Effect of Tonal Neutralization Rules on Native Speech Perception*

TSAN HUANG
University at Buffalo (SUNY) & The Ohio State University

0. Introduction

Many researchers discussed the interplay of phonology and speech perception (e.g. Hume and Johnson 2001), especially how perception helps to shape synchronic phonology and conditions historical sound change. In the other direction, despite claims for a universal map of inherent or positional perceptual salience in speech sounds (e.g. Steriade 2001), speech perception also appears to depend on listeners' linguistic experience. First of all, the native inventory of contrastive sounds may have an impact on speech perception. For instance, Japanese listeners, whose language has only one liquid sound, perceive the /r-l/ distinction differently from American English speakers (Miyawaki et al. 1975). Hume et al. (1999) found that while consonant-vowel transition seems to provide more place information for consonant place identification than stop burst for both American English and Korean listeners, the difference between the two kinds of stimuli is greater for Korean listeners. Hume *et al.* suggest that this is because the Korean listeners with a three-way stop consonant contrast, which is cued in part by the duration of aspiration, may be paying more attention to the CV transition between the burst and the vowel onset than do the English listeners, who have a two-way stop contrast. Second, the phonotactics of a language may have an effect on speech perception. Pitt (1998) found that phonotactic constraints biased native listeners' identification toward permissible sound sequences in English when perceiving continua whose two ends consist of a voicing or place contrast. Third, phonological rules operating in the listener's native language influence his/her perception as well. Fox (1992) found that English listeners fared poorly in identifying or discriminating vowels in the neutralizing context of /hVr(d)/. Fox suggests that knowledge of the phonological rule that neutralizes vowel contrast in this context may have affected listeners' ability to make perceptual decisions about vowel quality.

* This work was supported by Grant No. 5 R01 DC04421 to Keith Johnson from the National Institute on Deafness and Other Communication Disorders, and an Alumni Grant for Graduate Research to Tsan Huang from the Ohio State University in Spring 2003.

Past studies have shown that tonology may influence tone perception in a similar way. Gandour (1983, 1984) and Lee, Vakoch and Wurm (1996) showed that differences in lexical tone inventories may play a role in tone perception. In Gandour's (1983) study using synthesized f0 stimuli, speakers of Mandarin, Cantonese, Taiwanese, Thai, and English made dissimilarity judgments on tonal pairs. Results show that the tones were rated significantly differently by tone versus nontone language speakers, by Thai versus Chinese (Mandarin and Taiwanese) speakers, and by Cantonese versus Mandarin and Taiwanese speakers. Lee et al. (1996) used naturally recorded stimuli of Cantonese and Mandarin tones on word and nonword syllables. It was found that Cantonese and Mandarin listeners were better at discriminating tones in their own dialect and that the tone language speaking listeners did better than the English group.

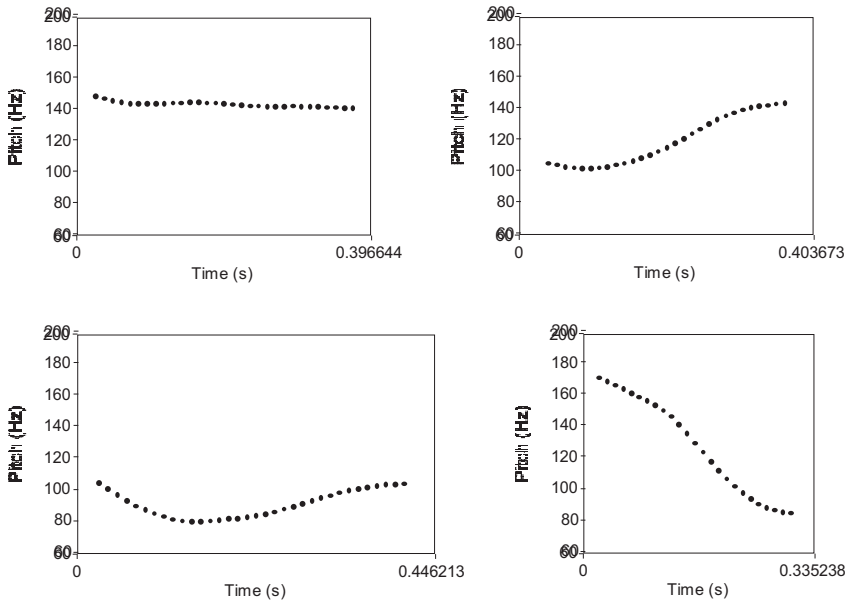
Gandour (1981, 1983) suggests that tone sandhi rules may also influence tonal perception. Using INDSCAL (Carroll and Chang 1970), Gandour (1981) analyzed confusion data from native listener identification of naturally produced Cantonese tones. He found that the high falling tone was placed midway between the level and the contour tones in the perceptual tone space. He argues that this is due to the fact that this tone has a high level allotone in Cantonese. Although the allotone was not present in the stimuli, allophony still interfered with listeners' perception. The effect of the same allophonic alternation showed up in Gandour's (1983) study, where Cantonese listeners perceived a /44/ (high level) contour to be similar to a /53/ (high falling). In the same experiment, Mandarin listeners perceived the /44/ contour to be similar to /35/ (rising), which, as Gandour points out, is due to the existence of the allophonic rule that turns a rising tone to a high level in Mandarin (Chao 1965; see also §1 below). In Huang (2001; see also Huang 2004), I have argued that the Mandarin T214 sandhi rule increases the confusability between T214 and T35 in native tone perception. The main findings of that study will be recapitulated in §2.

