Attention Computation in the Processing of L2 Cognition: Revisiting Learning Models

Yizhou Lan[1] and Will Xiangyu Li[2]

[1]School of Foreign Languages, Shenzhen University, Nanhai Avenue, Shenzhen 518000, China
[2]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China

Email: will.x.li@ieee.org

Abstract. Studies show that the way our brain processes incoming speech sounds has a lower-level grounding derived of acoustic similarity. Previous theoretical models of speech sound processing posit that higher-level cognitive process plays little role in perception and in successful and complete processing of speech sounds. The present study investigates if such models may be effectively extended to incorporate influences from higher level cognitive cues, such as voluntary attention, to certain acoustic dimensions of the speech sound stimuli. In this paper, we investigate the relationship in a qualitative way between the efficiency of the language processing and high-level perceptual mechanism through computational simulation of speech perception, and accuracy and reaction-time measurements. The results of experiments lead to an enhancement of existing statistical signal processing and perception models' predictions. Our findings revealed that acoustic similarity in speech sound signals merely does not accurately predict the acquisition outcome, and the enhancement of natural language learning can be achieved by effectively mining out the auxiliary cognitive cues in these signal processing activities.

1. Introduction

It has long been predicted in human speech processing research that human auditory perception of foreign sounds is of low-level processing in our cognition, without using higher-level cognition like consciousness and emotions [1] [2]. The voluntary attention to specific information during speech sound perception was considered irrelevant to human's processing of non-native, or second language (L2) phonetics. In psychological studies, many researchers consider language acquisition a process of statistical learning [3]. Theories of speech learning followed the footsteps of the discovery of distributive learning. The best representative is Catherine Best's PAM, or perceptual assimilation model [4] [5] [6]. It is the first model that explicitly predicted speech processing output of L2 learners by perceptual distance. Given 2 incoming candidates of non-native speech stimuli that learners need to compare and match with L1, Best [4] predicted altogether six possible outcomes. They are the final fixations of sound perception after that learners compare these L2 candidates to L1 ones. These six types are *Two Category* (TC), *Uncategorized-Categorized* (UC), *Category Goodness* (CG), *Both Uncategorizable* (UU), *Single Category* (SC),and *Non-linguistic Assimilation* (NA) types [5]. According to the model, phonetic categories are formed in infancy and they can interact with L2 categories when brain learns a second language in a 2-dimension perceptual map. When a speech

sound is given to an infant repeatedly, certain acoustic features of the incoming sound stimuli were then strengthened in multiple tokens in its memory, enhancing the magnetic power of that specific sound, which makes other similar incoming sounds the same perceptual category to that infant. The specifically strengthened sounds were regarded as prototypes or magnets, where incoming stimuli will become non-discernible because of the ease of categorizing it as the strong magnet. Extending the model to L2 learners, to add new L2 magnets into an L2 learner's cognition will be extremely difficult because the many existing magnets in L1 are potent magnets that will attract any similar incoming tokens, thus leading to stagnation of the learner's perceptual ability.

When two L2 sounds are assimilated into the same first language category, second language learners who cannot distinguish subtle acoustic differences face two possible categories. For example, when looking into L2 English learners from the Chinese background, Mandarin speakers to be specific. When they hear /r/ in English, it immediately associates with the native Mandarin sound /ʐ/, and that falls into a SC type where /r/ cannot be heard and learned because of the close distance between these two sounds [7] [8] [9] [10]. Although the model is highly explanatory, no computational component is available for attempting quantitative learning predictions.

To further computerize the prediction of learning outcomes, a mathematical model describing the processing result has been published by Guenther and Gjaja [11] [12]. The studies contained simulations modelling non-native precepts such as the extent to which a non-native percept is heard and correctly and efficiently processed may be strengthened by the repeated feeding of data that consolidates the relevant the activation of certain parts of our brain or facilitate perceptual patterns of similar auditory processing tasks. Other studies suggest that the acquired L2 categories may also have magnetic effect [13] [14] [15]. Eventually, the learning will attain equilibrium because a steady L2 category is concrete enough to be independent from the influence of L1 category's magnetic effect.

