Effect of several acoustic cues on perceiving Mandarin retroflex affricates and fricatives in continuous speech

Jian Zhu and Yaping Chen

Citation: The Journal of the Acoustical Society of America **140**, 461 (2016); doi: 10.1121/1.4955311
View online: https://doi.org/10.1121/1.4955311
View Table of Contents: https://asa.scitation.org/toc/jas/140/1
Published by the Acoustical Society of America

Effect of several acoustic cues on perceiving Mandarin retroflex affricates and fricatives in continuous speech

Jian Zhu[a] and Yaping Chen
Department of English, Beijing Foreign Studies University, 2 North Xisanhuan Avenue, Haidian District, Beijing 100089, China

(Received 31 October 2015; revised 8 June 2016; accepted 16 June 2016; published online 20 July 2016)

Relatively little attention has been paid to the perception of the three-way contrast between unaspirated affricates, aspirated affricates and fricatives in Mandarin Chinese. This study reports two experiments that explore the acoustic cues relevant to the contrast between the Mandarin retroflex series /tʂ/, /tʂʰ/ and /ʂ/ in continuous speech. Twenty participants performed two three-alternative forced-choice tasks, in which acoustic cues including closure, frication duration (FD), aspiration, and vocalic contexts (VCs) were systematically manipulated and presented in a carrier phrase. A subsequent classification tree analysis shows that FD distinguishes /tʂ/ from /tʂʰ/ and /ʂ/, and that closure cues the affricate manner. Interactions between VC and individual cues are also found. The FD threshold for separating /ʂ/ and /tʂ/ is susceptible to the influence of the following vocalic segments, shifting to lower values if frication is followed by the low vowel /a/. On the other hand, while aspiration cues /tʂʰ/ before /a/ and /ɤ/, this acoustic cue is obscured by gesture continuation when /tʂʰ/ precedes its homorganic approximant /ɻ/ in natural speech, which might cause potential confusion between /tʂʰ/ and /ʂ/. © *2016 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4955311]

[CGC] Pages: 461–470

I. INTRODUCTION

A. Perceptual cues for affricates and fricatives

It has long been known that affricates and fricatives with the same place of articulation form a single perceptual continuum (Gerstman, 1956). The two types of sounds only differing in manner of articulation can be distinguished by an array of acoustic cues when presented in monosyllables, notably frication duration (FD) (Kluender and Walsh, 1992; Mitani *et al.*, 2006; Repp *et al.*, 1978), amplitude rise slope (Mitani *et al.*, 2006), relative amplitude (Hedrick, 1997), initial burst (Dorman *et al.*, 1980), and amplitude rise time (Howell and Rosen, 1983; Mitani *et al.*, 2006; see Kluender and Walsh, 1992, for contradictory arguments). In continuous speech, they can also be distinguished by closure duration (Repp *et al.*, 1978; Dorman *et al.*, 1980).

However, these early studies only focus on the two-way distinction between affricates and fricatives in a couple of languages, particularly English. In comparison, the three-way distinction of unaspirated affricates, aspirated affricates and fricatives, as is found in Mandarin, is less well understood. Therefore, a study on this three-way contrast is quite necessary to alleviate the underrepresentation of aspirated affricates in the speech perception literature.

B. Affricates and fricatives in Mandarin Chinese

Mandarin Chinese has a three-way contrast of voiceless affricates and fricatives: the alveolo-palatal series /tɕ, tɕʰ, ɕ/ (j, q, x in Pinyin Romanization); the dental series /ts, tsʰ, s/ (z, c, s in Pinyin Romanization); and the retroflex series /tʂ, tʂʰ, ʂ/ (zh, ch, sh in Pinyin Romanization). Specifically, Mandarin Chinese contrasts two types of affricates: aspirated and unaspirated. A number of acoustic cues may be relevant to the affricate-fricative contrast in Mandarin: initial burst, amplitude rise time, amplitude rise slope, FD, closure, and aspiration. For example, the initial burst may be capable of distinguishing affricates and fricatives in Mandarin, as Liu *et al.* (2000) found that there was initial burst 72% of the time in affricates but only 1% in fricatives. Also, amplitude rise time is reported to be used by Mandarin listeners to separate affricates from fricatives (Tsao *et al.*, 2006)

Previous production studies suggest that FD may be one of the perceptual correlates that serve to distinguish affricates and fricatives in Mandarin, as the duration of both aspirated affricates and fricatives is considerably longer than that of unaspirated affricates in production (e.g., Feng, 1985; Qi and Zhang, 1982). However, whether duration separates aspirated affricates from fricatives is less clear, as production data show that the duration of aspirated affricates is only slightly shorter than that of fricatives (Feng, 1985; Liu *et al.*, 2000), if not longer (Qi and Zhang, 1982). Tsao *et al.* (2006) specifically tested the role of FD in distinguishing Mandarin affricates and fricatives in an AX (same-different) discrimination experiment, in which a set of synthesized /tɕ/, /tɕʰ/ and /ɕ/ varying in both FD and amplitude rise time were administered to Chinese listeners. They found that FD was utilized to distinguish between /tɕ/ and /tɕʰ/, and also between /tɕ/ and /ɕ/, but not between /tɕʰ/ and /ɕ/.

[a]Electronic mail: zhujianbw@gmail.com

Closure intervals before frication have been found to cue the manner of affricate in continuous speech (Dorman *et al.*, 1980; Repp *et al.*, 1978). Repp *et al.* (1978) found that inserting a silent interval between the word "say" and "shop" would cause listeners to report hearing "say chop," and that the participants' report of hearing "chop" increased as the silent interval increased.

Another potential factor that may influence the percept of affricates and fricatives is the VC. Mandarin Chinese has six vowel phonemes /i, y, u, ə, ɤ, a/ (Lee and Zee, 2003). Additionally, there are two "apical vowels" that are treated as allophones of the high front vowel /i/ because they can only be preceded by homorganic consonants—consonants with the same place of articulation. Traditionally, "apical vowels" are transcribed as two vowels /˔ʅ/ and /˔ʮ/ by many Chinese linguists (Wu and Lin, 1989; Wu *et al.*, 2015; Lin and Wang, 1992, pp. 43–44), but they are also argued to be syllabic fricatives /z̩/ and /z/ (Duanmu, 2007, pp. 34–35). A recent ultrasound study reported that apical vowels in Mandarin resemble approximants in actual realization and thus should be transcribed as the approximant /ɻ/ and /ɻ̩/ (Lee-Kim, 2014). We agree with Lee-Kim (2014) and the symbol /ɻ/ for retroflex approximant is used throughout this article.

In Mandarin Chinese, when affricates and fricatives precede their homorganic approximants (i.e., /ʂɻ/ and /tʂʰɻ/), the tongue configuration remains almost unchanged from the consonant to the following vocalic segment (Lee-Kim, 2014; Wu and Lin, 1989, p. 140; Wu *et al.*, 2015, p.159). Because of gesture continuation of this kind, the aspiration in aspirated affricates is found to display two distinct features depending on the following vocalic segment. If the tongue constriction of the following vocalic segment is the same as that of its preceding aspirated affricate, the feature [+aspiration] is realized as lengthened frication, otherwise it is realized as aspiration (Wu and Lin, 1989, p. 140; Wu *et al.*, 2015, p. 159). Following this analysis, if /tʂʰ/ precedes its homorganic approximant /ɻ/, the abstract feature [+aspiration] is realized as frication in surface form, rendering it similar to a fricative. But when /tʂʰ/ precedes other vowels, aspiration in that aspirated affricate remains intact. Thus, while aspiration may be the acoustic cue of aspirated affricates, this cue is not available when an aspirated affricate precedes its homorganic approximant.