In the present study, we re-tested the phonology and perception interplay in the domain of lexical tone perception in Standard Beijing Mandarin and further investigated at what level(s) such influences were present. The data were compared with those of Huang (2001). The theoretical implications of these empirical data for speech perception models will also be discussed.

1. Background: Tones and Tone Sandhis in Standard Mandarin

Standard Mandarin has four lexical tones. Chao (1965) describes them as high level [55], mid-rising [35], low falling-rising [214], and high falling [51]. The numbers in the square brackets indicate the idealized pitch values of these tones on a five-level scale. I shall refer to them as T55, T35, T214 and T51, respectively. (Figure 1 shows the f0 traces of these tones.) There are also the so-called neutral-toned syllables in Mandarin, which are not specified for tone underlyingly.

Figure 1 F0 traces of T55 (upper left panel), T35 (upper right), T214 (lower left), and T51 (lower right), produced in monosyllables by a male Beijing speaker. Lengths of the X-axes reflect the relative durations.



Underlying full tones may be modified under the influence of their tonal environment. In the third tone sandhi, T214 becomes T35 when immediately followed by another T214 (Chao 1965, Duanmu 2000):

(1) The T214 Sandhi Rule
/T214.T214/ → [T35.T214][1]

Since an underlying /T35.T214/ sequence is also realized as [T35.T214], the paradigmatic contrast between T35 and T214 is lost before a following T214, creating many homophonous surface pairs. For example, /*hao*^{214}.*mi*^{214}/ 'good rice' is not distinguishable from /*hao*^{35}.*mi*^{214}/ 'millimeter', since both surface as [*hao*^{35}.*mi*^{214}]. The neutralization of T214 and T35 is complete perceptually (Wang and Li 1967, Peng 1996).

In the second tone sandhi rule, T35 becomes T55 when following a T55 or T35 and preceding a full-toned syllable (Chao 1965). This rule is optional and is not taught to second language learners.

(2) The T35 Sandhi Rule
/T55.T35.T*x*/ → [T55.T55.T*x*], or
/T35.T35.T*x*/ → [T35.T55.T*x*],

