Towards a Revised Motor Theory of L2 Speech Perception

Yizhou Lan
Department of Chinese, Translation and Linguistics
City University of Hong Kong
Tat Chee Avenue, Kowloon, Hong Kong
ylylan2-c@my.cityu.edu.hk

Abstract

This study aims to review, through experiment proof of a salient effect of articulatory gestures on L2 perception, the time-honored but still put-to-sideways motor theory of speech perception. On one hand, previous studies in support to motor theory were largely done by tests of mismatch in duplex perception of acoustic/speech data; or by L1 development observations. On the other hand, L2 learning studies had seldom followed the motor theory framework. The current study employed two experiments on experienced L2 English speakers from a Cantonese L1 background to finish discrimination tasks on both 1) same allophone [tr] and [tʃ] but with different gestural overlapping in real words 2) the crucial acoustic cue of distinguishing the gestural differences of the same contrast by native speakers in isolation -- namely, the CV transitions. Results showed that non-native speakers could perform native-like in experiment 2 but not in experiment 1. Though both experiments contain the same acoustic information, only experiment 1 contains the entire gestural information. It is concluded that, at least, errors in second language acquisition has a gestural basis, which might partly support the motor theory from a new perspective.

1 Introducing the theoretic dispute

Acoustic-based perception mechanisms claim that human speech is perceived by a psycho-acoustic device which is capable of normalizing incoming sound tokens and extracting acoustic cues from acoustic sounds to form phonological categories (Pisoni, 1985; Kuhl, 2000). But the myth these theories failed to give explicit clarification to lies in the multiplicity and high variability of acoustic signal in one same percept of speech sound. Upon this possible discrepancy, it is suggested by motor theorists that the human percept for speech sounds lies in the articulatory gestures and production is based on that accordingly (Liberman and Mattingly, 1985).

Inconsistencies between the two theories of speech perception lie in what the primitive percept of speech is and the nature of processes of perception are. Acoustic perception theorists insist that human beings actively detect the acoustic information in the flow of speech, which is recognized as speech sounds. In motor theory, however, sound waves are but the product of intended articulatory gestures, which constitute an independent "language module". In terms of process, the acoustic perception of speech inevitably introduces two systems consisting of phones, the physical property of acoustic signals; and phonemes, the mental representation or classification of meaningful sound units (Ladefoged, 1993). However, the motor theory believes that we only perceive speech sounds (not other acoustic signal) through gestures because only linguistic sounds own gestural properties.

Despite the difference, an important common ground shared by both models is that both models separate phonetics (physical stimuli) and phonology (mental representation) with different instruments. For acoustic models, the two systems are separated by two levels of processing; for motor theory, a completely torn-apart module was introduced by claiming that the ability to detect gestures is "purely linguistic" and differs from acoustic perception fundamentally (ibid.).

Previous studies supporting the motor theory of speech perception had largely adopted the methodology of duplex perception (Rand, 1974; Whalen & Liberman, 1987) to show that segmentation of speech sounds by using acoustic detail is not plausible for human language perception because experiments has shown that humans perceive CV transitions (primarily stops

27th Pacific Asia Conference on Language, Information, and Computation pages 136－142

and fricatives) in speech sounds (part of a word) more accurately and context-dependent than non-speech acoustic sounds, like bird chirps.

More recent studies on animal perception of language (Kuhl, 2000) provided arguments against the motor theory because the ability to perceive gestures, as it was put, can also be captured by other mammals. On its basis, Best (1995) brought forward another gesture-based theory of speech perception entitled the direct-realist view. Its basic viewpoint, different from the motor theory, is that language perception is not innate, because although without intended gestures, other animals can still distinguish human vowels. Rather, human beings perceive speech by generalizing others' gestures, no matter he or she have such knowledge of gesture.