Different VCs may also cause listeners to recalibrate the categorical boundary on a perceptual continuum (Repp, 1982; Repp and Liberman, 1987). For instance, it has been found that, when perceiving stops, listeners shifted their perception toward longer voice onset time (VOT) for the high VC (Nearey and Rochet, 1994). Production data show that retroflex affricates and fricatives are considerably longer before /ɻ/ than before /a/ and /ɤ/ (Feng, 1985). This can be traced back to the aerodynamics of vowel production. Ohala (1983, p. 204) noted that close vowels (i.e., high vowels)—induce a higher velocity of the oral airflow, which, in turn, causes greater air turbulence and perhaps longer frication of the segment. Given the production data, we hypothesize that longer FD is necessary for perceiving fricatives before high and mid vowels than before low vowels, but it seems that the effect of VCs in perceiving frication has not been systematically explored.

C. The current study

Currently, few studies have probed into the acoustic correlates of the three-way contrasts of affricates and fricatives in Mandarin. Tsao *et al.* (2006) have examined the role of FD and amplitude rise time, but other acoustic cues and interactions between cues and VCs are in need of further study. Thus the present study examines the perceptual correlates of the retroflex affricate-fricative series by testing the following cues—FD, closure, aspiration and their interactions with VCs. Our experiment aims to demonstrate that perception of the retroflex affricate-fricative series in continuous speech can be the temporal integration of the following acoustic cues: (1) silent intervals that immediately precede the frication noise; (2) the frication noise itself; and (3) acoustic properties of the ensuing vocalic segment.

The goals of this study are threefold. First, it aims at bringing more information to bear on the affricate-fricative contrast through the investigation of the three-way contrast of affricates and fricatives in Mandarin. Second, we want to explore the role of VCs in perceiving frication noise. To our knowledge, few studies have tested the role of VC in perceiving the affricate-fricative contrast. Finally, we include the special "apical vowel" to investigate its perceptual implications within the phonological system of Mandarin Chinese.

Two identification experiments are designed to test the role of FD, closure, and aspiration in the perception of three retroflexes /tʂ, tʂʰ, ʂ/ in Mandarin Chinese, and the possible influences of VCs. Three vocalic segments—/ɻ/, /ɤ/ and /a/—are chosen as part of the phonetic context within which the target stimuli occur. They are all unrounded vowels that can occur after retroflexes: /ɻ/ has the same tongue configuration as the retroflexes while the other two vocalic segments—the mid vowel /ɤ/ and the low vowel /a/—are successively lower in vowel height. It is hypothesized that the categorical boundary may shift to higher values, or longer FD in mid vowel /ɤ/ than in low vowel /a/. The retroflex approximant /ɻ/ may even require an even higher threshold of frication because /ɻ/ as a continuation of the retroflex gesture tends to induce even longer frication than /ɤ/ and /a/.

Experiment 1 tests the role of FD, closure and the VC. Stimuli that vary in FD, closure duration and VCs are created and presented to participants in a three-alternative forced-choice identification task. Experiment 2 is a partial replication of experiment 1, with aspiration included in the stimuli to examine whether it is a perceptual correlate for perceiving /tʂʰ/ responses in /ɤ/ and /a/ contexts.

II. EXPERIMENT 1

A. Method

1. Recording

A 23-year-old male native speaker of Beijing Mandarin was recorded in a quiet room. The speaker did

not report any speech or hearing disorder. He was instructed to read through a list of nine sentences presented in Pinyin Romanization. The target syllables were all embedded in the final position of a carrier phrase—"wo3 shuo1"/uɔ3 ʂuo1/ (meaning "I say" in Mandarin and the number indicating the lexical tone of that syllable: 1 stands for the high level tone and 3 the fall-rise tone). There are nine different target syllables all in high level tone presented in Pinyin Romanization: zhi (/tʂɻ/), chi (/tʂʰɻ/), shi (/ʂɻ/), zhe (/tʂɤ/), che (/tʂʰɤ/), she (/ʂɤ/), zha (/tʂa/), cha (/tʂʰa/), sha (/ʂa/). Each sentence was repeated twice at normal speed, thus a total of 18 sentences were recorded (3 vowels × 3 consonants × 2 repetitions). These sounds were directly recorded into a laptop (Lenovo Z360, Lenovo, China; Sound card: Realtek ALC272, Realtek, Taiwan) using Audacity software via a head-mounted microphone (Philips SHM7110U, Philips, Netherlands) with a sampling rate of 44 100 Hz and 16-bit quantization.

Segmentation was performed by the first author using Praat 5.4.09 (Boersma and Weenink, 2015). For the target syllables, four acoustic events were identified: closure, frication, aspiration and the vocalic period. Closure was identified as the near silent period between the end of the periodic source and the onset of high frequency noise. For affricates, if the initial burst was visible, closure was delimited by the end of periodic wave and the initial burst. The boundaries of frication were identified at the onset of random noise and the onset of voicing in the region of the first formant of the following vowel. If aspiration was present, its boundaries were marked by the cessation of high frequency noise and the onset of voicing in the region of the first formant of the following vowel. While both frication and aspiration are aperiodic noise, they are distinguishable in the waveform and the spectrogram. In the waveform, aspiration is much weaker in amplitude than the preceding frication and the following vocalic segments. In the spectrogram, the spectral energy of frication usually concentrates at high frequency regions, but aspiration tends to show a formant pattern similar to the following vowel, albeit much weaker in amplitude (Clements and Khatiwada, 2007). We had found in our recording the superimposed aspiration (SA) where the aspiration noise blended into the initial glottal pulses of the following vowel (Mikuteit and Reetz, 2007). Since a sharp distinction did not exist between the superimposed portion and the following vowel, we decided to treat the superimposed aspiration as a property of the vocalic segment (Clements and Khatiwada, 2007) and used the onset of voicing to separate aspiration and the vowel (Peterson and Lehiste, 1960). Finally, the vocalic period was marked by the presence of clear formant trajectories. Production data are presented in Table I, and they were also regarded as reference values in stimuli manipulation. The mean duration of the recorded sentences was 726 ms.

TABLE I. Mean and SD of closure duration (Closure), frication duration (F Dur), aspiration duration (Aspiration) and duration of the vocalic segments (V Dur) across nine consonant-vowel combinations.[a]

	Closure (ms)		F Dur (ms)		Aspiration (ms)		V Dur (ms)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
tʂɻ	53	4.24	77	4.22	—	—	201	2.84
tʂɤ	71	12.72	53	7.77	—	—	209	3.56
tʂa	60	1.41	38	0.00	—	—	217	0.07
tʂʰɻ	31	3.95	141	0.84	—	—	185	1.56
tʂʰɤ	31	6.36	105	0.14	34	2.63	210	1.14
Tʂʰa	32	2.12	57	1.20	31	3.21	200	1.34
ʂɻ	—	—	163	0.91	—	—	182	0.35
ʂɤ	—	—	155	0.21	—	—	200	1.90
ʂa	—	—	112	0.07	—	—	188	1.41

[a]Standard deviation (SD).

2. Stimuli

The three acoustic cues examined in this experiment were FD, closure and VCs. FD was treated as a quasi-continuous variable varying in five steps, while closure was treated as binary, either present or absent. The frication was also followed by three different VCs, /ɻ/, /ɤ/, and /a/. A set of 30 stimuli was created with Praat by resynthesizing the recorded natural utterances. All these stimuli were embedded in the final position of a carrier phrase to make the effect of closure salient to listeners.

For the carrier phrase, one sentence was randomly selected. The initial part of the utterance, /uɔ3 ʂuo1/ ("I say"), was used as the carrier phrase for all target syllables in the experiment. The total duration of the carrier phrase was 355 ms with an average f0 of 135 Hz and its intensity was scaled to 75 dB sound pressure level (SPL). This carrier phrase was chosen mainly because it was semantically unpredictable and sounded natural. However, the confounding /ʂ/ in the carrier phrase, which might bias listeners toward certain responses, was taken into consideration when we interpreted our results. Since the exact same carrier phrase was used in both experiments, its influence should remain the same for all stimuli.