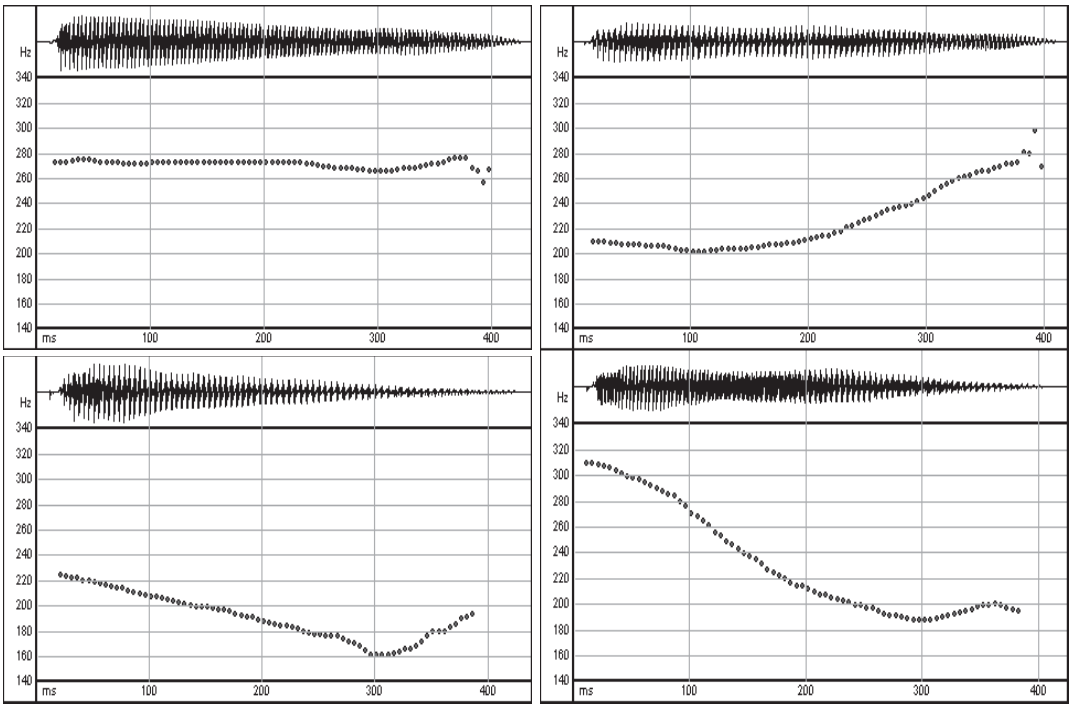
[1] I shall use a period [.] between two syllables produced as a sequence/word, and a hyphen [-] between an ordered pair of monosyllables. A notation with a slash [/] between two monosyllables, as in T55/T35, covers both T55-T35 and T35-T55. Tones may appear as raised diacritics in words, as in /*hao*^{214}.*mi*^{214}/ 'good rice'.

where T*x* is any non-neutral tone. /*cong*^{55}.*you*^{35}.*bing*^{214}/ → [*cong*^{55}.*you*^{55}.*bing*^{214}] '(Chinese) onion pancakes' is a familiar example (Chao 1965:36). This rule also leads to paradigmatic neutralization: the contrast between T55 and T35 is lost.

2. Summary of Huang (2001)

Huang's (2001) study tested the hypothesis that native phonology may influence speech perception, using natural speech tokens of Mandarin tones and Chinese- and English-speaking listeners. The stimuli used in that experiment were the first syllables cut from recorded disyllabic nonsense sequences and had the same segmental shape /ba/ with varying tones (Figure 2). An AX discrimination task was used. While each stimulus pair was played (at a 300ms inter-stimulus interval, or ISI), listeners made simple "same"/"different" judgments.

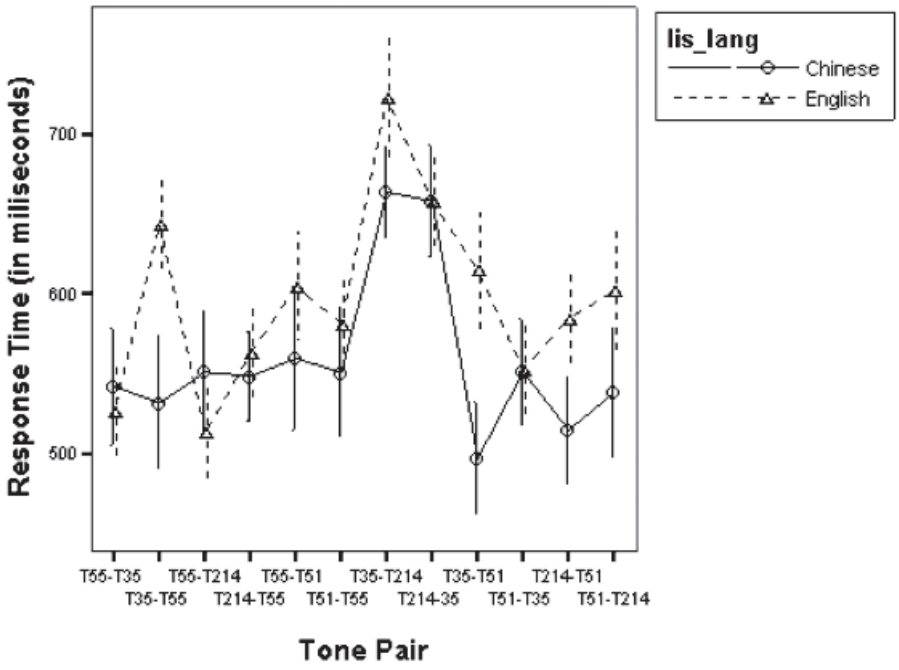
Figure 2 F0 traces of the stimulus tones used in Huang (2001). Upper left: T55; Upper right: T35; Lower left: T21(4) – final rise cut off in non-final position; Lower right: T51.