Even so, the direct realist theory faces two challenges. Firstly, it did not specify what are the gestures being utilized as categories, not like motor theory's predecessors' work with articulatory gestures (Browman and Goldstein, 1987, 1992), and is inherently phonemic. The other limitation is that it did not fully explain how sounds are learned, although there are hints that it was through frequency-based statistical learning. Maybe the cause was the fear to be labeled another auditory-based theory, because statistical learning of speech sounds is inherently normalization of psycho-acoustic data. Both challenges cannot be resolved by only using L1 data. The reason is shown in the section below.

2 Employing L2 as a condition to unveil the motoric nature of speech perception

Second language acquisition of speech is believed to be influenced by the native language of the learners. Especially, experienced learners who are considered near-native in proficiency will often establish stable intermediate categories in an audio-based learning model, the most widely renowned being the Speech Learning Model (SLM, Flege, 1987; Flege et al., 2003). In essence, L2 provides another dimension to testify language perception models by providing an intermediate, if not impoverished, level between L1 and L2 in the speaker's ontogeny (Major, 2002), and thus may depict different perceptual accuracy in acoustical or phonological tasks.

The motor theory is not exactly what others (Massaro and Chen, 2008) has criticized that perception comes through multiple sources. According to Liberman and Mattingly (1985) "...the string of phonetic segments is overlapped in the sound ... [with] no acoustic boundaries. Until and unless the child (tacitly) appreciates the gestural source of the sounds, he can hardly be expected to perceive, or ever learn to perceive, a phonetic structure." Under an experiment design of L2 perception, it will be even more demanding for L2 speakers to tactically retrieve intended gestures which are different from that in L1.

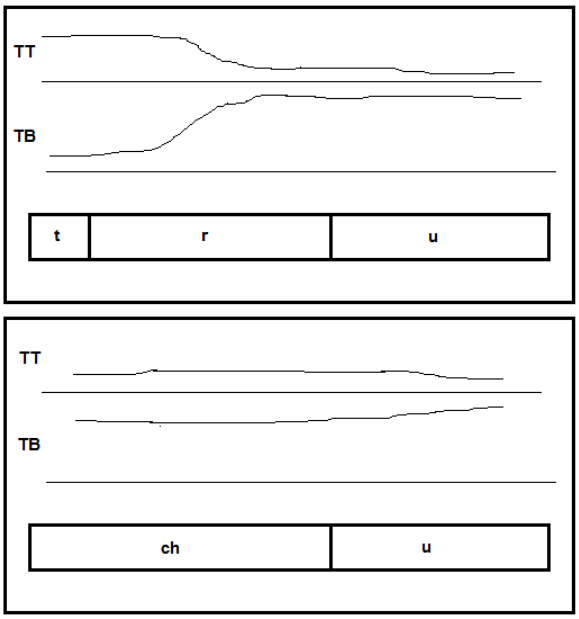
The basic rationale of motor theory is that gestures are invariant (and that acoustics are too variable), and thus more prone to be regarded as the percept under the ecological mechanism of human perception (Galantucci et al., 2006). This claim has been more amplifiably proven by this experiment because variations in gestures have caused serious perceptual problems, but not the 'crucial' acoustic cue of formant transition in L2 perceivers.

However, empirical studies seldom provided counter-evidence to the claims it has made. Nor did the auditory-motor debate ever been explicitly carried on in the scope of L2 acquisition. Actually, using L2 as an examining condition for the speech perception theories has its own inherent merits. Investigating this question through L2 has a very profound implication towards which of the two theories are more explanatory. In results in L1 that distinguishes accuracy in acoustic/speech sound perception, we can either say the salient different result of perceiving full CV words and CV transitions is because of the normalization of acoustic sound into speech sound category through extensive statistical learning; or, alternatively, we can also say that gestures are the distal objects that humans perceive directly as categories. However, in L2, it is easier to see whether pure acoustic sounds are perceived as linguistic sound, or if gestures play a part too. If the latter is true, the learnability of L2 speakers in one sound may be discovered to be different in different gestural environments. This is something L1 data cannot provide since L1 perceptions are almost always accurate in linguistic settings; even native listeners hear purely acoustic sounds. The current study examines the tongue tip and tongue body gestures of /r/ in CrV, which may vary in degrees of overlapped gestural constellations introduced by vowel contexts (/i/, where gestures are not heavily loaded and /u/, where gestures are more in conflict).