The fricative used to generate stimuli was excised from one of the two recorded sentences that contained the target /ʂɻ/ syllable in the final position. Its FD was 158 ms. Since previous studies suggest that amplitude rise time is one of the perceptual cues for distinguishing affricates and fricatives (e.g., Tsao *et al.*, 2006; Howell and Rosen, 1983), a flat amplitude envelope of the fricative /ʂ/ was generated by multiplying the original amplitude envelope with its reciprocal (Mitani *et al.*, 2006). The intensity envelope of /ʂ/ was first generated by Praat's function "To Intensity…" with default settings except that the time step was set to 0.001 s. Then the intensity tier was converted to amplitude tier with Praat's function "To Amplitude Tier," which was then inverted to its reciprocal using the formula "1/self." The inverted amplitude tier and the original fricative were combined by multiplying them together with Praat's "Multiply" function. Then its initial and final portions of 10 ms were chopped off, leaving the middle part of the stable frication as the stimulus. The intensity of the resynthesized frication noise was scaled to 75 dB SPL. The FD continuum varying from 60 to 180 ms in a 30 ms step was constructed with

reference to the data in Table I using "To Manipulation" with Praat's default settings.

To check whether there was any change of spectral properties due to manipulation, we performed linear predictive coding (LPC) analysis of the stimulus and all recorded tokens by using Praat's functions "To Spectrum…" and "LPC smoothing…" (Number of peaks: 22; Pre-emphasis from: 50 Hz) to generate the smoothed spectral envelopes. Figure 1 presents the LPC envelopes of 18 recorded tokens of /tʂ/, /tʂʰ/, /ʂ/ in three VCs and the spectral envelope of the resynthesized fricative. The LPC envelope of the stimuli was highly similar to those of the recorded syllables.

Next, three clear utterances were selected, each containing one of the three target syllables—/ʂʅ/, /ʂɤ/ and /ʂa/. The average f0 of the three utterances was 147 Hz. The vocalic segments of the target syllables, /ʅ/, /ɤ/, /a/ were then excised from the target syllables in the sound editing interface. Manipulation of vocalic segments was done using Praat's function—"To Manipulation…" with the time step set to 0.01 s and the pitch ceiling and floor to 500 and 75 Hz, respectively. Then their duration was normalized to 150 ms, their F0 to 150 Hz, and their average intensity to 75 dB SPL. The first three formant frequencies of the three vocalic segments were as follows: /ʅ/—395, 1768, and 2218 Hz; /ɤ/—519, 1151, and 2851 Hz; /a/—945, 1398, and 2434 Hz.

A silent interval of 30 ms was created by setting its amplitude to 0 Pa. This silent interval was later inserted after the offset of the carrier phrase /uɔ3 ʂuo1/ and before the onset of target fricative /ʂ/ in order to mimic the silence created by closure. Finally, the carrier phrase, a period of silence (if any), frication of varying duration and the vocalic segments were concatenated together. Sample spectrograms are given in Fig. 2.

3. Participants

The participants in experiment 1 were 20 university students in their 20 s (ten males, ten females). They were all from Northern China where Mandarin is spoken. We also assessed the participants' production through talking before the experiments to make sure they can distinguish retroflexes and their non-retroflex counterparts in production. None of them had reported any speech or hearing disorder. After the experiment, they received small gifts for their participation.

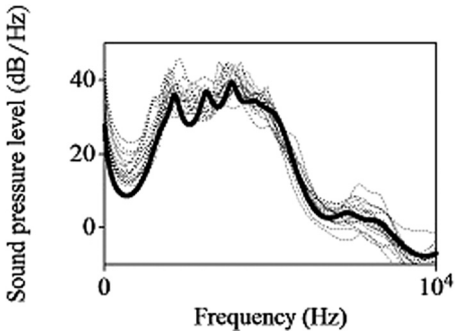

FIG. 1. LPC envelopes of retroflex affricates and fricatives. Thin dotted lines represent the LPC envelopes of the 18 recorded tokens of /tʂ/, /tʂʰ/, /ʂ/ across three VCs, whereas the thick dark line represents the LPC envelope of the resynthesized fricative used in the stimuli.

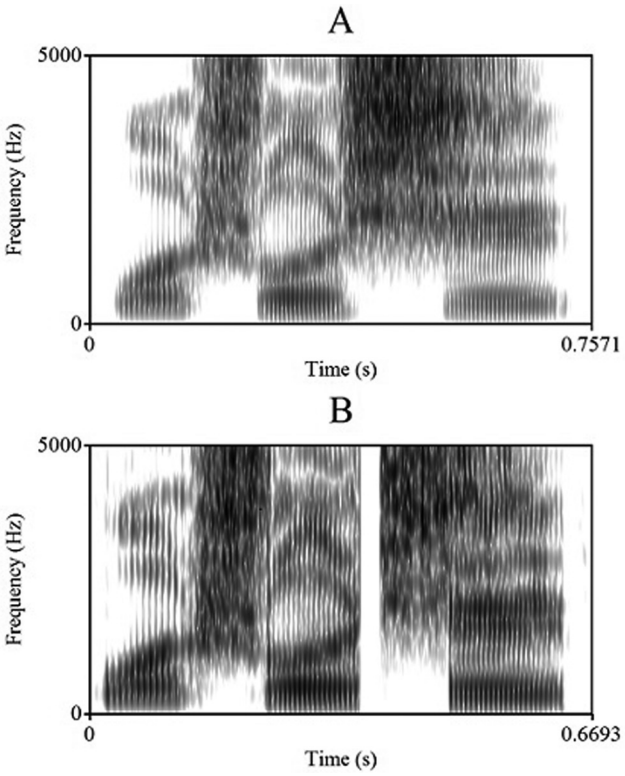

FIG. 2. Sample spectrograms. (a) Spectrogram of a recorded sentence, /uɔ3 ʂuo1 ʂʅ1/. (b) Spectrogram of one of the stimuli in experiment 1. The carrier phrase /uɔ3 ʂuo1/ was followed by 30 ms of silence, 90 ms of frication and the vocalic segment /ʅ/.

4. Procedure

The experiment, conducted in two quiet rooms in a university, used within-subject design and the experimental procedure was programmed using PsychoPy (Peirce, 2007). Before the experiment, instructions were displayed on the computer screen in Chinese and explained to the participants. Then participants were asked to perform a three-alternative forced-choice (3AFC) task that began with a practice task, in which nine of the recorded sentences that contained /tʂ/, /tʂʰ/, /ʂ/ across three VCs were delivered to them binaurally in the same manner as in the formal experiment. Listeners would hear these target syllables embedded in the carrier phrase, /uɔ3 ʂuo1/ ("I say"). For each trial, a fixation point lasting for 1000 ms appeared first in the center of the screen, alerting the participants of the forthcoming sound. Then each stimulus was played once with the screen subsequently presenting horizontally the ordered Pinyin Romanization labels of these three sounds for the participants to choose from (zh for /tʂ/, ch for /tʂʰ/, sh for /ʂ/). Participants were instructed to respond to the audio stimuli by pressing "v,""b," and "n" keys on the keyboard, each corresponding to a category of the stimuli: /tʂ/, /tʂʰ/, and /ʂ/, regardless of the following vowel. Each participant received a total of 150 trials [3 VCs × 2 closure conditions

(CLOs) × 5 steps of FD × 5 repetitions], presented in quasi-random order and separated into five blocks. Each repetition of the 30 stimuli was presented in a single block continuously. Participants were allowed to take untimed breaks between blocks when the program prompted them to choose between taking a break and proceeding to the next block. The entire session lasted from 10 to 12 min.