Both the judgment accuracy and reaction time (RT) were recorded. The results showed that T35 and T214 – i.e., the two tones involved in the T214 sandhi – were perceptually more confusable (attracting more mistakes and inducing longer RTs) than any other tone pairs for both listener groups. As error rates were very low for both listener groups (overall 4.5% for the Chinese listeners and 5.25% for the AE listeners), there was no statistically significant difference among the tone pairs.

A repeated measures analysis of variance (ANOVA) was performed on the RT data for the correct "different" responses, with tone pair (i.e., T55/T35, T55/T214, T55/T51, T35/T214, T35/T51, and T214/T51) as the within-subject variable (12 levels), and listener language (i.e., Chinese and English) as the between-subject variable (2 levels). No significant difference was found between language groups, [$F(1, 21) = .76$, $p = .393$]. But there was a significant effect with tone pair types, sig.[$F(7.487, 157.221) = 13.382$, $p < .001$, partial $\eta^2 = .389$]. The interaction of language and tone pair was also significant, sig.[$F(7.487, 157.221) = 3.295$, $p = .002$, partial $\eta^2 = .136$].

Figure 3 RT plot (in milliseconds) for the correct "different" tone pair responses. Error bars show one standard error.



Pairwise comparison for each language group showed that T35/T214 were the most confusable for the Chinese listeners and were significantly different from all other pairs ($p < .05$), while T35-T51 was the least confusable and significantly different from all other pairs except for T35-T55 and T214-T55. While T35/T214 were also the most confusable for the AE group, T214-T35 was not significantly different from T35-55, T35-T51, or T51-T214. They also found three tone pairs to be the least confusable, namely T55-T35, T55-T214, and T51-T35, which did not stand out in the Chinese listeners' data at all (see Figure 3).

Pairs T35-T55 and T35-T51 are quite different for the two listener groups. As seen in the ANOVA report above, these were among the least confusable for the Chinese listeners but the more confusable for the AE listeners. A T-test on the RT data shows that these between-group differences are significant: $t = -2.136$, $p = .045$, and $\eta^2 = 0.178$ for T35-T55, and $t = -2.254$, $p = .035$, and $\eta^2 = 0.195$ for

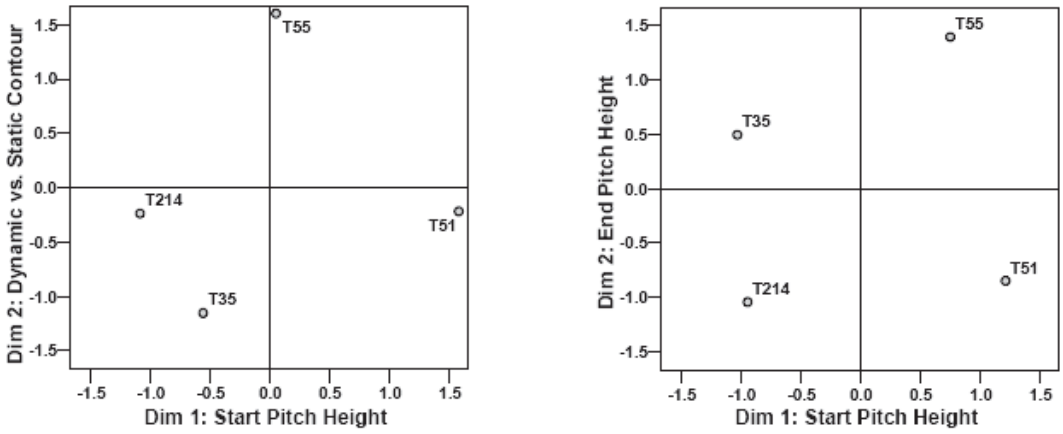
T35-T51. What is special about these pairs is that the pitch offset of the first tone (T35 in both cases) and the pitch onset of the second tone (T55 or T51) are very similar in height. This seems to affect the AE listeners' perception, but not the Chinese listeners', whose RT curve is fairly flat, except for T35-T214 and T214-T35, while that for the AE listeners has more obvious maxima and minima, some of which are attributable to this factor (e.g. T35-T55, T55-T51 and T51-T214). It is likely that the AE listeners, with no lexical tone categories in their lexicon, were more sensitive to the pitch onsets and offsets and used them as phonetic cues to discriminate the tones (Wang 1976, Stagray and Downs 1993). The more similar these points are, the more confusable the tones are for the AE listeners, as in the case of T35-T55 and T35-T51. On the other hand, the Chinese listeners may have perceived the f0 contour on a monosyllable as an indivisible unit and thus ignored such phonetic details of the contour to a certain extent.

