presents previous and current experimental findings and analyzes the production process in Section 2. Then it explains the AAM in Section 3 and concludes the paper in Section 4.

2. EXPERIMENTAL DATA

This section presents previous and current experimental data in support of the AAM. Due to space limit, only the data involving English rhotic sounds are presented. Section 2.1 describes the data and associated data collection and processing methods; Section 2.2 presents and compares results derived from the 2008 and current data; Section 2.3 analyzes the crucial acoustic simulation process in the L2 production and provides evidence for the AAM.

2.1 Data description, collection, and processing

Both English and Mandarin Chinese have rhotic sounds, but they are phonologically and phonetically different. In English, the rhotic sound /ɹ/ (corresponding to the letter *r*) is an approximant and can follow alveolar stops to form a cluster such as /tɹ/ (as in *tr*) in syllable initial position, whereas this combination is illegal in Mandarin Chinese.

In order to investigate how Chinese L2 learners produce the novel English /ɹ/ in novel syllable position, [17] conducted an acoustic study of English words *tree* (/tɹi/) and *true* (/tɹu/) produced by Chinese speakers. The Mandarin word *qù* (/tɕ^{h}y/, 'go') was used to contrast the word *tree*. Also, the Mandarin word *chù* (/tʂ^{h}u/, 'touch') and the English word *chew* (/tʃ^{j}u/) were used to contrast the word *true*. In addition, native production of the English test words was also recorded for comparison. The speech data were collected from 19 undergraduate students of Chinese nationality from a Canadian university, and they are all international students coming from China. They produced each test word in isolation and repeated 4 times. Each word token was transcribed into IPA symbols based on its spectrogram display in the sound analysis software Praat ([4]). This can be done because different sounds have different spectral patterns. For example, the acoustic cue of /ɹ/ is a dipping F3, and the high, front vowel /i/ has a lower F1 and higher F2 than the high, back vowel /u/ ([16]). For affricates such as /ʧ, tʂ, tɕ/, it takes a little more effort to identify their fricative part, as their spectral patterns are not always distinct. Therefore, a main acoustic parameter of fricatives, COG (Center Of Gravity, a measure of the spectral energy concentration location in a spectrum) was used to help decide their place of articulation ([14]).

Figure 1 respectively illustrates a native English male speaker and a Chinese male speaker's production of *tree*. Their difference can be seen clearly in the two spectrograms. On the left, the voiced part of *r* (i.e., [ɹ]) shows up as a typical dipping F3. In contrast, the F2, F3, F4 all go up as the sound changes from /ɥ/ to /i/.

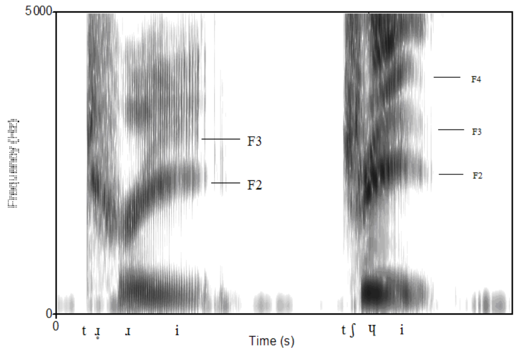

Figure 1: Spectrograms of *tree* produced by a native (left) and a Chinese (right) speaker

Figure 2 respectively illustrates a native male speaker's production of *true* (left) and a Chinese male speaker's production *true* (center) and *chù* (right). The difference can be seen clearly in the three spectrograms. On the left spectrogram, the voiced part of *r* (i.e., [ɹ]) gain shows up as a typical dipping F3. In contrast, the center and right spectrograms show that the fricative part for /ʧ/ (supposed to be *tr*) has less energy concentration at the lower frequency region (indicated by the two arrows) and hence appears lighter than that of /tʂ/.

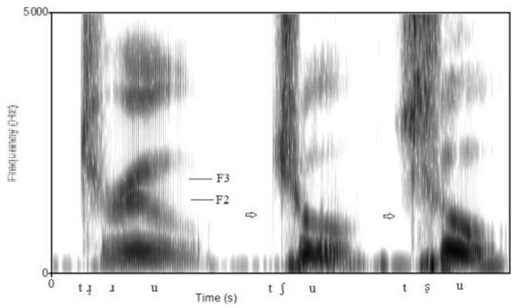

Figure 2: Spectrograms of *true* (left) produced by a native speaker and *true* and *chù* (center and right) produced by a Chinese speaker

As a supplement, the auditory judgment was performed by two trained phoneticians, one of them being the author and the other a native English speaker.