However, more recent studies, with the aid of technological advances, are interested in the directly measuring of the affective aspects in non-native sound processing. Such studies have shown that only a simple distance-attractive system cannot suffice in predicting language acquisition results, and voluntary attention may play a part [16] [17] [18]. These studies suggest that when learners pay better focus, the acoustic features from the non-native speech may be heard with improved accuracy. A few recent studies following the ones above are starting to focus on the higher-level cognitive processes displayed in the neurological process of non-native perceivers when they process L2 speech sounds [1] [18] [19] [20]. These studies has given good reference to the present study that investigates acoustic perceptual distance and attention.

2. Methods

The current investigation explores the effect of human cognitive/affective factor of voluntary attention on the outcome of speech perception in the form of behavioral experimental study and simulation. There were 24 native English speakers and 24 speakers speaking Hong Kong Cantonese. Each group consisted of 12 male and 12 female participants with an average age of 22.5. Both groups of speakers were free from deficits, and both studied English for an average duration of 10 years. They are informed of the experimental procedure prior to the study, with written consent obtained from participants. Table 1 shows the exact confusion pairs in the study. All these sound pairs are SC types which are difficult for that group of learners to learn. In all, 6 pairs of perception test are done.

Table 1. Phonemic contrasts of Cantonese, Mandarin and English that are of poor discrimination predicted by PAM.

	Cantonese	**English**
Cantonese	N/A	/w/ and /r/
Mandarin	/ʐ/ and /j/	/ʐ/ and /r/
English	/θ/ and /f/	N/A

2.1. Sound Materials

Stimuli of the behavioral study include pseudo-words with the target consonants /r/-/w/-/ɻ/ and /θ/ -/f/-/s/ paired and embedded in the CVC syllable structure closed by /t/, with vowel varying among or in /i, ɑ, u/ conditions. The synthesizer uses the automatic synthesizing function to create artificial speech sounds for listeners to identify impartially so that learners can be exposed by English sounds attuned to exact parameters we set for training. Formant frequency is the main acoustic property used to discern sonorants, i.e., vowels and approximants (e.g. /r/-/w/) in the temporal domain. Spectral moments, however, is the main sibilant, or non-sonorant acoustic property to measure the frequency of the hissing sounds in the spectral domain [21]. The sound syntheses were done individually through Praat in its Klatt synthesizer function [22]. For approximant sounds, i.e. /r/ and /w/, the primary factor to distinguish them is formant frequencies [23].

2.2. Procedure

In the speech recognition (identification and discrimination) task, participants were asked to pick out the correct stimulus based on its own native phonemes and to rate whether the stimulus was a good, or same example of the category on a scale of 1(bad) to 7(good). And calculate the average goodness of the category. The total number of experiments for each participant is 8 consonants ×3 vowels ×5 repetitions = 120. The identification test is divided into four stages because of the large number of trials, which can lead to fatigue among participants.

In the discrimination task, participants faced three pairs of stimuli - the choice discrimination task. Inter-stimulus interval (ISI) was 300 milliseconds. Adding the same amount of filler to the experiment, a total of 960 stimuli were compiled into 320 trials, each consisting of three stimuli. No stimulus is paired to itself. The trial was divided into four stages. The process places the stimulus in two-dimensional Euclidean space so that the perceptually similar stimuli are placed together, while different stimuli are placed far away, using the inverse monotonic function to maximize the fitting between distance and similarity. When the experiment starts, participants hear groups of trials consisting of three stimuli and need to judge whether the last stimulus is similar to the first or second one, each contributing to 50% of chance. A practice session is placed before the trial sessions to help the participants understand the structure and procedure of the test. The responses to the practice session were not analyzed.

The discrimination task is repeated after the attention of the learners to acoustic data is instructed to be attuned to the sensitive acoustic aspects of the sounds by listening to the change of acoustic parameters separately with corrective feedback. Different from previous studies which mainly used covert stimuli-response training with largely the same testing material and procedure, the new protocol utilizes meta-cognitive methods to instruct to learners overtly about the sensitivity the sound correlates, and the articulatory movements the acoustic features correspond to. The instruction is done by an experienced linguist to ensure correctness. The overall discrimination accuracy and response time of listeners before after the shifting-of-attention procedure are reported in the next section.