Obviously, these different processing strategies were not always to the advantage of either group of listeners: for T55-T35 and T55-T214, the AE listeners used the phonetic cues more efficiently and scored shorter RTs. But the Chinese listeners made good use of contour information in pairs T35-T55 and T35-T51. This difference in strategies is actually a very telling one, because it suggests that the long RTs for the T35/T214 pairs may have resulted from different factors for the two groups. That is, T35 and T214 were confusable for the Chinese listeners not because of the phonetic similarity between the tones that affected the AE listeners, but because of the tone sandhi in their native phonology which neutralizes the contrast between these two tones in one environment.

The RTs for T35/T214 are much longer relative to the other tone pairs in the Chinese listeners' data, while the inter-pair RT differences for the AE listeners are less pronounced. This is best visualized using INSCAL (Carroll and Chang 1970). The analysis also brought out the two most important tonal characteristics (labeled along the two dimensions in Figure 4) in each listener group's perception. Notice that the AE listeners paid attention to both f0 onsets and offsets, while the Chinese listeners seemed to rely on f0 onsets only.

As the perceptual distance *d* was computed with the reciprocal function $d = 1/RT$ (Shepard 1978), the distance between T35 and T214 is noticeably much shorter in the Chinese space. Recall that the within-group pairwise comparisons also showed that for the Chinese listeners, T35/T214 were significantly different from all other tone pairs. This seemingly surprising pattern can be explained if, as Peng (1996) found, some surface [T35] syllables may be linked to both /T35/ and /T214/ morphemes (perhaps as part of a compound) in the Chinese listeners' lexicon. It is worth noting that with such a complex mental representation of the tonal category (or rather categories) of certain morphemes and a one-to-many mapping of surface tone to underlying tone categories, the boundary between the T35 and T214 categories may be blurred and the confusion between these tones may exist beyond just the sandhi environment, which is why there is not much difference between the RTs for T35-T214 and T214-T35 for the Chinese listeners.

Figure 4 The Chinese (left panel; stress = 0.189, RSQ = 0.89) and English (right panel; stress = 0.169, RSQ = 0.91) listeners' perceptual spaces of the four tones as revealed by the INSCAL analysis.



3. The Present Study: A Speeded AX Task

Fox (1984) found that faster response led to decreased language effects. In particular, he showed that a response latency shorter than 500 ms blocks the lexical effect on perception. We decided to investigate whether a speeded task would reduce the sandhi effect on tone perception as shown in Huang (2001).

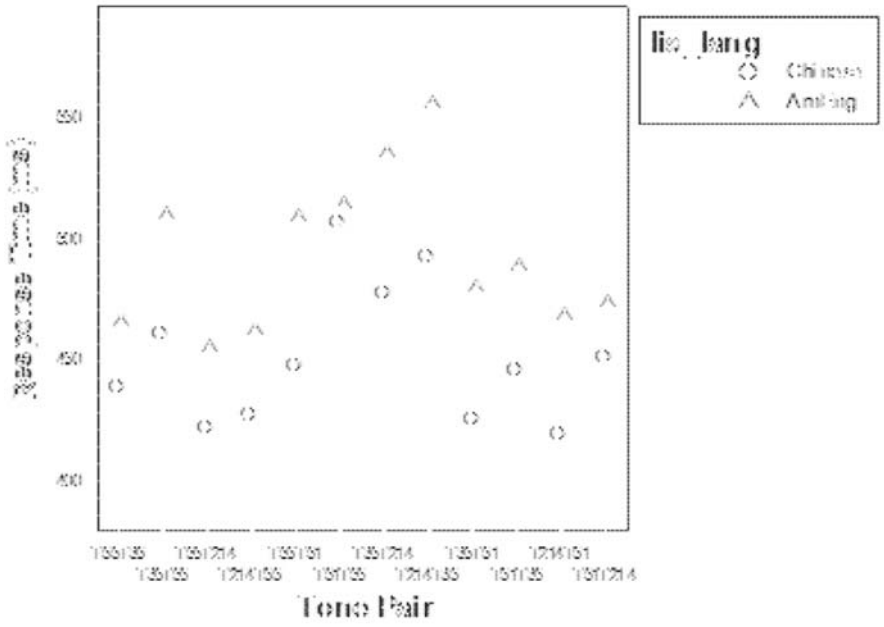
3.1. Procedures

A group of 24 high school students from Beijing and a control group of 20 native AE-speaking undergraduate college students from Columbus, Ohio participated in the present study. The task was the same as that used in Huang (2001), except that they were asked to respond within 500ms. (They could check their performance on the computer screen in front of them. As can be seen in Figure 5 below, they did very well, having most RT values below or around 500ms.) The inter-stimulus interval was also shortened to 100ms. The stimuli were recorded as monosyllables and had the segmental makeup of /ba/. (The f0 traces shown in Figure 1 were from this recording.) As in the previous study, error rates and RT were recorded.