3 Gestural difference in Cantonese L2 speech of English *tr-* cluster

Cantonese speakers were reported by previous literature to have an inclination to mispronounce English C-r clusters. They either deleted the [r] or substituted it to [w] (Hung, 2002; Chan, 2006). However, for alveolar clusters (tr- and dr-), previous studies showed that considerable affrication was a feature of their production (Lan and Oh, 2012). According to SLM, Cantonese speakers should be able to perceive them in a *tr-*/*ch-* contrast in the initial position, given that they had ample experience in using English.

Even for native speakers, the acoustic signals of [r] in C-r production with the two vowel contexts are very similar. However, the *tr-* clusters in two vowel conditions, /i/ and /u/, were observed to have different gestures. The gestural difference can be shown in the following four schematic scores (following Browman and Goldstein, 1987) of gestures of CV syllables in *true*, *chew*, *tree* and *Chee*, respectively (See Figure 1). TT stands for Tongue Tip constriction degree. If the tongue tip moves forward or frontward, the magnitude would be high; TB stands for Tongue Body constriction degree. If the tongue body moves backward, the magnitude would be high.



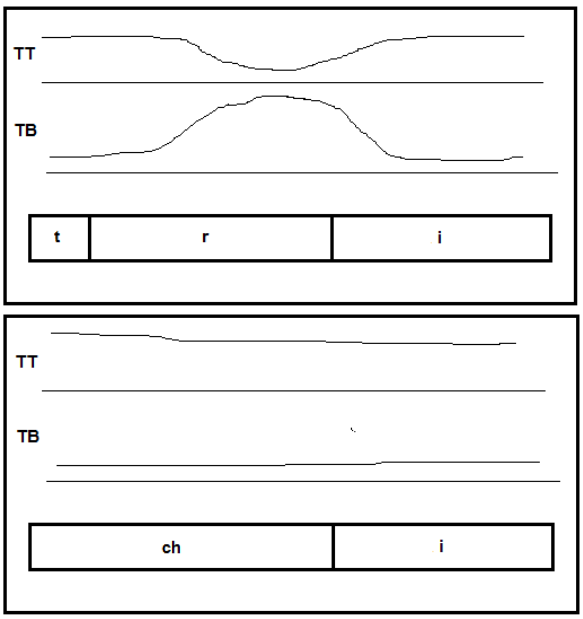


Figure 1: Schematic gestural scores for *true*, *chew*, *tree* and *Chee*, from top to bottom.

Note that the contrast of gestural scores for the [r] part in *tr-i* and *ch-i* is clear, because the [r] in /i/ environment shows both TT backward and TB retraction; whereas in *ch-i*, TT was always in forward position and TB always in rest position. However, the contrast of in *tr-u* and *ch-u* is more opaque because the TT and TB for both *tr-* and *ch-* words are eventually attaining the same position. Temporal overlap has made the sound contrast even more indiscernible to L2 learners.

One possible concern is, as has been pointed out earlier in this section, that although gesturally the [r] productions varied considerably for TT and TB constellations in /i/ and /u/ contexts, the acoustic properties of these two environments, nevertheless, were invariant in both conditions. Thus phonetically, the two conditions cannot constitute an allophonic variation. The two spectrograms in Figure 2 show that both sounds had considerable F3 rise, which is a signature characteristics for the presence of /r/.