5. Data analysis

The experimental data were analyzed with the Classification and Regression Tree (CART) (Breiman *et al.*, 1984). The classification tree is a subcategory of CART used to fit categorical data. It works through recursive binary splitting to partition the predictor space into a set of small regions and makes predictions according to the most commonly occurring class of that region (see Chapter 8 in James *et al.*, 2013, for an overview). Results of the CART can be summarized as a set of rules and displayed graphically as tree diagrams, which are easy to interpret. Compared with the widely used multinomial logistic regression model, the classification tree can handle complex relationships between variables without assuming certain distributions and coding dummy variables. We also find classification trees particularly suitable for handling perception data because the process of making predictions by evaluating each data point against a set of rules probably reflects how listeners integrate multiple acoustic cues into a single perceptual entity. One can simply inspect the branches of the tree to see how different acoustic cues interact before a specific prediction is reached. While the classification tree does not return p-values for assessing statistical significance, the model can be assessed through prediction accuracy.

A classification tree (tree 1) was grown to investigate the potential contributions of FD, Closure (CLO, absent vs present) and VC (/ɻ/ vs /ɤ/ vs /a/). Another classification tree, tree 2, was fitted to investigate the possible effects of participants and blocks, with subject and block included as additional categorical predictors. To prevent the model from learning too much from the random noises in the data (overfitting), we increased the threshold for splitting: thirty observations were required to split a node, and at least ten observations were required in all terminal nodes (Shih and Lu, 2015). Then tenfold cross-validation was performed to simplify the tree structure ("tree pruning"), so as to ensure the generalizability and interpretability of the tree. In a tenfold cross-validation, the original data are partitioned randomly into ten folds of approximately equal size. One fold of data is used as a validation set to evaluate the classification accuracy, while the remaining nine folds of data are used to fit the tree. This process is repeated ten times with a different validation set each time and finally the tree that minimizes the cross-validation error is selected (James *et al.*, 2013). These routines in the classification tree analysis were performed with the R packages "rpart" (Therneau *et al.*, 2015) and "rpart.plot" (Milborrow, 2015), implemented in R (R Core Team, 2015). Another R package "ggplot2" was used to produce figures (Wickham, 2009).

B. Results

The results are summarized in Figs. 3(a), 3(b), and 3(c), which are produced by calculating the mean and standard deviation of the percent responses across participants. It is noted that, with the increment of FD, the percent /ʂ/ responses tend to increase [Fig. 3(c)] while the percent /tʂ/ responses show a decreasing trend [Fig. 3(a)]. The presence of closure effectively boosts the affricate responses across different VCs and FD [Figs. 3(a) and 3(c)]. Participants' responses are also modulated by VCs, as the /a/ context has elicited more /ʂ/ responses and fewer /tʂ/ responses than the other VCs [Figs. 3(a) and 3(c)]. However, these three cues

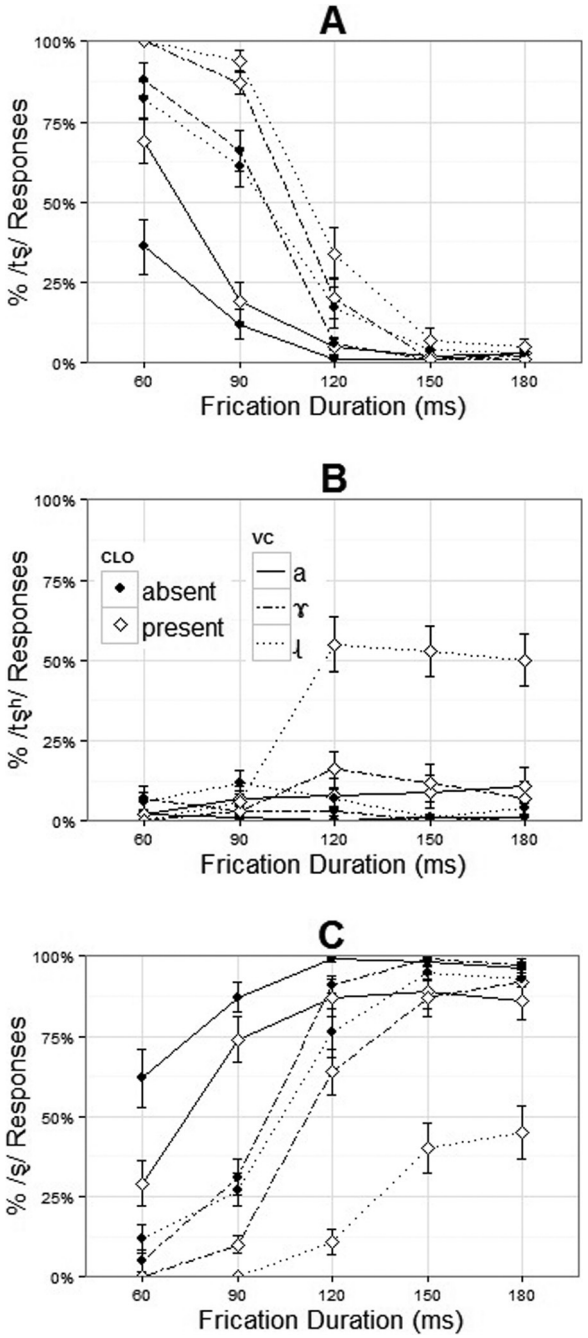

FIG. 3. Results of experiment 1. (a) Mean and SD of the percent /tʂ/ responses; (b) mean and SD of the percent /tʂʰ/ response; (c) mean and SD of the percent /ʂ/ responses. SD is represented by error bars.

seem not to affect the percent /tʂʰ/ responses generally, except that the presence of closure and the /ɻ/ context elicit more /tʂʰ/ responses than other vocalic segments as the FD lengthens [Fig. 3(b)].

The classification accuracy of tree 1 and tree 2 were 82.6% (2477 out of 3000 instances) and 83.5% (2505 out of 3000 instances), respectively, both much higher than chance level (33.3%). In tree 2, block did not contribute to the classification task and subject only slightly improved the classification accuracy. Therefore, only tree 1 is analyzed here (Fig. 4).

Each non-terminal node in the tree represents the threshold used to partition the population into two subgroups. The root node on top represents the threshold for splitting the whole population into two subpopulations, whereas the terminal nodes at the bottom specify the predicted results and classification accuracy. The classification tree can be interpreted as a series of "if-else" statements. If a particular data point satisfies the threshold at one non-terminal node, then it goes to the left branch, otherwise it passes to the right branch, and this process continues until it reaches the terminal node. For instance, the rightmost branch of the tree can be interpreted as the following: if a stimulus has FD shorter than 105 ms and if its VC is not /a/, then it will be directly classified as "zh" (/tʂ/). The number below (678/800) indicates the classification accuracy, with the denominator 800 showing the total number of responses under this condition and the numerator, 678, the number of responses that are correctly identified by the classification tree.

Examination of tree 1 (Fig. 4) shows that FD distinguishes /tʂʰ/ and /ʂ/ from /tʂ/. Stimuli that have longer frication than the threshold of 105 ms are judged either as /tʂʰ/ or /ʂ/. When the FD is shorter than 105 ms, the stimuli is directly classified as /tʂ/ if it does not precede vowel /a/, but even when the frication precedes /a/, the frication necessary for labeling /ʂ/ ($\geq$75 ms) is still longer than that for labeling /tʂ/ ($<$75 ms), if the influence of closure is ignored for the moment. This implies that the FD threshold for separating /ʂ/ and /tʂ/ is susceptible to the changes of the following vocalic segments: the threshold is generally higher in /ɻ/ and /ɤ/ context (approximately 105 ms) than that in /a/ context (approximately 75 ms).

Tree 1 also indicates that closure contributes to the distinction between affricates and fricatives (see Fig. 4). As can be seen from the two nodes that rely on closure to split the subpopulation, given that all previous conditions are satisfied, the presence of closure would bias the classification tree to identify the stimuli as affricates and the absence of which would cause the stimuli to be identified as fricatives.