3.2. Results and Analyses

Since the stimuli were played at a comfortable volume with no background noise, the error rates were very low, 5.13% and 5.88% overall for the Chinese and the AE listeners, respectively. The pairs that attracted the most errors were T35/T214 (9.35%) and T55/T51 (7.61%) for the Chinese listeners, and T55/T51 (8.75%) and T35/T214 (7.5%) for the AE listeners. Figure 5 shows the group RT plots.

Figure 5 RT (in milliseconds) for the correct "different" tone pair responses. Error bars show one standard error.



In comparison with the RT plots in Figure 3, we may notice that the AE listeners had very similar curves for both datasets, except that here T55/T51 have longer RTs and that T51-T35 is at a maximum. These seem explainable in terms of the differences in the stimuli used in the two studies: the differences in the f0 offsets of the first stimulus (T55 or T51) and the onset of the second (T55, T51 or T35) in these pairs are smaller in the present study. The curves for the Chinese listeners' datasets are also very similar, except that the points for T35-T55 and T51-T55 are now at maxima. The explanation of matching f0 offset and onset cannot be invoked for the Chinese listeners, because the f0 offset of T35 was also similar to the onset of T55 in the 2001 stimuli. In addition, unlike the AE listeners, the Chinese listeners have a shorter RT for T55-T51 than T51-T55 here. One possible explanation for the RT disparity for T35-T55 in the two studies is that: We have a group of Beijing listeners in the present study, who may have the T35 sandhi in their speech. As for T51-T55, recall that the stimuli used in this study were recorded as monosyllables. When played at a 100ms ISI, they might have been heard as two consecutive syllables in normal speech. As a result, the Beijing listeners might have undone the downstep effect of T51 (Xu 1997) and effectively "raised" the onset of the following T55. As the Chinese listeners tend to rely on the onset pitch to predict the whole tonal contour, T55 might have been mistaken as T51 when it was played after a T51.

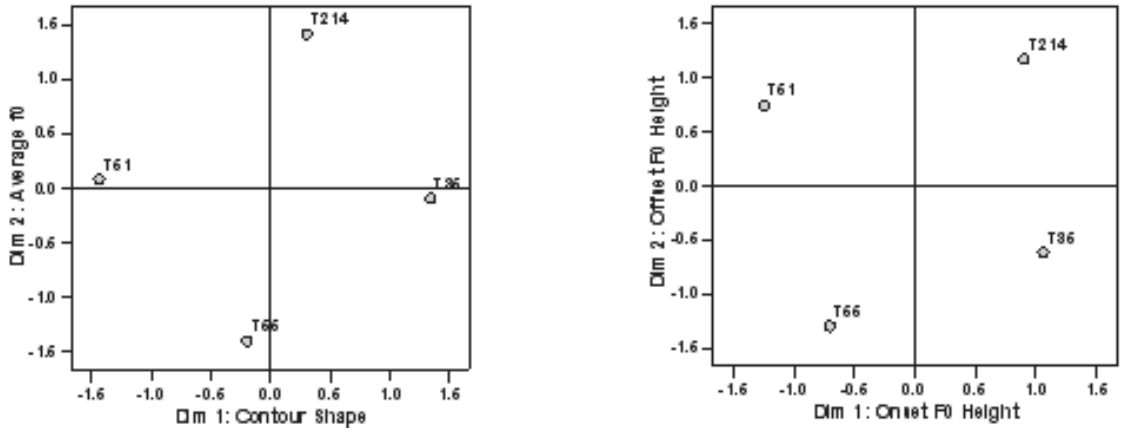
A repeated measures ANOVA on the RT data of correct "different" responses found no significant between-subject effect, $F(1, 41) = 2.48$, $p = .123$. The within-subject factor of tone pair had a significant effect, sig.[$F(9.396, 385.225) = 21.455$, $p < .001$, partial $\eta^2 = .344$]. There was also a significant effect with the interaction

of language and tone pair, sig.[F(9.396, 385.225) = 2.136, p = .024, partial η^2 = .05]. Results from paired comparisons using independent samples T test were rather unremarkable, with only T35-T55 and T35-T214 showing some marginal between-group differences.