A similar method was used to collect current data from 27 undergraduate students in a Chinese university. Different from 19 speakers in the 2008 group, which all had at least one year experience of living and studying in Canada, 27 speakers in the current group have no oversea English learning experience. Therefore, their production may be less accurate than that of the 2008 group, but given the shared monosyllabic feature of the Chinese language, the present study assumes that Chinese speakers' production will be subject to the same syllable-based L1 processing constraint, regardless of speakers' individual language background.

2.2 Results

The production data of 2008 and current data were analyzed in terms of percentage of tokens associated with each type of production over all tokens of a test word. Results from the 2008 data are shown in Table 1.

Table 1: Results from the 2008 data

Token	Native-like production	Non-native production	
		English *ch*-like	Mandarin *ch*-like
Tree (/tɹi/)	25% ([tɹ̥i])	62.5% ([ʧᶣi])	12.5% ([tʂʷi])
true (/tɹu/)	12.5% ([tɹ̥u])	25% ([ʧᶣu])	62.5% ([tʂu])

Table 1 reveals that the production of English words *tree* and *true* is of three types: native-like (25% [tɹ̥i] & 12.5% [tɹ̥u]), English *ch*-like (62.5% [ʧᶣi] & 25% [ʧᶣu]), and Mandarin *ch*-like (12.5% [tʂʷi], 62.5% /tʂu/). The latter two types show that the substitution occurs not simply with English [ʧ] or Mandarin [tʂ] but also with a glide /ɥ/ or /w/ inserted in between the affricate and the vowel.

In [17], the insertion of an extra glide was interpreted merely as a way of simulating the rhotic sound and the process was considered segment-based; whereas the present study reinterprets it as a result of acoustic simulation of the whole syllable rather than individual segments and hence syllable-based. That is to say, the former interpretation assumes that simulation is a process of substitution on a segment by segment basis, whereas the current interpretation assumes that the substitution process is based on the whole syllable so that when determining which segment is to be used in L2 syllable production, the process does not care whether certain acoustic cues come from one segment or whether an L2 segment can be matched with an L1 segment. All it needs to do is find a segment or segment combination to recover these cues. Therefore, L2 learners may use different segmental combinations to realize the same L2 syllable, often at the price of changing the syllable structure. Glide insertion also occurs in the current data (see Table 2).

Table 2: Results from the current data

Token	Non-native production		
	English *ch*-like	Mandarin *ch*-like	Mandarin *q*-like
tree	56% ([ʧᶣi]) 15%([ʧᶣei])	22% ([tʂʷi])	7%[tɕᶣi]
true	82% ([ʧᶣu])	7%([tʂu]) 11%([tʂᶣu])	0

Results in Table 2 are somewhat different from those in Table 1, which is not surprising given the different groups of speakers and over ten years' time span. First, native-like production tokens are completely missing in the current data. As mentioned in Section 2.1, this is very likely due to the fact that all Chinese speakers in the current group lack of exposure to an English speaking environment. Second, new types of production emerge in the current data: 15% tokens of *tree* were rendered as [ʧᶣei] and 7% similar to the Mandarin word *qu* ([tɕᶣi]; 82% tokens of *true* were rendered as [ʧᶣu], similar to the English word *chew* (/ʧʲu/), and 11% as [tʂᶣu]. These new types are acoustically less similar than other types, again suggesting that speakers have insufficient language exposure or use.

Last, for the same types of production shared by the 2008 and current data, the proportion is different: it decreases from 62.5% to 56% for [ʧᶣi] and from 62.5% to 7% for [tʂu], whereas it increases from 12.5% to 22% for [tʂʷi] and from 25% to 82% for [ʧᶣu]. The most dramatic change occurs in the production of the word *true*: there are much fewer tokens of [tʂu] but much more tokens of [ʧᶣu] in the current data than in the 2008 data. This change suggests that over the years, palatalization has become a common strategy used to coarticulate the affricate and the high, back vowel /u/, perhaps more by Chinese speakers with less English proficiency and lower-level English exposure.

To sum up, it is difficult for Chinese speakers to produce the English *tr* accurately because of the novel sound combination, which can be viewed as a phonotactic constraint. To avoid the constraint, they tend to substitute it with an acoustically similar combination of affricates and glides. Moreover, which combination they use as a substitute does not seem to be random. The following section analyzes what factors may have helped shape the substitution patterns.

2.3 Analysis

The above results suggest that Chinese speakers tend to use Mandarin sounds and structure to simulate English words and the simulation process involves substitution and modification at both segmental and syllabic levels.