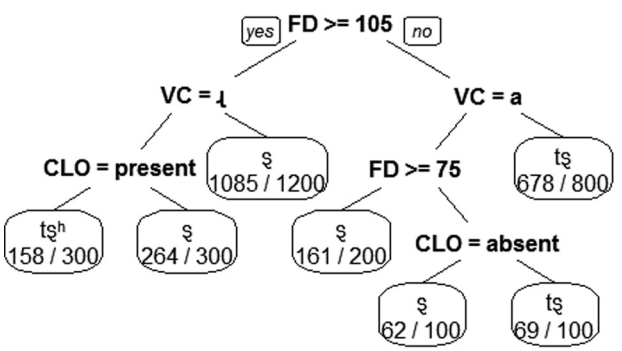

FIG. 4. Tree 1. Classification using FD, VC, and CLO.

The context in which the affricate-fricative series is produced also affects how it is perceived by participants. On the one hand, as is discussed above, the VC affects where the threshold of FD for separating /ʂ/ and /tʂ/ falls along the perceptual continuum. On the other hand, the percept of /tʂʰ/ is also directly influenced by the context in which the sound is presented. It can only be identified when its FD is longer than 105 ms and when it is followed by /ɻ/ and preceded by a period of closure. But the accuracy was only 52.6% (158/300). Inspection of Figs. 3(b) and 3(c) shows that /tʂʰ/ and /ʂ/ are ambiguous under these conditions, presumably because of the influence of the retroflex approximant /ɻ/.

C. Discussion

The results of experiment 1 confirm that FD and closure are perceptual correlates of /tʂ/, /tʂʰ/ and /ʂ/. FD serves to distinguish between the aspirated affricate /tʂʰ/ and the unaspirated affricate /tʂ/, and between the fricative /ʂ/ and the unaspirated affricate /tʂ/. But it cannot distinguish between /tʂʰ/ and /ʂ/. This is consistent with what has been found for the perception of alveolo-palatal affricates and fricatives in Mandarin Chinese (Tsao *et al.*, 2006). Though production data suggest that fricatives tend to be slightly longer than aspirated affricates in Mandarin Chinese (Feng, 1985; Liu *et al.*, 2000; Wu *et al.*, 2015, p. 150), this difference is not exploited to distinguish between aspirated affricates and fricatives. The possible reason may be that the difference in FD between /tʂʰ/ and /ʂ/ is not as salient as that between /tʂ/ and /ʂ/ and between /tʂʰ/ and /tʂ/, as presented in Table I and other studies (e.g., Feng, 1985).

Closure also contributes to the contrast between affricates and fricatives in connected speech. The presence of silence tends to bias the participants to label the stimulus as an affricate [Figs. 3(a), 3(b), and 3(c)], which is reminiscent of previous findings that insertion of silence cues the affricate manner in English both in word-final and sentence-final positions (Repp *et al.*, 1978; Dorman *et al.*, 1980).

In addition, the results of this experiment confirm our hypothesis that VCs affect the perception of the retroflex affricate-fricative series. The influence of VC is twofold. First, it affects the location of categorical boundary between /ʂ/ and /tʂ/. As borne out by the classification tree analysis summarized in Fig. 4, the frication threshold that separates /ʂ/ from /tʂ/ was 105 ms in /ɤ/ and /ɻ/ context but it has been lowered to 75 ms in /a/ context. The gradient changes in perception roughly correspond to the changes in vowel height and roughly corroborates the production data, which showed that frication tends to be shorter before low vowels than before mid and high vowels (Feng, 1985). But this categorical boundary did not seem to differ for /ɤ/ and /ɻ/.

In this experiment, we also find interesting results concerning the perception of the retroflex aspirated affricate, /tʂʰ/. Even when the closure cues the affricate manner, participants only reported the percept of /tʂʰ/ in the /ɻ/ context,

but not in /ɤ/ and /a/ contexts. As the /tʂʰ/ responses in the apical VC are around 55%, which is above the 33% chance level, the response pattern cannot simply be attributed to misclassification due to participants' inattentiveness. Some of the participants reported after the experiment that some stimuli sounded like /tʂʰ/ at the onset but gradually shifted toward /ʂ/ at the offset, especially in the /ɻ/ context. Confusion of /tʂʰ/ with /ʂ/ when preceding /ɻ/ might be due to the gesture continuation which smears out the cue of aspiration.

But it is worth noting that some of the results, such as the lack of /tʂʰ/ responses in /a/ and /ɤ/ contexts [Fig. 3(b)], could be affected by the procedure of stimuli manipulation. Specifically, the frication that was originally generated from /ʂ/ and its flattened amplitude envelope could bias the participants to label the stimuli as /ʂ/. In addition, listeners might categorize the stimuli with reference to the potentially confounding /ʂ/ in the carrier phrase, which may prompt them to be more cautious in classification because the resynthesized frication in the stimuli differs from the potentially confounding /ʂ/. The frication threshold for separating /ʂ/ from /tʂ/ when the frication is not preceding /a/ is located at 105 ms (the root node of tree 1; Fig. 4), possibly because listeners have utilized the duration of the /ʂ/ (100 ms) in the carrier phrase to calibrate the categorical boundary. However, despite these potential biases, the results still indicate that participants are quite sensitive to the acoustic cues under investigation.

III. EXPERIMENT 2

In experiment 2, we investigate the contribution of aspiration to the percept of /tʂʰ/ in /ɤ/ and /a/ contexts. To test the hypothesis that the absence of aspiration has greatly reduced the /tʂʰ/ responses in the /ɤ/ and /a/ contexts, we replicate experiment 1 with aspiration included in /ɤ/ and /a/ contexts to see if the inclusion of this additional cue will boost the /tʂʰ/ response. The vocalic segment /ɻ/ is excluded in experiment 2 because aspiration is not available when /tʂʰ/ precedes /ɻ/ in natural speech. In data analysis, comparable data from both experiments 1 and 2 are pooled together to assess the contributions of these acoustic cues.

A. Method

1. Stimuli

A set of 20 stimuli that varied in FD, CLOs and VCs was generated (2 VCs × 2 closure durations × 5 steps of FD). The same carrier phrase and the same fricative stimulus in experiment 1 were used. The only difference was that the frication continuum ranged from 30 to 150 ms in 30 ms steps in experiment 2. Since experiment 1 has shown that shorter FD is necessary for perceiving /tʂ/ in /a/ context, each step of the FD was reduced by 30 ms, in order to make the three responses more balanced.

The aspiration and vocalic segments concatenated to the carrier phrase were recreated. First, two clear utterances that contained /tʂʰa/ and /tʂʰɤ/, respectively, were selected. Then, for the two syllables /tʂʰa/ and /tʂʰɤ/, the aspiration and the vocalic segments that immediately followed the frication were spliced from the original recording as a whole. As the aspiration before vowel /ɤ/ was 33 ms and that before /a/ was 35 ms, aspiration was normalized to 30 ms. Next, the vowels in the spliced sound clips were normalized to 150 ms in duration, their f0 to 150 Hz and their average intensity to 75 dB SPL as in experiment 1. The first three formant frequencies of the three vocalic segments were as follows: /ɤ/—469, 1115, and 2705 Hz; /a/—929, 1451, and 2452 Hz.

The same silent interval in experiment 1 was used. Finally, the carrier phrase, a period of silence (if any), frication of varying duration, the sound clips containing aspiration and the following vocalic segments were concatenated together. Sample spectrograms are given in Fig. 5.

2. Participant

The same participants in experiment 1 participated in experiment 2.

3. Procedure

The experimental procedure was the same as that in experiment 1. Each of the 20 stimuli was presented binaurally with five repetitions, resulting in a total of 100 trials (2 VCs × 2 CLOs × 5 steps of FD × 5 repetitions).