Separate ANOVA analyses on the RT data for each group revealed that tone pair types had a significant effect for both groups (p < .001). The AE listeners found T214-T35, T35-T214 and T35-T55 the most confusable, while T55/T51 fell in the middle of the confusability rank. For the Chinese listeners, T51-T55 was the most confusable, followed by T214-T35, T35-T55 and T35-T214.

Group INSCAL spaces are shown in Figure 6. As in the earlier study, the AE listeners' tone space has very clearly defined dimensions. The effects of the T214 and T35 sandhi rule and the possible downstepping effect of T51 on T55 are not visible in the Chinese listeners' space, precisely because all of them might be at work. If the sandhi effects did not go away completely even in this simple AX task with a short 100ms ISI and a response time constraint, it is evidence for the contention that language-specificity exists in early levels of speech processing.

Figure 6 INSCAL spaces for the Chinese (left panel; RSQ = .918, stress = .157) and the AE listeners (right panel; RSQ = .918, stress = .164).



4. General Discussion

As is evident from the experimental data reported above, linguistic experience can lead to language-specific patterns in speech perception, which should be accounted for in any model of speech perception. For Steriade (2001), the universal perceptual salience map of speech sounds does not change for speakers of a particular language. Rather, language-specific patterns arise from different constraint rankings. However, leaving language-specificity to different constraint rankings does not offer an adequate theoretical explanation for the phenomenon, as rankings derived from empirical data only describe the patterns but do not

reveal the driving mechanisms. Some other force(s) must be working along with the perceptual map in determining language-specific patterns.

Guenther and colleagues (Guenther and Gjaja 1996, Guenther et al. 1999, Guenther and Bohland 2002) propose that the language-specificity in speech perception has a neurophysiological basis. In their neural model of an auditory cortical map, the formation of the map is determined by stimulus input and type of training. In particular, Guenther et al. (1999) found that categorical training in psychophysical experiments using nonspeech-like bandpass-filtered acoustic noise in different frequency ranges led to smaller cortical representation of – hence, decreased sensitivity to – stimuli in the training range, while discrimination training led to larger cortical representation – hence, increased sensitivity – in the training range. Functional magnetic resonance imaging (fMRI) studies by Guenther and Bohland (2002) provided further supporting evidence for this assertion. If an auditory warping similar to what Guenther et al.'s model describes existed, it would certainly serve the linguistic purpose well, as the warping directs neural activities to distinguishing between-category differences while ignoring irrelevant within-category differences. But the model as it stands now cannot account for the different degrees of language-specific effects as seen in the two studies discussed above.

In Johnson's (2004) lexical distance model, the universal perceptual distances assumed for speech sounds need not be altered by linguistic experience to account for language-specific effects, which simply emerge from the lexicon, as incoming signals are compared directly against phonetically detailed forms stored there. The model computes overall perceptual distance (d) from two sources, namely inherent auditory similarities between two stimuli (d_a), and aggregated average difference in lexical activations by the two stimuli (d_l, computed as the difference in the amounts of activation of the lexicon caused by these stimuli, with a constant k gating the influence of this lexical distance on perception under different experimental conditions); or $d = d_a + k \times d_l$. It is claimed that the model has the ability to distinguish discrimination performance from categorization performance, the former of which can be found in a minimal uncertainty task such as the speeded AX task reported here (no lexical access, perceptual distance computed almost exclusively from auditory distance) and the latter of which in tasks involving higher memory load such as AXB identification (lexical forms consulted). Johnson's (2004) fricative perception data from a rating task and a speeded AX discrimination task by Dutch and AE listeners support this claim.

Neither the neural model nor the lexical distance model explicitly discusses the issue of how neutralization rules may affect discrimination of two contrastive sounds (or tones) that are neutralized in a certain environment. Within Guenther et al.'s model, we may imagine a "noisy" training condition under which stimuli categorized into an abstract representation of A may sometimes have to be also categorized as B. This double-identity status may weaken the contrast between the relevant categories. Within Johnson's lexical distance model, because of the cross-representation of two sounds (e.g. T35 and T214 in Mandarin), a T35 or

T214 input may activate lexical items containing either a /T35/ or /T214/ form in the lexicon. Consequently, the lexical distance between /T35/ and /T214/ is predicted to be smaller than if there is no such neutralization rule. Both models need further refining to account for the data reported here.

References

Carroll, J. D. and J.-J. Chang. 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika* 35(3): 283-319.