The non-native production of *tree* ([tɹ̥i]) as [ʧᶣi] reveals that an L2 syllable can be reconfigured based on salient acoustic cues. Here the rounded, high, front glide /ɥ/ is a Mandarin sound and its insertion is apparently for the purpose of simulating the lip-rounding feature of [ɹ]. Since the roundedness cue can be satisfied not only by /ɥ/ but also by /w/, using [tʂʷi] to substitute [tɹ̥i] is also viable and an even better substitute, as all three sounds exist in Mandarin.

Then, why did a majority of the speakers choose [ʧᶣi] over [tʂʷi] (56% vs. 22%)? In the former choice, the affricate [ʧ] is not even a Mandarin sound, but the speakers are able to produce it without any difficulty. A plausible explanation is that the former production must have been shaped by some coarticulation mechanism. It happens that a so-called strong anticipatory effect does occur in Mandarin syllable production, and it can cause a regressive assimilation process in place of articulation of the involved segments ([8], [26]). Given that the three sounds in [ʧᶣi] have a place of articulation all at the front, this regressive place assimilation mechanism must have been at work in the production of *tree* and may turn [tʂʷi] further into [ʧᶣi]. The whole process can be summarized as follows: [tɹ̥i] as intake → [tʂʷi] as closest substitution → [ʧᶣi] as a result of regressive place assimilation.

Unlike *tree*, the production of *true* ([tɹ̥u]) does not need to insert a rounded glide for the purpose of simulating the lip-rounding feature of [ɹ], as the following rounded vowel /u/ can serve the same purpose. Then why do a large proportion of tokens (82% [ʧᶣu]) from the current group still have the glide /ɥ/ inserted? As mentioned in Section 2.2, palatalization has become a trendy strategy to coarticulate front consonants and back vowels. Based on [13], it is actually a general coarticulatory mechanism which helps alleviate the conflict between anterior and posterior tongue gestures. That is to say, when speakers coarticulate alveolar or post-alveolar affricates like [tʃ] and [tʂ] and the high, back vowel /u/, they tend to insert a palatal glide in order for the tongue gesture to transit easily from the front to the back.

Another piece of evidence comes from the native production of the word *chew*, which is transcribed as /tʃu/ in dictionaries but commonly produced as [tʃʲu] instead. Apparently, the glide /ʲ/ is inserted for ease of articulation.

As for the question why most tokens of *true* in the current data use non-rhotic [tʃ] (82% [ʧᶣu]) rather than rhotic [tʂ] to substitute the rhotic /tɹ/, the answer again lies in the Mandarin specific anticipatory effect: the speakers may start out producing the word as [tʂu], which is acoustically close to the Mandarin word *chù*, and then insert the palatal glide /ᶣ/ to ease the articulation, turning the word into [tʂᶣu], but since the palatal glide has a front articulation, it causes regressive assimilation in the affricate, turning it into [ʧ]. The whole process can be summarized as follows: [tɹ̥u] as intake → [tʂu] as closest substitution → [tʂᶣu] for ease of articulation → [ʧᶣu] as a result of regressive place assimilation.

So far, the above analysis shows that acoustic simulation is not a simple segment-by-segment substitution process: its output can be constrained by both general and the Mandarin specific coarticulatory mechanism, that is, regressive place assimilation.

3. THE PROPOSED MODEL

The above section has provided evidence in support of the syllable-based acoustic simulation process. This section illustrates the process in the proposed AAM.

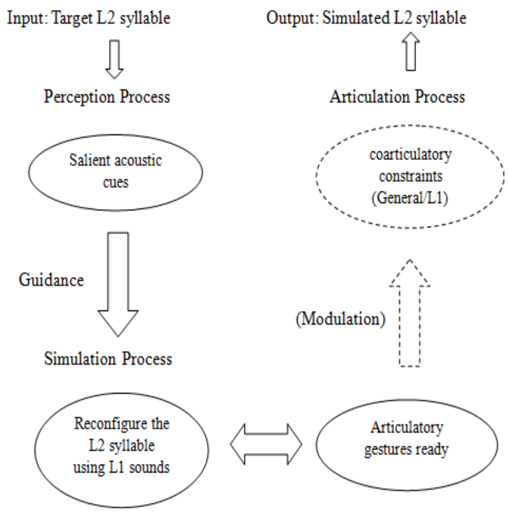

Figure 1: The Acoustic-Articulatory Model of L2 Syllable Production

As shown in Figure 1, the whole process starts from input of a target L2 syllable and ends with output of a simulated L2 syllable. It includes two components: acoustic and articulatory.