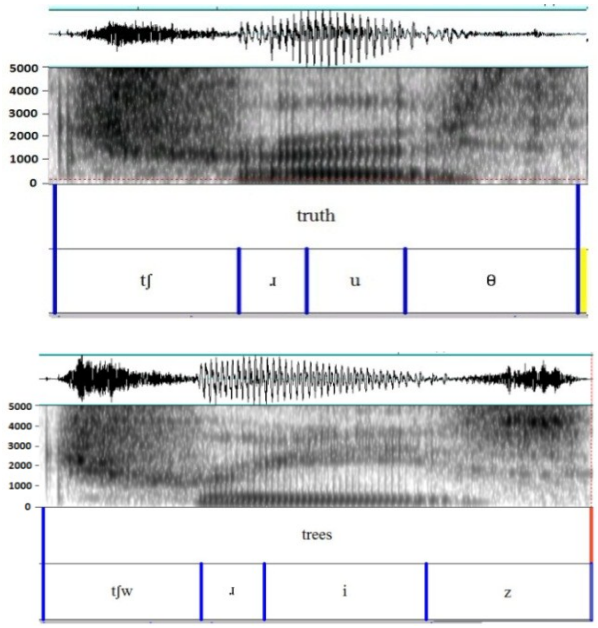


Figure 2: Spectrograms of truth and trees by native English speakers.

Apart from the impressionistic data, 92 of these tokens (46 *trees* and 46 *truth* productions) by native English speakers were analyzed for F3 in the [r] part and the results were sent to an independent variable t-test. Result showed that the difference of F3 in two vowel context was insignificant [t=-2.09, df=90, p=.305].

4 Experiment protocols for current study

The two experiments employed a contrast of word perception and non-speech acoustic detection respectively. In experiment 1, speakers were presented *tr-i* and *tr-u* sounds with *ch-i* and *ch-u* sounds as contrasts for discrimination perception. And in experiment 2, the CV formant transition parts are elicited from the speech and participants were asked to distinguish the acoustic segments from *tr-i* and *ch-i*, as well as *tr-u* and *ch-u*.

If the results are in support to auditory perception, as suggested by SLM, then the perceptual accuracy, no matter high or low, should be the same for L2 speakers because in both vowel contests, [r] sounds fully represents the acoustic data which is needed for L2 speakers to successfully/unable to distinguish the target sound contrast. The accuracy rate depends on the degree to which Cantonese speakers categorize the /r/ sound into phonemes correctly.

If the results support the classic motor theory, provided the difficulty in gestural contrast of *tr-* in the /u/ context by the learners, then the perceptual accuracy for full words should be better than the acoustic differences because of prior duplex experiments on native speakers has shown that acoustic perception of elicited "perceptual cues" should be poorer if not supported by the information of intended gestures by complete words. Also, the higher predicted accuracy may be attributed to the motor theorists' belief that human perception of speech sounds is modular and universal, which enables the universal grammar to help L2 learners perceive the intended gestures. A further prediction is that the accuracy for vowel contrasts of /i/ and /u/ should be different because of the different gestural difficulty demonstrated by the previous section.

4.1 Participants

Participants were three adults (2 females and 1 male, mean age=27.5) working as administrative staff at City University of Hong Kong. They all spoke English fluently as their working language. None of them had exposure to other foreign languages except English. All participants were right-handed with no reported hearing or motor-control defects. They did not have prior exposure to musical training. For controlling, three native monolingual English speakers (2 females and 1 male, mean age=26.5) from California, U.S. also participated in the study and went through the same procedure.

4.2 Stimuli

The perception tests were carried out in the Phonetics Lab, City University of Hong Kong. The listening perception materials used in two experiments are elicited from the same set of language productions by a native speaker. Stimuli words were produced by another Native American English speaker in a carrier sentence of "Now I say_____".

Words for both experiments were designed as minimal pairs of trVC and chVC (e.g. *trep-twep*). Stimuli differ in five vowel contexts, /i/, /ɛ,æ/, /u/, /ʌ/, and /ɔ/. Each word was repeated for three times by the native English speaker and then the most clearly pronounced utterance was selected as an experiment word. Stimuli for experiment 1 were the words themselves. However, in experiment 2, only the CV transition, or /r/ part, which was defined strictly as the start of voicing to the steady state of vowel, was used. In both experiments, test tokens were added with the equal numbers of fillers. In each experiment, stimuli were repeated for 10 times and randomized. In total, 600 tokens were tested (6

participants × 2 experiments × 5 vowels × 10 repetitions).