4. Data analysis

Four classification trees were constructed to investigate the contributions of each acoustic cue and possible effects of blocks and participants. The settings for fitting classification

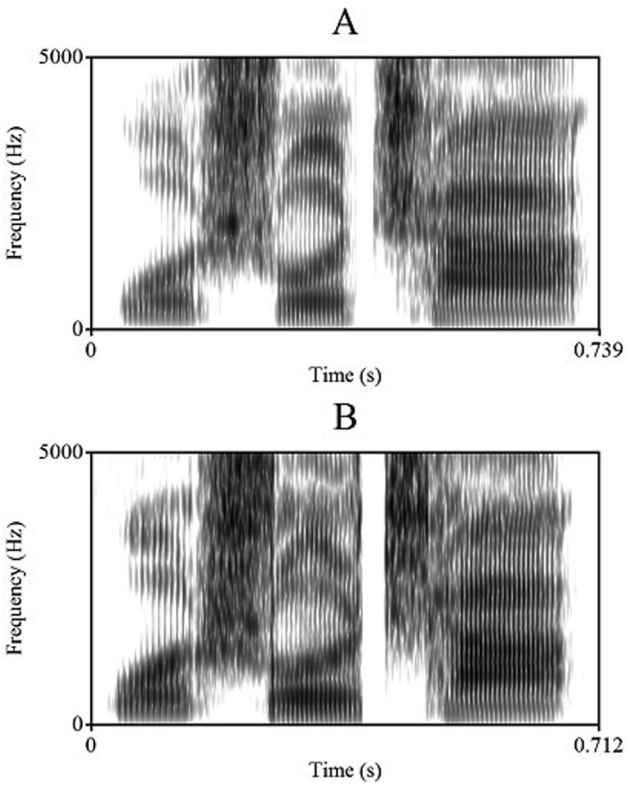

FIG. 5. Sample spectrograms. (a) Spectrogram of /uɔ3 ʂuo1 tʂʰa1/. (b) Spectrogram of one of the stimuli in experiment 2. The carrier phrase /uɔ3 ʂuo1/ was followed by 30 ms of silence, 60 ms of frication, 30 ms of aspiration and the vocalic segment /a/.

trees were the same as those in experiment 1. All data from experiment 2 were used to train tree 3. Identification responses (/tʂ/ vs /tʂʰ/ vs /ʂ/) were treated as the dependent variables, while VC (VC, /ɻ/ vs /ɤ/ vs /a/), FD, and CLO (absent vs present) were included as predictors. Tree 4 was an expanded version of tree 3 with subject and block included as additional predictors.

To fit tree 5 and tree 6, we pooled together data in /a/ and /ɤ/ contexts from both experiment 1 and 2. In order to make the data comparable, we excluded data points involving FD of 180 and 30 ms from the analysis. A new predictor, aspiration condition (ASP), was included in fitting tree 5 and tree 6. Stimuli from experiment 1 (which did not have aspiration) were coded as unaspirated, and data from experiment 2 were coded as aspirated. In tree 5, FD, VC (/ɻ/ vs /ɤ/ vs /a/), CLO (absent vs present) and ASP (absent vs present) were treated as predictors. Tree 6 was an expanded version of tree 5 with subject and block as additional predictors.

B. Results

The results are summarized in Fig. 6 by computing the mean and standard deviation of the percent responses across participants. The influence of VCs is quite noticeable, with the responses elicited by the /a/ context being predominantly /tʂʰ/ [Fig. 6(b)]. The responses are distributed more evenly across three categories in the /ɤ/ context [Figs. 6(a), 6(b), and 6(c)]. While the effects of FD and closure remain largely similar to those reported in experiment 1, the inclusion of aspiration in stimuli seems to alter the trade-off between these cues. For instance, closure seems to contribute little to the /tʂ/ responses [Fig. 6(a)].

The classification accuracy of tree 3 was 74.4% (1487 out of 2000 instances) and that of tree 4 was 79.0% (1579 out of 2000 instances), suggesting some effects of individual variability. Since tree 3 (Fig. 7) has already accounted for a large proportion of data and has a much simpler tree structure, it is used in the following analysis.

The influence of VC is noticeable, as the /a/ context strongly biases participants towards perceiving /tʂʰ/ such that tree 3 classifies every target consonant that precedes the vowel /a/ as /tʂʰ/. In contrast, if the target consonant is in /ɤ/ context, FD distinguishes /tʂ/ from /tʂʰ/ and /ʂ/, as target consonants with FD shorter than 75 ms were directly classified as /tʂ/. Closure is only used to separate affricates from fricatives when frication is longer than 75 ms, suggesting that this acoustic cue may be redundant when frication is brief.

To examine the role of aspiration, tree 5 and tree 6 were grown. The classification accuracy rate of tree 5 was 77.8% (2488 out of 3200 instances) and that of tree 6 was 83.4% (2668 out of 3200 instances). Block had no influence on the classification accuracy but the inclusion of Subject as a predictor did improve the accuracy by about 6%. As presented in Fig. 8, tree 5 is a combination of tree 3 and part of tree 1. We only focus on the factor aspiration here. Following the left branches, a stimulus with aspiration and in /a/ context would be directly classified as /tʂʰ/. If the target syllable in a stimulus is in /ɤ/ context, in addition to aspiration, it should

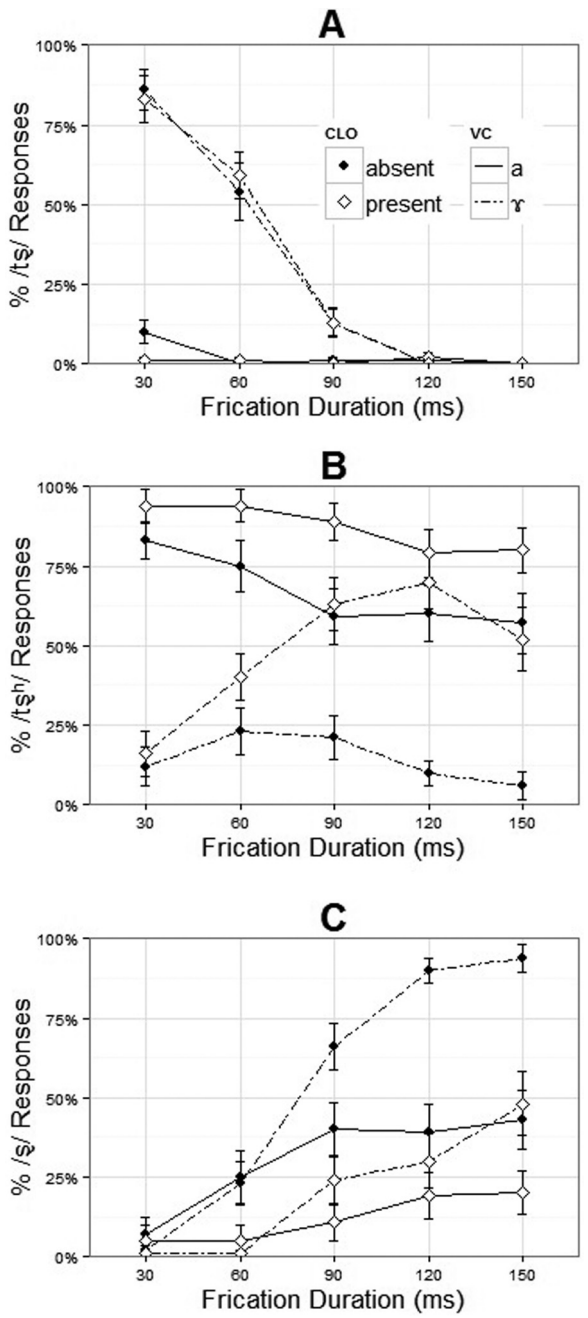

FIG. 6. Results of experiment 2. (a) Mean and SD of the percent /tʂ/ responses; (b) mean and SD of the percent /tʂʰ/ responses; (c) mean and SD of the percent /ʂ/ responses. SD is represented by error bars.