The acoustic component is subsumed under the perception process and it involves associating acoustic cues of an L2 syllable with an L1 syllable. Specifically, it includes two stages: the first stage is for L2 learners to capture acoustic cues salient to them from the input signal; the second stage is the simulation process. According to [24], acoustic cues serve as intake for further processing. Guided by acoustic intake, the simulation process is responsible for reconfiguring the target L2 syllable using available L1 sounds and also making sure that the syllable conforms to L1 phonotactic conventions.

The articulation component is responsible for coordinating articulatory gestures and preparing them for production. It also includes two stages: the first stage is to prepare gestures required of realizing the reconfigured syllable; the second stage is to subject relevant gestures to the modulation effect of general and L1 specific coarticulatory mechanisms.

Note that the output form is not solely influenced by coarticulatory constraints (hence the dotted shapes in Figure 1), as [18] found that prosodic constraints may be also involved. In addition, previous language experience can influence how speakers perceive an L2 syllable, and inaccurate perception will affect subsequent processing, resulting in production errors ([10], [20]).

As for how exactly perception affects production, recent psycho- and neuro-linguistic evidence ([15], [23]) has pointed to a non-conventional, direct-realist view of speech perception ([12], [25]). This view assumes a direct association between perception and articulation, so in L2 production, for example, speakers will perceptually assimilate articulatory gestures underlying L1 and L2 phonemes rather than phonemes themselves ([3]).

The present study acknowledges this alternative view and considers it compatible with the proposed AAM: whether the syllable is represented by a combination of phonemes or gestures, the premise that the syllable is the basic unit for perceptual assimilation will not be undermined. If the direct-realist view is adopted in the AAM, then acoustic simulation becomes acoustically informed articulatory simulation and the whole process needs to be adjusted accordingly. Limited by the current research focus, the present study will leave it to future work.

In contrast to the PPM, the AAM is constructed based on the notion of L1 substitution at the syllabic level; it treats the syllable as the basic unit of speech perception and assumes an acoustic simulation process different from that of the PPM: substitution is syllable- rather than segment-based from the very beginning, and it aims at recovering acoustic cues rather than retrieving individual segments contained in the whole syllable. Therefore, L2 learners do not need to differentiate individual L2 segments and try to match them one by one with L1 segments. Instead, they just need to find L1 segments or segment combinations that can emit the acoustics cues they hear as a whole. In the production of *tree*, for example, learners use glide insertion simply to recover the acoustic cue associated with lip-rounding. They do not substitute /ɹ/ directly with a similar L1 sound, indicating that their goal is to simulate the native production of the whole L2 syllable based on acoustic cues, whatever particular L2 segments emit the cues. It is in this sense that the simulation process is considered syllable rather than segment-based. As a result, equivalence classification between L1 and L2 segments assumed by the PPM does not necessarily occur in L2 speech learning of Chinese speakers.

Concerning the respective roles of acoustic cues and coarticulatory mechanisms in L2 syllable production, the proposed AAM assumes that the former is to provide guidance for syllable reconfiguration and the latter is to modulate articulatory gestures for final production. It is due to the dynamic nature of coarticulation that segmental substitution patterns vary. The present study considers these roles essential in L2 syllable production and hopes that future research can provide more data in support of the AAM.

4. CONCLUSION

This paper has shown that an L2 syllable can be simulated as a whole under the guidance of acoustic cues and the associated articulatory gestures can be modulated not only by general but also by language specific coarticulatory mechanisms. The AAM has summarized the whole process of L2 syllable production and provided a new perspective for the research of L2 speech acquisition.

5. ACKNOWLEDGES

This work is supported by the 10th Chinese Foreign Language Education Fund, No. ZGWYJYJJ10A122.

6. REFERENCES

[1] Archibald, J. 2005. Second language phonology as redeployment of L1 phonological knowledge. *Canadian Journal of Linguistics* 50 (3), 285-314.

[2] Barrios, S., Jiang, N., Idsardi, W. J. 2016. Similarity in L2 phonology: Evidence from L1 Spanish late-learners' perception and lexical representation of English vowel contrasts. *Second Language Research* 32, 367-395.

[3] Best, C. T., Goldstein, L. M., Nam, H., Tyler, M. D. 2016. Articulating what infants attune to in native speech. *Ecological Psychology* 28 (4), 216-261.

[4] Boersma, P., Weenink, D. 2015. *Praat, a System for Doing Phonetics by Computer*. http://www.praat.org [version 6.0.05].