4.3 Procedure

Both experiments utilize the discrimination paradigm of the sounds in the minimal pairs. In this paradigm, three consecutive words (e.g., *treek/tweek/tweek*) were played, where the third word was identical to either the first or the second one. The participants were asked to circle the correct word on the answer sheet. The inter-stimulus intervals (ISI) were set at 250 milliseconds for both tasks.

To resolve a possible problem that might hinder reliability of stimuli induced by acoustic differences other than from the critical consonant part, the original vowel parts of the stimuli were replaced with the identical vowel which was sectioned from one token so that vowel quality remained consistent for the tasks. For instance, the [i] in one clear production of "treek" was used for all tokens with /i/ vowels in both experiments.

5 Results

5.1 Results by participant groups

For the sake of contrasting the two experime and highlighting the difference, the results w first presented with Cantonese and native Engl contrast and then by experiments.

Native English speakers showed an aver accuracy rate of 98.8% in discerning the *tr-/* contrast in words (N=300, std=.111). T difference between experiment 1 and 2 was 10 and 97.5%, which was statistically significant [2.259, df=298, p<.05]. The difference betw subjects was not significant [F(2, 297)=.3 p=.740]. The effect of vowel was not signific in experiment 2 [F(4, 145)=1.021, p=.398]. It v not significant in experiment 1 either.

For native Cantonese speakers, the overall accuracy rate was 81% (N=300, std=.397). The difference between experiment 1 and 2 was 66% and 95% [t=5.534, df=298, p<.0001]. The difference between subjects was insignificant [F(2, 297)=1.557, p=.214]. The effect of vowel was not significant in experiment 2 [F(4, 145)=.511, p=.728]. However, it was significant in experiment 1. [F(4, 145)=3.031, p<.05]. Among the vowel members, Tukey's post-hoc tests showed that the difference of vowel /i/ and /u/ were significant [/i/: md=.45, std.E=.145, p=.02; /u/: md=.45, std.E=.145, p=.02] (See Figure 3).

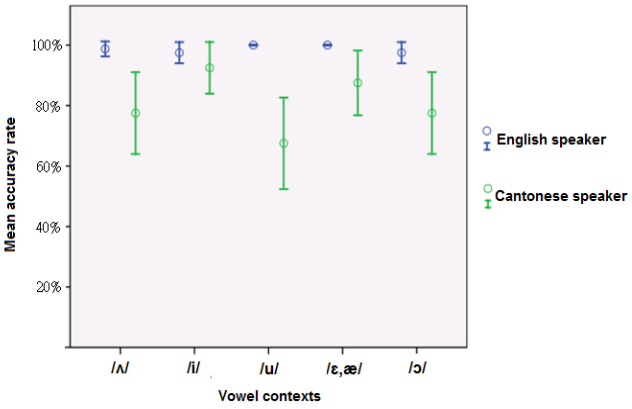


Figure 3: Accuracy rates of English and Cantonese speakers plotted by vowel types.

5.2 Results by experiments

The comparison of Cantonese and native English speakers' accuracy rate in each experiment was done, too. For experiment 1, the difference was significant [t=10.116, df=298, p<.0001]. However, for experiment 2, the difference was insignificant [t=1.136, df=258, p=.257] (See Figure 4).

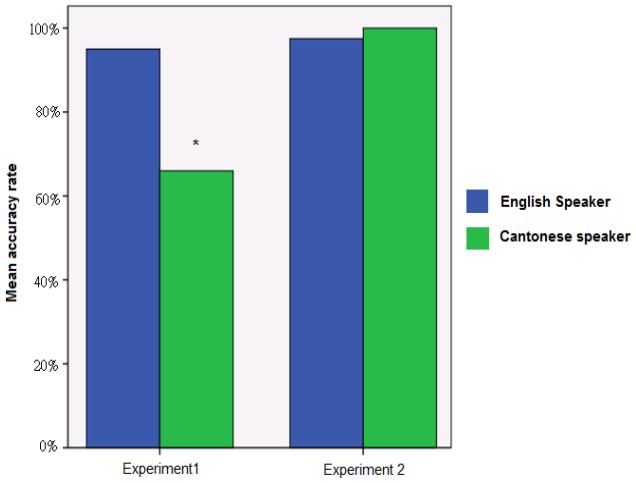


Figure 4: Accuracy rate of two experiments plotted by English/Cantonese speakers.