Chao, Yuen Ren. 1965. *A Grammar of Spoken Chinese*. Berkeley and Los Angeles: University of California Press.

Duanmu, S. 2000. *The Phonology of Standard Chinese*. Oxford: Oxford University Press.

Fox, Robert Allen. 1984. Effect of lexical status on phonetic categorization. *Journal of Experimental Psychology: Human Perception and Performance* 10(4): 526-540.

Fox, Robert Allen. 1992. Perception of vowel quality in a phonologically neutralized context. In Y. Tokura, E. Vatikiotis-Bateson, and Y. Sagisaka (eds.) *Speech Perception, Production and Linguistic Structure*, 21-42. Tokyo: Ohmsha and IOS Press.

Gandour, Jack. 1981. Perceptual dimensions of tone: Evidence from Cantonese. *Journal of Chinese Linguistics* 9: 20-36.

Gandour, Jack. 1983. Tone perception in Far Eastern languages. *Journal of Phonetics* 11: 149-176.

Guenther, F. H. and J. W. Bohland. 2002. Learning sound categories: A neural model and supporting experiments. *Acoustical Science and Technology* 23(4): 213-220.

Guenther, Frank H. and Marin Gjaja. 1996. The perceptual magnet effect as an emergent property of neural map formation. *Journal of the Acoustical Society of America* 100: 1111-1121.

Guenther, F.H., F.T. Husain, M.A. Cohen, and B.G. Shinn-Cunningham. 1999. Effects of categorization and discrimination training on auditory perceptual space. *Journal of the Acoustical Society of America.* 106: 2900-2912.

Huang, Tsan. 2001. The interplay of perception and phonology in Tone 3 sandhi in Chinese Putonghua. In E. Hume and K. Johnson (eds.) *Studies on the Interplay of Speech Perception and Phonology* (OSU working papers in linguistics No. 55), 23-42.

Huang, Tsan. 2004. Language-specificity in auditory perception of Chinese tones. Ph.D. dissertation, The Ohio State University.

Hume, E. and K. Johnson. 2001. A model of the interplay of speech perception and phonology. In E. Hume and K. Johnson (eds.) *The Role of Perception in Phonology*. New York: Academic Press.

Hume, E., K. Johnson, M. Seo, G. Tserdanelis, and S. Winters. 1999. A Cross-linguistic study of stop place perception. *Proceedings of the XIVth International Congress of Phonetic Sciences*, 2069-2072.

Johnson, Keith. 2004. Cross-linguistic perceptual differences emerge from the lexicon. In A. Agwuele, W. Warren, and S.-H. Park (eds.) *Proceedings of the 2003 Texas Linguistics Society Conference: Coarticulation in Speech Production and Perception*, 26-41. Sommerville, MA: Cascadilla Press.

Lee, Y.-S., D.A. Vakoch, and L.H. Wurm. 1996. Tone perception in Cantonese and Mandarin: A cross-linguistic comparison. *Journal of Psycholinguistic Research* 25: 527-542.

Miyawaki, K., W. Strange, R. Verbrugge, A.M. Lieberman, J.J. Jenkens, and O. Fujimura. 1975. An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception and Psychophysics* 18: 331-340.

Peng, S.-H. 1996. Phonetic implementation and perception of place coarticulation and tone sandhi. Ph.D. dissertation, The Ohio State University.

Pitt, Mark A. 1998. Phonological process and the perception of phonotactically illegal consonant clusters. *Perception & Psychology* 60(6): 941-951.

Shepard, R. N. 1978. The circumplex and related topological manifolds in the study of perception. In S. Shye (ed.) *Theory Construction and Data Analysis in the Social Sciences*. San Francisco: Jossey-Bass.

Stagray, J. R. and D. Downs. 1993. Differential sensitivity for frequency among speakers of a tone and a nontone language. *Journal of Chinese Linguistics* 21(1): 143-163.

Steriade, D. 2001. A perceptual account of directional asymmetries in assimilation and cluster reduction. In E. Hume and K. Johnson (eds.) *The Role of Perception in Phonology*. New York: Academic Press.

Wang, W. S.-Y. 1976. Language change. *Annals of the New York Academy of Sciences*, 61-72.

Wang, W. S.-Y. and K.-P. Li. 1967. Tone 3 in Pekinese. *Journal of Speech and Hearing Research* 10: 629-636.

Xu, Yi. 1997. Contextual tonal variations in Mandarin. *Journal of Phonetics* 25(1): 61-83.

Tsan Huang
609 Baldy Hall
SUNY/Buffalo
Buffalo, NY 14260

thuang3@buffalo.edu