[5] Brown, C. 1998. The role of the L1 grammar in the L2 acquisition of segmental structure. *Second Language Research* 14 (2), 136-193.

[6] Cebrian, J. 2002. *Phonetic Similarity, Syllabification and Phonotactic Constraints in the Acquisition of a Second Language Contrast.* Doctoral dissertation, University of Toronto, Canada.

[7] Chang, C. B. 2015. Determining cross-linguistic phonological similarity between segments: The primacy of abstract aspects of similarity. In: Raimy E., Cairns C. E. (eds.). *The Segment in Phonetics and Phonology*. John Wiley & Sons, Inc. 199-217.

[8] Chen, X. 1997. Study of coarticulation of Mandarin segments. *Studies of Chinese Language* 5, 345-350.

[9] Colantoni, L., Steele, J. 2008. Integrating articulatory constraints into models of second language phonological acquisition. *Applied Psycholinguistics* 29 (3), 489-534.

[10] Escudero, P. 2009. Linguistic perception of 'similar' L2 sounds. In: Boersma P., Hamann S. (eds.). *Phonology in Perception*. Berlin: Mouton de Gruyter. 151-190.

[11] Flege, J. E. 1995. Second language speech learning: Theory, findings and problems. In: Strange, W. (ed.), *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*. Timonium: York Press. 233-277.

[12] Fowler, C. 1986. An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics* 14, 3-28.

[13] Gick, B., Wilson, I. 2006. Excrescent schwa and vowel laxing: Cross-linguistic responses to conflicting articulatory targets. In: Goldstein, L., Whalen, D. H., Best, C. T. (eds.), *Papers in Laboratory Phonology 8: Varieties of Phonological Competence*. Berlin, New York: Mouton de Gruyter. 635-660.

[14] Jongman, A., Wayland, R., Wong, S. 2000. Acoustic characteristics of English fricatives. *J. Acoust. Soc. Am.* 108 (3), 1252-1263.

[15] Kearney, E., Guenther, F. H. 2019. Articulating: The neural mechanisms of speech production. *Language, Cognition and Neuroscience* 34 (9), 1214-1229.

[16] Ladefoged, P., Johnson, K. 2011. *A Course in Phonetics (6th edition)*. Boston, USA: Wadsworth, Cengage Learning.

[17] Li, Y. 2008a. The phonetically based segmental substitution and syllable modification in L2 speech: Acoustic evidence from Chinese speakers' production of English *tr* and the dark *l*. *Proc. 8th PCC*, Beijing.

[18] Li, Y. 2008b. *Mandarin Speakers' Production of English and Mandarin Post-vocalic Nasals: An Acoustic Approach.* MA thesis. University of Victoria, Canada.

[19] Li, Y., Cai, J. 2018. Speech transfer research: Theoretical models and Empirical studies. *Second Language Learning Research* 1, 113-130.

[20] Major, R. C. 2001. *Foreign Accent: The Ontogeny and Phylogeny of Second Language Phonology*. Mahwah, NJ: Lawrence Erlbaum Associates.

[21] O'Séaghdha, P. G. 2015. Across the great divide: Proximate units at the lexical-phonological interface. *Japanese Psychological Research* 57 (1), 4-21.

[22] O'Seaghdha, P. G. Chen, J., Chen, T. 2010. Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition* 115, 282-302.

[23] Skipper, J. I., Devlin, J. T., Lametti, D. R. 2017. The hearing ear is always found close to the speaking tongue: Review of the role of the motor system in speech perception. *Brain and Language* 164, 77–105.

[24] Vanderweide, T. 2005. *Cue-based Learning and the Acquisition of Pre-vocalic Clusters.* Doctoral dissertation, University of Calgary, Canada.

[25] Whalen, D. H. 2019. The motor theory of speech perception. *Oxford Research Encyclopedia of Linguistics*.

[26] Wu, Z., Sun, G. 1989. Experimental study of coarticulation of unaspirated stops in Mandarin CVCV syllables. *Report of Phonetic Studies of Chinese Academy of Sciences.*

[27] Xin, X., Lan, T., Zhang, Q. 2020. Assimilation mechanism of the phonological encoding process in L2 speech production of Chinese-English bilinguals. *Acta Psychologica Sinica* 52 (12), 1377-392.

LI Ya PhD She is an associate professor of Foreign Language Department, Zhejiang Ocean University. Her research interests include speech prosody, L1 & L2 language acquisition, and Chinese tones.
E-mail: liyawy@zjou.edu.cn