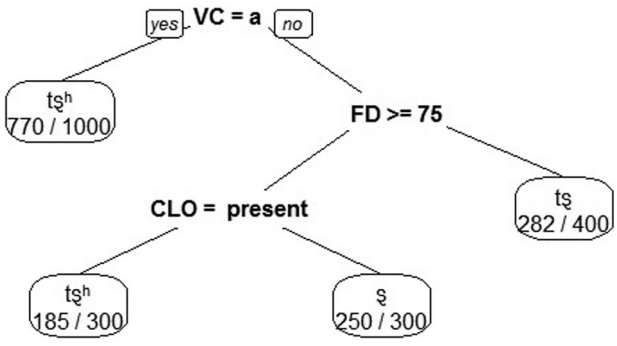

FIG. 7. Tree 3.Classification using FD, VC, and CLO.

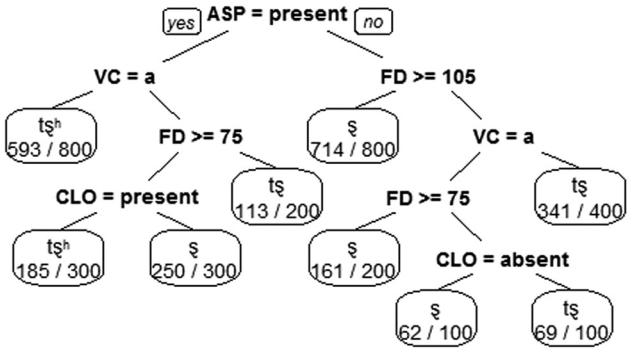

FIG. 8. Tree 5. Classification using FD, VC, ASP, and CLO.

also have a stretch of frication longer than 75 ms and the presence of closure to be classified as /tʂʰ/. These results indicate that aspiration is necessary for perceiving aspirated affricate in /a/ and /ɤ/ context.

C. Discussion

Results of tree 3 and tree 5 presented above generally show that aspiration contributes substantially to the percept of /tʂʰ/, suggesting that it is an acoustic correlate of the retroflex aspirated affricate in /a/ and /ɤ/ context. Tree 5 again reveals that closure cues the affricate manner, and that FD serves to differentiate unaspirated affricates and fricatives.

It is worth noting that tree 5 classified all stimuli in /a/ context as /tʂʰ/ with an accuracy rate of 74% (593/800), which may suggest that aspiration strongly biased participants to perceive /tʂʰ/ in /a/ context. That these vocalic segments were originally excised from those preceded by aspirated affricates may also contribute to these classification results. Since we have decided to treat the superimposed aspiration as a property of the vocalic segments, the aspiration on the initial portion of the vowel might further strengthen the bias towards aspirated affricates. But even if aspiration is present, listeners still perceive /tʂ/ and /ʂ/ from time to time and tend to give more /tʂ/ and /ʂ/ responses in /ɤ/ context than in /a/ context. This can be explained by the phonetic trading relations between these cues, in which aspiration is overridden by short frication in the /ɤ/ context, biasing listeners to give /tʂ/ responses. By the same token, /ʂ/ responses may be a direct result of aspiration overridden by long and stable frication and absence of closure. But this trading relation between cues seems to differ in terms of VCs, as Figs. 6(a) and 6(c) indicate that frication may override aspiration more frequently before /ɤ/ than before /a/.

IV. GENERAL DISCUSSION

The current study explores how FD, closure, and aspiration are utilized to cue the three-way contrast between /tʂ/, /tʂʰ/, and /ʂ/ in Mandarin Chinese, and whether VCs modulate the utilization of these acoustic cues.

With regard to the role of acoustic cues, results of our experiments show that FD, closure and aspiration jointly cue affricate-fricative contrast in Mandarin Chinese, confirming some of the findings about these cues in Mandarin and other languages (e.g., Dorman *et al.*, 1980; Tsao *et al.*, 2006; Repp *et al.*, 1978). Specifically, both /tʂʰ/ and /tʂ/ can be distinguished from /ʂ/ by closure in continuous speech. Aspirated affricates can be further cued by the presence of aspiration in /a/ and /ɤ/ context. FD is found to be an acoustic cue that serves to distinguish unaspirated retroflex affricates from aspirated ones, and unaspirated retroflex affricates from retroflex fricatives, a finding reported by Tsao *et al.* (2006) when they examine the contrast between the alveolo-palatal series.

The finding about the contribution of FD may have implications for the production study of Mandarin affricates. It is reported that the unaspirated retroflex affricate /tʂ/ is extremely short, almost comparable to an aspirated stop (Qi and Zhang, 1982; Ran, 2007). The extremely short /tʂ/ in Mandarin Chinese is probably shaped by the pressure to maximize the phonological contrast. Since FD cues the contrast between /tʂ/ and /tʂʰ/ as well as the contrast between /tʂ/ and /ʂ/, the short FD can make /tʂ/ more distinct from the other two homorganic consonants. As /tʂʰ/ tends to be slightly shorter than /ʂ/ (e.g., Feng, 1985; Liu *et al.*, 2000), /tʂ/ must be shortened even further so as to be distinguishable from /tʂʰ/. A similar case has also been reported by Utman and Blumstein (1994). The labio-dental fricative /f/ in Ewe has higher amplitude of frication than the same fricative in English, presumably because Ewe speakers have to maintain the contrast between labio-dental fricative /f/ and bilabial fricative /Φ/, a phoneme not present in English. Our results may provide yet another piece of evidence showing that speech perception and production shape each other within a phonological system.

Our results regarding the influence of VCs on the acoustic cues reveal the trade-off between the place and the manner contrasts. Lee-Kim (2014) suggests that one possible advantage of the retroflex and dental approximant, i.e., /ɻ/ and /ɻ̪/, is to maximize the place contrast between retroflex and dental fricatives in Mandarin. Since the gesture remains almost unchanged from the preceding fricative to the following homorganic approximant, the approximant contains additional cues to the place of articulation of the preceding homorganic fricative, thereby enlarging the perceptual distance between sibilants of different places of articulation.

We argue that this perceptual advantage is not without cost. While the homorganic approximant enhances the place contrast, it also increases the ambiguity between /tʂʰ/ and /ʂ/, rendering these two sounds confusable to some extent. As the feature [+aspiration] is realized as lengthened frication instead of aspiration when /tʂʰ/ precedes /ɻ/ (Wu *et al.*, 2015, p. 159), the absence of aspiration can diminish the distinction between /tʂʰ/ and /ʂ/ on some listening conditions. The perceptual implication is that listeners have to dispense with one less acoustic cue in perceiving the contrast between /tʂʰ/ and /ʂ/ in /ɻ/ context. As seen in the leftmost branch in tree 1 (Fig. 4), frication longer than 105 ms compounded with closure might cue /tʂʰ/ in the /ɻ/ context, but the low classification accuracy (158/300) indicates that this classification result is associated with relatively high uncertainty. The confusability between /tʂʰ/ and /ʂ/ may be the reason why frication preceded by a short period of silence was more

often judged as /tʂʰ/ in /ɻ/ context than in other VCs [Fig. 3(b)]. If the offset of /tʂʰ/ is similar to /ʂ/, it is possible that listeners tend to rely more on the consonant onset to distinguish /tʂʰ/ from /ʂ/.