6 Discussions

The results of the two experiments may help giving some evidence to, if not settle, some of the theoretical disputes. For both experiments, native English speakers performed almost perfectly. The uniform high perception rate is not fruitful to support either of the competing theories. The analyze-worthy result lies in the comparison of English and Cantonese speakers as well as the

comparison between two experiments for Cantonese speakers, together with the effect of vowel contexts. It was shown that the first experiment witnessed a significantly different perceptual accuracy in two vowel contexts, with Cantonese speakers having a lower accuracy rate and a bigger discrepancy between vowel contexts; whereas the accuracy in second experiment was equally high in two groups and the high accuracy rates were not affected by vowel contexts. This showed that articulatory gestures in context might help establish categories and influence the acquisition of speech sounds, rather than acoustic information only. Therefore, the acoustic model cannot explain all of the L2 phonological acquisition patterns.

However, the results were also not in line with a purely motor theory either, because the traditional motor theory will predict that word perception should be better than acoustic perception because linguistic aids are provided. Instead, the results showed that word perception rate is poor for experienced Cantonese speakers.

A new "gestural-learning model" for L2 perception, based on Best's direct-realist theory, is hereby brought about. It has three major hypotheses. 1) perception of speech sounds is neither purely acoustical nor linguistically innate; 2) the process of learning of speech sounds is in fact the learning gestures through a distributive manner, which is influenced by the sensitiveness to gestural categories, and specifically, number and density of the categories being intervened with each other; 3) The learning process of an L2 ontogeny is gradual and gradient.

The model offers a way to explain for the results of this study. It may explain (1) why the accuracy rate in experiment 2 is better than experiment 1. In experiment 2, no gestural information is used so it's natural to perceive acoustic, non-linguistic sounds correctly because the focus is on acoustic detail; (2) why /i/ showed a higher accuracy rate than /u/. Since L2 learners are hard or insensitive to internalize much tokens of the gestural information in /u/ because of the complexity of the gesture. /i/ tokens are more salient to be perceived and are thus more prone to have gone through distributive learning. However, /u/ tokens are often neglected by its gestural complexity and thus be equivalently categorized with the *ch-* category, resulting in less distributive learning.

The major difference between the two classic theories and the current model is that language is neither purely linguistic nor acoustic. It involves a gradual learning process of intended gestures and gesture constellations. The direct- realist theory (Best 1995, Best et al, 2001) has already mentioned that the gestures in speech perception could also be learned through experience and not inherently acquired by the linguistic module. More than that, the current model combines the distributive learning model with the scope of second language speech learning, and adopts a gradual perspective into the learning process.

The possible drawback for the motor theory to reconcile to a distributive acquisition model is because of the idea that linguistic perception is modular and different from acoustic perception. This is partly real as confirmed by the results of this study. However, in this way, phones and phonemes are so apart that L2 speakers cannot learn phonemes through phones because they lack the certain intended gestures in development. Nevertheless, the results, as has discussed earlier, suggest that L2 speakers can still perceive more than 80% of the tokens correctly in some vowel contexts. This proves the ability for L2 learners to extract gestural information from L2 linguistic experience, hence the new model of speech perception. The table below is a sketch of the three models being compared (SFee Table 1).

Acoustic-based	Motor theory	Gestural learning
Frequency-based statistical learning	Purely innate as a single modular/device	Frequency-based statistical learning
Normalized prototype-another type of invariant	Direct perception of distal gestures	Direct perception of distal gestures

Table1: Comparison of three theoretic models.