3. Results

Results primarily report the Cantonese and English speakers' performances on the tasks after the attuning of attention. For /w/-/r/-/ɻ/ and /f/-/θ/-/s/ contrast respectively, the below findings are evident:

For the /w/-/r/ contrast, the higher F3 accuracy results has indicated that, near the /w/-/r/ boundary, English speakers are more sensitive than Cantonese ones in terms of F3. Such high sensitivity only exists near the boundary of this contrast for English speakers. Averaged for both contrasts, although the accuracy levels are significantly higher near category boundaries than within the boundaries for both groups of speakers, the levels of such significance are much higher for English ($p<.001$) than for Cantonese ($p<.05$). The sensitivity of cantonese listeners to English /r/ ($p<.001$) and /w/ ($p<.05$) was significantly higher than that of cantonese listeners. English listeners have always been highly sensitive on the phoneme boundary between the two categories ($p<.001$). However, for /r/ and /ɻ/, both

Cantonese and English speakers show similar pattern of accuracy variations by both attuning their perceptual salience to F2 (p=.134).

For the /f/ and /θ/ pair, Cantonese listeners show greater sensitivity to M2, which is not a decisive factor for the distinction of /f/ and /θ/; and are less sensitive to the M1 dimension (p<.001). Because the M1 feature has higher sensitivity than M2, it is implied that M1 has the bigger influences towards the perceptual accuracy in this sound contrast, and can thus be seen as the perceptual cue of these fricatives for Cantonese speakers. For English speakers, however, the M1 dimension is more sensitive (p<.001). Again in terms of acoustic dimensions, a mismatched pattern is shown for Cantonese and English speakers. For /s/ and /θ/, both English and Cantonese speakers nevertheless show sensitivity to M1.

General accuracy rates after the attuning of attention for the four chosen sound pairs are shown in Figure 1, layouting two types of voice contrasts. The four groups of figures (a)(e),(b)(f), (c)(g) and (d)(h) respectively represent the perceptual accuracies of Cantonese speakers in /r/-/ɻ/, /w/-/r/, /s/-/θ/ and /f/-/θ/ from top to bottom. Within confidence levels of 1-7, 4 represents estimated perceptual category boundary. The four groups of figures (i)(m), (j)(n), (k)(o) and (g)(p) respectively represent the perceptual accuracies of English speakers in /r/-/ɻ/, /w/-/r/, /s/-/θ/ and /f/-/θ/ from top to bottom. Within confidence 1-7, 4 represents estimated perceptual category boundary. The two-dimensional perceptual maps use thermos-display to further indicate perceptual accuracy.

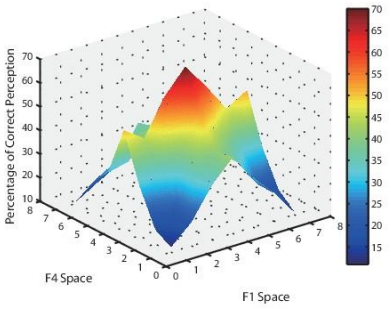
Overall, the perceptual accuracy patterns of the two speaker groups on both sound contrasts are composed below in Figure 1. the results for /r/ and /ɻ/ in Figures 1-(a)(e) and (b)(f) shows that speakers from Hong Kong and America had results of acquisition that are similar. The rise of accuracy rates for both Cantonese and English speakers towards the category center show that they are clearly two separate sounds. However, in Figures 1-(c)(g) and (d)(h), which cover the /w/ and /r/ pair, the results are not in line with those of English speakers in that the accuracy is not changing significantly with F3 but with F2. Even at the center of categorization by English speakers, Cantonese speakers only attain an accuracy rate slightly above chance. This shows that the category formation of /w/ and /r/ is sensitive to F2 instead of F3.

The accuracies for /s/-/θ/ and the /f/-/θ/ pairs are to be explained similar as above. For the /f/-/θ/ pair in Figures 1-(i)(m) and (j)(n), Cantonese has more sensitive dental features /θ/ than English who consider /s/ more similar to Figure 1-(k)(o) and (l)(p). Considering the above, we can discover that attention to calculated facets of the acoustic stimuli can and do influence perceptual accuracy.