Finally, both experiments 1 and 2 indicate that listeners show remarkable perceptual sensitivity to coarticulatory effects induced by the VC. The threshold of FD that separates /tʂʰ/ and /ʂ/ from /tʂ/ is found to be 105 ms in /ɤ/ and /ɻ/ context but this threshold is lowered to 75 ms in /a/ context (Figs. 4, 7, and 8), which seems to parallel the production data in Table I and in Feng (1985). This is reminiscent of the results obtained by Nearey and Rochet (1994), who found an overall correlation between production and perception of VOT in stops in both English and French. Our study demonstrates that participants can also compensate for the temporal difference in frication induced by different VCs, even when the vocalic segment occurs later temporally than the frication in a syllable. Though it remains unclear whether such compensation is achieved through directly parsing the signal into speech gestures or through analyzing the auditory objects (see Beddor, 2015, for a brief summary of the debate between gestural and auditory theories), listeners' compensation for coarticulation shows that they are capable of tracking the fine phonetic details encoded in the speech signal in a way that roughly parallels the dynamics of articulation.

ACKNOWLEDGMENTS

We thank Mo Li for recording all the stimuli and Ning Yan, Yue Wang, Zhe Chen, and Zichong Yin for their assistance. We are also grateful to Ruixue Ma, Qi Hao, and others for kindly participating in our experiments. Last but not least, we would like to express our utmost gratitude to the editor of JASA and three anonymous reviewers for their insightful comments, which have contributed substantially to the improvement of this manuscript.

Beddor, P. S. (**2015**). "Experimental phonetics," in *The Oxford Handbook of Linguistic Analysis*, edited by B. Heine and H. Narrog, 2nd ed. (Oxford University Press, Oxford, UK), pp. 503–524.

Boersma, P., and Weenink, D. (**2015**). "Praat: Doing phonetics by computer (version 5.4.09) [computer program]," http://www.praat.org/ (Last viewed 3 June 2015).

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (**1984**). *Classification and Regression Trees* (Chapman and Hall/CRC, Boca Raton, FL), pp. 1–368.

Clements, G. N., and Khatiwada, R. (**2007**). "Phonetic realization of contrastively aspirated affricates in Nepali," in *Proceedings of ICPhS XVI*, pp. 629–632.

Dorman, M. F., Raphael, L. J., and Isenberg, D. (**1980**). "Acoustic cues for a fricative-affricate contrast in word-final position," J. Phonet. **8**, 397–405.

Duanmu, S. (**2007**). *The Phonology of Standard Chinese*, 2nd ed. (Oxford University Press, Oxford, UK), pp. 34–35.

Feng, L. (**1985**). "Duration of consonants, vowels and lexical tones in Beijing Mandarin (in Chinese)," in *Experimental Studies of Beijing Mandarin*, edited by J. Lin and L. Wang (Peking University Press, Beijing, China), pp. 131–195.

Gerstman, L. J. (**1956**). "Noise duration as a cue for distinguishing among fricative, affricate, and stop consonants," J. Acoust. Soc. Am. **28**(1), 160.

Hedrick, M. (**1997**). "Effect of acoustic cues on labeling fricatives and affricates," J. Speech, Lang. Hear. Res. **40**(4), 925–938.

Howell, P., and Rosen, S. (**1983**). "Production and perception of rise time in the voiceless affricate/fricative distinction," J. Acoust. Soc. Am. **73**(3), 976–984.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (**2013**). *An Introduction to Statistical Learning* (Springer, New York), pp. 307–336.

Kluender, K. R., and Walsh, M. A. (**1992**). "Amplitude rise time and the perception of the voiceless affricate/fricative distinction," Percept. Psychophys. **51**(4), 328–333.

Lee, W. S., and Zee, E. (**2003**). "Standard Chinese (Beijing)," J. Int. Phonet. Assoc. **33**(1), 109–112.

Lee-Kim, S. I. (**2014**). "Revisiting Mandarin 'apical vowels': An articulatory and acoustic study," J. Int. Phonet. Assoc. **44**(3), 261–282.

Lin, T., and Wang, L. (**1992**). *An Introduction to Phonetics* (in Chinese) (Peking University Press, Beijing, China), pp. 43–45.

Liu, H. M., Tseng, C. H., and Tsao, F. M. (**2000**). "Perceptual and acoustic analysis of speech intelligibility in Mandarin-speaking young adults with cerebral palsy," Clinic. Linguist. Phonet. **14**(6), 447–464.

Mikuteit, S., and Reetz, H. (**2007**). "Caught in the ACT: The timing of aspiration and voicing in East Bengali," Lang. Speech **50**(2), 247–277.

Milborrow, S. (**2015**). "rpart.plot: Plot 'rpart' models: An enhanced version of 'plot.rpart' [computer program]," https://cran.r-project.org/web/packages/rpart.plot/index.html (Last viewed 20 May 2016).

Mitani, S., Kitama, T., and Sato, Y. (**2006**). "Voiceless affricate/fricative distinction by frication duration and amplitude rise slope," J. Acoust. Soc. Am. **120**(3), 1600–1607.

Nearey, T. M., and Rochet, B. L. (**1994**). "Effects of place of articulation and vocalic context on VOT production and perception for French and English stops," J. Int. Phonet. Assoc. **24**(1), 1–18.

Ohala, J. J. (**1983**). "The origin of sound patterns in vocal tract constraints," in *The Production of Speech*, edited by P. F. MacNeilage (Springer, New York), pp. 189–216.

Peirce, J. W. (**2007**). "PsychoPy—Psychophysics Software in Python," J. Neurosci. Methods **162**(1), 8–13.

Peterson, G. E., and Lehiste, I. (**1960**). "Duration of syllable nuclei in English," J. Acoust. Soc. Am. **32**(6), 693–703.

Qi, S., and Zhang, J. (**1982**). "Analysis of consonant durations in Mandarin Chinese (in Chinese)," Acta Acust. **7**(1), 8–13.

Ran, Q. (**2007**). "On the nature of the apical affricate in Mandarin Chinese: From the perspective of duration (in Chinese)," Chin. Linguist. **3**, 69–77.

R Core Team (**2015**). "R: A language and environment for statistical computing," http://www.R-project.org (Last viewed 20 May 2016).

Repp, B. H. (**1982**). "Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception," Psychol. Bull. **92**(1), 81–110.

Repp, B. H., and Liberman, A. M. (**1987**). "Phonetic category boundaries are flexible," in *Categorical Perception*, edited by S. Harnad (Cambridge University Press, Cambridge, UK), pp. 89–112.

Repp, B. H., Liberman, A. M., Eccardt, T., and Pesetsky, D. (**1978**). "Perceptual integration of acoustic cues for stop, fricative, and affricate manner," J. Exp. Psychol. Hum. Percept. Perform. **4**(4), 621–657.

Shih, C., and Lu, H. Y. D. (**2015**). "Effects of talker-to-listener distance on tone," J. Phonet. **51**, 6–35.

Therneau, T. M., Atkinson, B., and Ripley, B. (**2015**). "rpart: Recursive partitioning [computer program]," https://cran.r-project.org/web/packages/rpart/index.html (Last viewed 20 May 2016).

Tsao, F. M., Liu, H. M., and Kuhl, P. K. (**2006**). "Perception of native and non-native affricate-fricative contrasts: Cross-language tests on adults and infants," J. Acoust. Soc. Am. **120**(4), 2285–2294.

Utman, J. A., and Blumstein, S. E. (**1994**). "The influence of language on the acoustic properties of phonetic features: A study of the feature [strident] in Ewe and English," Phonetica **51**(4), 221–238.

Wickham, H. (**2009**). *ggplot2: Elegant Graphics for Data Analysis* (Springer, Amsterdam, the Netherlands), pp. 1–213.

Wu, Z., and Lin, M. (**1989**). *Experimental Phonetics: A Summary* (in Chinese) (Higher Education Press, Beijing, China), pp. 135–149.

Wu, Z., Lin, M., and Bao, H. (**2015**). *Experimental Phonetics: A Summary (Expanded Edition)* (in Chinese) (Peking University Press, Beijing, China), pp. 153–163.