One limitation of the study is that it failed to provide longitudinal data as direct evidence to support the third hypothesis of the model. However, from the experiment we see that for different vowel contexts, the accuracy rate was different, and the overall accuracy rate for the *tr-* category is 66%, which is in between perfect (100%) and chance (50%), representing an intermediate and gradual level of learning. Limitation also lies in the small number of participants and languages.

7 Conclusions

The study summarizes the different predictions the traditional acoustic approach and motor theory would give to Cantonese L2 speakers' perception

of *tr*- cluster in two vowel contexts. The result shows that Cantonese speakers perform poorly in real-word perception tests but near-ceiling in acoustic sound perception. This shows that acoustic sound is not a basis for L2 speech perception and the results supports the motor theory that speech is not perceived through sounds exclusively. However, the result that L2 speakers having an intermediate rate of successfully perceiving the L2 sounds raises questions towards motor theory's claim that the language modular is innate and cannot be shaped by experience.

Through these results, a new model of gestural learning was proposed through the discussions above. This model would bring fine-grained gestural percepts and frequency-based normalizing process of category formation together. Further investigations, such as more sound contrasts from more L1 and L2 linguistic backgrounds, as well as real-time EMA or fMRI imaging of L2 speakers' articulations may be done to testify it in detail.

Acknowledgments

The author is grateful to Dr. Dong Yanping for her advice, which gave birth to an initial idea of this study.

References

A. M. Liberman and I. G. Mattingly. 1985. The motor theory of speech perception revised. Cognition 21 (1), 1–36.

B. Galantucci, C. A. Fowler, and M. T. Turvey. 2006. The motor theory of speech perception reviewed. Psychonomic bulletin & review, 13(3), 361-377.

C. P. Browman and L. Goldstein. 1987. Tiers in articulatory phonology, with some implications for casual speech. Haskins Laboratories Status report on speech research, 1-30.

C. P. Browman and L. Goldstein. 1992. Articulatory phonology: an overview. Phonetica, 49, 155-180.

C. T. Best. 1995. A Direct Realist View of Cross-Language Speech Perception. In Strange, W. (ed.). Speech perception and linguistic experience: Issues in cross-language research, 171-204.

C. T. Best, G. W. McRoberts, and E. Goodall. 2001. Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. Journal of Acoustics Society of America, 109(2), 775–94.

D. B. Pisoni. 1985. Speech perception: Some new directions in research and theory. Journal of the Acoustical Society of America, 78, 381-388.

D. H. Whalen. 1981. Effects of vocalic formant transition and vowel quality on the English [s]-[S] boundary. Journal of the Acoustical Society of America, 69, 275-282.

D. W. Massaro and T. H. Chen. 2008. The motor theory of speech perception revisited. Psychonomic bulletin & review, 15 (2), 453-457.

J. E. Flege. 1987. The production of "new" and "similar" phones in a foreign language: evidence for the effect of equivalence classification, Journal of Phonetics, 15, 47-65.

J. E. Flege, C., Schirru., and I. R. A. MacKay. 2003. Interaction between the native and second language phonetic subsystems, Speech Communication, 40, 467-491.

P. K. Kuhl. 2000. A new view of language acquisition. Proceedings of the National Academy of Sciences, 97(22), 11850-11857.

P. Ladefoged. 1993. A course in Phonetics. Harcourt Brace Jovanovich College Publishers.

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

T. C. Rand. 1974. "Letter: Dichotic release from masking for speech". The Journal of the Acoustical Society of America, 55 (3), 678–680.

T. N. Hung. 2002. Language in contact: Hong Kong English phonology and the influence of Cantonese. In Kirkpatrick, A. (ed.). Englishes in Asia: Communication, Identity, Power and Education. Melbourne: Language Australia,191-200.

V. A. Mann and A. M. Liberman. 1983. Some differences between phonetic and auditory modes of perception. Cognition, 14, 211-235.

Y. Chan. 2006. Strategies used by Cantonese speakers in pronouncing English initial consonant clusters: Insights into the interlanguage phonology of Cantonese ESL learners in Hong Kong, IRAL proceedings, 44, 331-355.