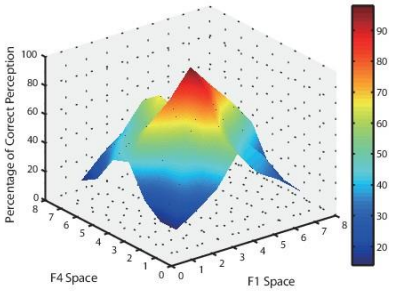
4. Discussions

We may conclude from the clear contrast between the results before and after ATS that non-native perceivers' including of attention has a large bear on L2 sound processing. Thus, one thing is clear: the by the correct computation and exploitation of key acoustic feature and stimulating attention, the processing accuracy of non-native sounds can be improved.

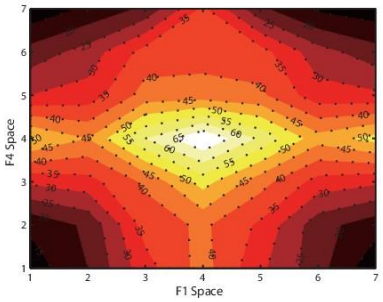
A further theoretic implication from the study posits that L2 learning is to be modeled in a multi-dimensional system, where a simple distance-based algorithm may not suffice. In his well-know complexity theory of language learning, Larsen-Freeman [24] held the view that one language production task derived from a simple mapping may generate numerous unexpected patterns in real-time phenomena, since the rule mappings may be affected by other attractors seemingly random to the issue. In the present study, Cantonese speakers' accuracy data clearly shows that the seemingly over-simplistic feature components (such as formant frequencies, spectral moments or vowel conditions [22] [23]) in one sound contrast may result in huge perception discrepancies.


Percentage of Correct Perception
F4 Space
F1 Space

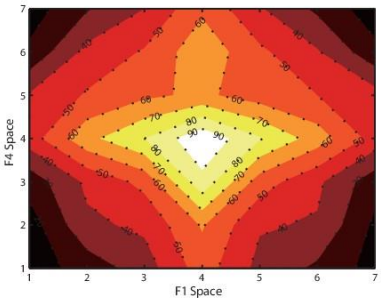
(a)


Percentage of Correct Perception
F4 Space
F1 Space

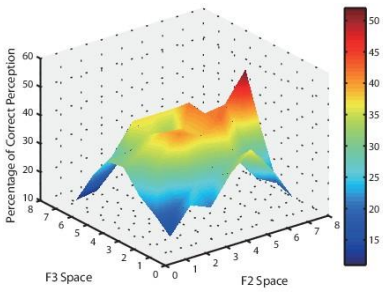
(b)


F4 Space
F1 Space

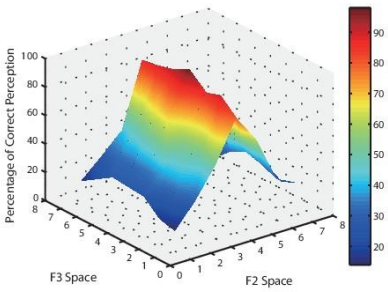
(e)


F4 Space
F1 Space

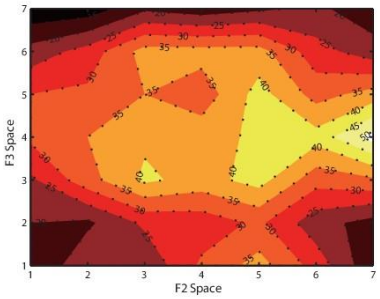
(f)


Percentage of Correct Perception
F3 Space
F2 Space

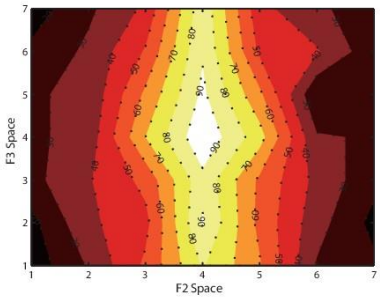
(c)


Percentage of Correct Perception
F3 Space
F2 Space

(d)


F3 Space
F2 Space

(g)


F3 Space
F2 Space

(h)

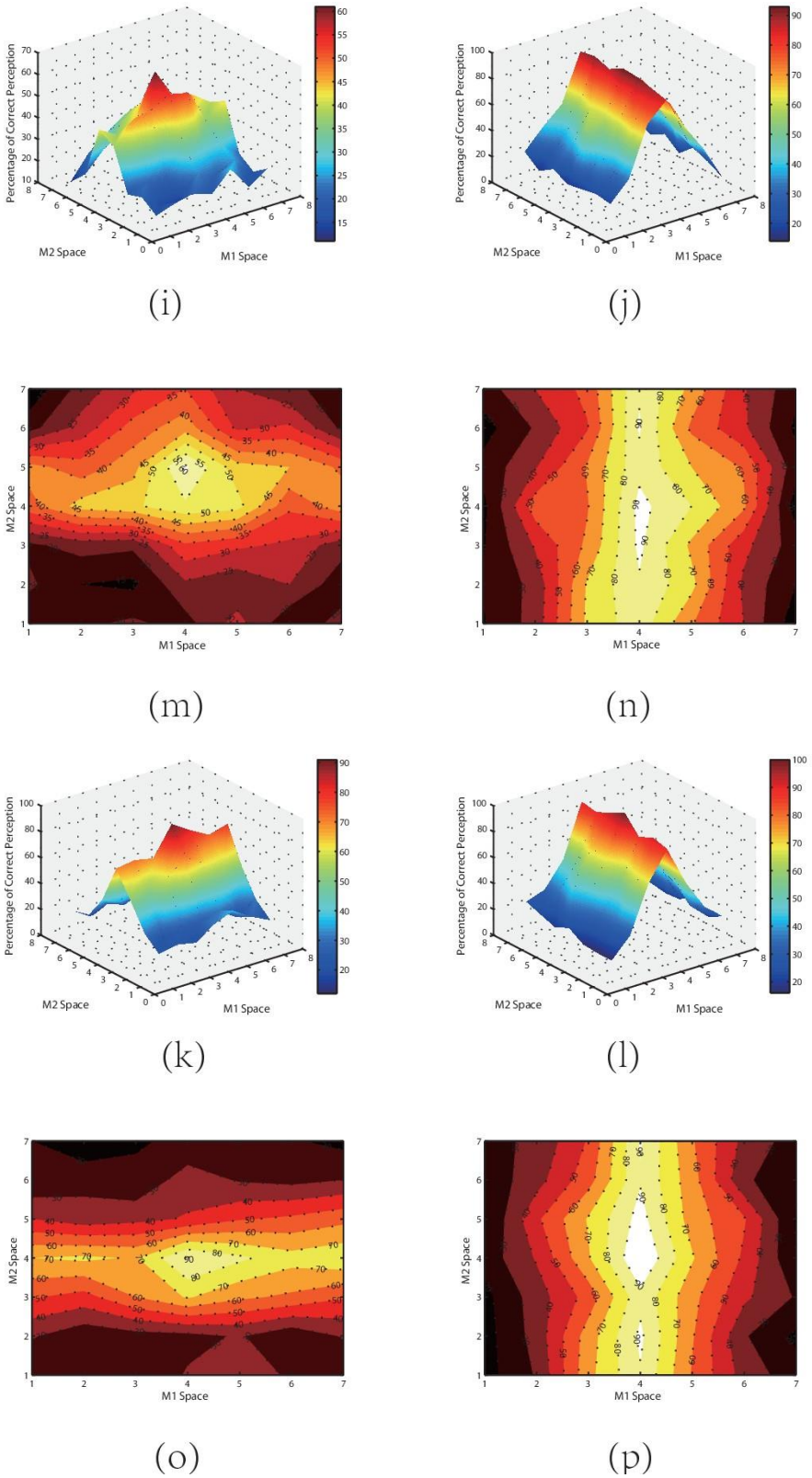


Figure 1. Percentage of correct perception with confidence (1-7) of chosen stimuli by Cantonese speakers in upper eight panels, and percentage of correct perception with confidence (1-7) of chosen stimuli by English speakers in each cell in the lower eight panels.

5. Conclusion

The present study has examined the role of human cognitive/affective factor of voluntary attention in L2 speech acquisition by utilizing behavioral experiments and computational analyses, and has proven the existence of the attentional effect through witnessing the learners' distinct fluctuation of L2 accuracy when presented with programmed formant/spectral frequency variations. We have demonstrated that the brain's higher-level processing functions play an important role in human cross-linguistic perception. We supplemented PAM's prediction by measuring behavioral traits under various attention conditions, and extended the prediction of the model to appropriately include the determinants of autonomous attention [17] [18]. The role of intentional attention in the quantitative investigation of computational analysis of the perceptual accuracy of data will lead to the revision of current learning models.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant No. 61601226, and in part by Natural Science Foundation of Jiangsu Province of China under Grant No. BK20160850. This work was also supported in part by the Shenzhen University Young Scientist Start-up Fund. Part of the preliminary findings were presented at the 2018 IEEE ICSPCC conference.

References

[1] Diehl R L, Lotto A J and Holt L L 2004 Speech perception *Annual Review of Psychology* ***55*** pp 149-179
[2] Pisoni D B 1985 Speech perception: Some new directions in research and theory *The Journal of the Acoustical Society of America* ***78*****(1)** pp 381-388
[3] Baddeley A 1992 Working memory: The interface between memory and cognition *Journal of Cognitive Neuroscience* ***4*****(3)** pp 281-288
[4] Best C T 1995 A direct realist view of cross-language speech perception, *In Strange W (ed.). Speech perception and linguistic Experience: Issues in Cross-Language Research* pp 171-204
[5] Best C T, McRoberts G W and Goodell E 2001 Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system *The Journal of the Acoustical Society of America* ***109*****(2)** pp 775-794
[6] Best C T, & 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* pp 13-34
[7] Chang Y C, Hong J and Hallé P 2007 English cluster perception by Taiwanese Mandarin speakers *ICPhS XVI* pp 797-800
[8] Deterding D 2006 The pronunciation of English by speakers from China *English World-Wide* ***27*****(2)** pp 175-198
[9] Flege J E, Munro M J and MacKay I R 1995 Factors affecting strength of perceived foreign accent in a second language *The Journal of the Acoustical Society of America* ***97*****(5)** pp 3125-3134
[10] Flege J E and Liu S 2001 The Effect of experience on adults' acquisition of a second language *Studies in Second Language Acquisition* ***23*****(4)** pp 527-552
[11] Guenther F H and Gjaja M N 1995 The perceptual magnet effect as an emergent property of neural map formation *Working Papers of Boston University Center for Adaptive Systems and Department of Cognitive and Neural Systems* pp 1-31
[12] Guenther F H and Gjaja M N 1996 The perceptual magnet effect as an emergent property of neural map formation *The Journal of the Acoustical Society of America* ***100*****(2)** pp 1111-1121
[13] Kuhl P K and Iverson P 1995 Linguistic experience and the "perceptual magnet effect" *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* ***39*****(4)** pp 121-154

[14] Kuhl P K 2000 A new view of language acquisition *Proceedings of the National Academy of Sciences* ***97*****(22)** pp 11850-11857
[15] Iverson P, Kuhl P K, Akahane-Yamada, R Diesch E, Tohkura Y I, Kettermann A and Siebert C 2003 A perceptual interference account of acquisition difficulties for non-native phonemes *Cognition* ***87*****(1)** pp B47-B57
[16] Golestani N, Paus T and Zatorre R J 2002 Anatomical correlates of learning novel speech sounds *Neuron* ***35*****(5)** pp 997-1010
[17] Strange W 2011 Automatic selective perception (ASP) of first and second language speech: A working model *Journal of Phonetics* ***39*****(4)** pp 456-466
[18] Francis A L and Nusbaum H C 2002 Selective attention and the acquisition of new phonetic categories *Journal of Experimental Psychology: Human Perception and Performance* ***28*****(2)** pp 349-366
[19] Lan Y and Li W XY 2014 Personality, category, and cross-linguistic speech sound processing: a connectivistic view *The Scientific World Journal* pp 1-7
[20] Setter J 2010 *Hong Kong English* (Edinburgh: Edinburgh University Press)
[21] Kent R D and Read C 2003 The *Acoustic Analysis of Speech* (Albany: Delmar Learning)
[22] Li F, Munson B, Edwards J, Yoneyama K and Hall K 2011 Language specificity in the perception of voiceless sibilant fricatives in Japanese and English: Implications for cross-language differences in speech-sound development *The Journal of the Acoustical Society of America* ***129*****(2)** pp 999-1011
[23] Klatt D H and Klatt L C 1990 Analysis, synthesis, and perception of voice quality variations among female and male talkers *The Journal of the Acoustical Society of America* ***87*****(2)** pp 820-857
[24] Larsen-Freeman D 1997 Chaos/complexity science and second language acquisition *Applied Linguistics* ***18*****(2)** pp 141-165