Effects of speaker variability and noise on Mandarin fricative identification by native and non-native listeners

Chao-Yang Lee, a) Yu Zhang, and Ximing Li
Division of Communication Sciences and Disorders, Ohio University, Athens, Ohio 45701

Liang Tao and Z. S. Bond
Department of Linguistics, Ohio University, Athens, Ohio 45701

(Received 6 September 2011; revised 3 April 2012; accepted 26 May 2012)

Speaker variability and noise are two common sources of acoustic variability. The goal of this study was to examine whether these two sources of acoustic variability affected native and non-native perception of Mandarin fricatives to different degrees. Multispeaker Mandarin fricative stimuli were presented to 40 native and 52 non-native listeners in two presentation formats (blocked by speaker and mixed across speakers). The stimuli were also mixed with shape-specified noise to create five levels of signal-to-noise ratios. The results showed that noise affected non-native identification disproportionately. By contrast, the effect of speaker variability was comparable between the native and non-native listeners. Confusion patterns were interpreted with reference to the results of acoustic analysis, suggesting native and non-native listeners used distinct acoustic cues for fricative identification. It was concluded that not all sources of acoustic variability are treated equally by native and non-native listeners. Whereas noise compromised non-native fricative perception disproportionately, speaker variability did not pose a special challenge to the non-native listeners.

I. INTRODUCTION

A fundamental issue in speech perception is how listeners process acoustic variability to uncover speakers’ intended sounds. The process may be explicated by examining the sources of acoustic variability (e.g., variability arising from speaker differences), characteristics of the listeners (e.g., native vs. non-native), and how the two factors interact. In other words, the nature of speech perception may be elucidated by examining how sources of acoustic variability affect speech perception by listeners of various characteristics.

In this study, we examined the perception of multi-speaker Mandarin fricatives by native Mandarin listeners and non-native (English-speaking) listeners in quiet and noisy listening conditions. Our goal was to evaluate whether speaker variability and noise would affect native and non-native perception to different degrees. Intuitively, acoustic variability would affect non-native listeners to a greater extent because of their limited knowledge of the target language. However, there is evidence that not all sources of acoustic variability affect non-native speech perception disproportionately. Because speaker variability and noise are two of the most common sources of acoustic variability, investigating how they affect native and non-native speech perception would contribute to our understanding of speech perception.

A. Processing speaker variability and noise by native and non-native listeners

Acoustic characteristics of phonologically identical utterances can be vastly different across speakers. Therefore, how listeners uncover speakers’ intended sounds despite such variation is a foundational issue in speech perception (Johnson, 2005). Although non-native perception is usually assumed to be more susceptible to acoustic variability, there is some evidence that speaker variability affects, in fact, native and non-native perception similarly. For example, Bradlow and Pisoni (1999) and Takayanagi et al. (2002) showed that the effect of speaker variability on auditory word recognition was comparable between native and non-native listeners. In the suprasegmental literature, Lee et al. (2009) and Lee et al. (2010a) also showed that speaker variability disrupted Mandarin tone identification similarly for native and non-native listeners. It appears that speaker variability does not present a special challenge to non-native listeners. Furthermore, the literature on learning non-native sound contrasts also showed that training with multi-speaker stimuli enables non-native learners to generalize learning to unfamiliar speakers (Lively et al., 1993; Hardison, 2003). This finding suggests that learners develop exemplar-based, speaker-specific sound representations before shifting attention to contrastive aspects of non-native sound categories (Lively et al., 1993). That is, adapting to different speakers seems to be an ability possessed equally by native and non-native listeners.

By contrast, many studies showed disproportionate difficulty of non-native listeners in speech recognition in other adverse conditions such as noise (Mayo et al., 1997; Meador et al., 2000; Nábělek and Donahue, 1984; Takata and Nábělek, 1990; Van Wijngaarden et al., 2002). While this is consistent with the idea that non-native listeners have more difficulty in processing acoustic variability, there is also evidence that noise does not disrupt non-native phoneme recognition to a greater degree. Cutler et al. (2004) showed that
the effect of noise on phoneme recognition is comparable between native (English) and non-native (Dutch) listeners. In an investigation of Mandarin tone identification by native and non-native listeners, Lee et al. (2010a) also showed that the effects of noise and language background did not interact, suggesting that non-native tone perception was not affected to a greater extent than native tone perception. Cutler et al. (2004) suggest that the commonly reported disproportionate difficulty of non-native perception in adverse conditions does not arise from difficulties at the phoneme identification level. Bradlow and Alexander (2007) further suggest that because non-native difficulties with noise accumulate across all levels of spoken language comprehension, the disproportionate effect of noise on non-native perception may not surface at a relatively low level of processing (e.g., identifying segmental phonemes).

Although these interpretations seem to provide a coherent account of the noise findings by resorting to the level of processing, it is still not clear why certain adverse conditions, but not others, affect native and non-native speech perception differently. This point can be illustrated by a series of studies on Mandarin tone perception by native and non-native (English-speaking) listeners in various adverse conditions. In particular, fragmented acoustic input disrupted non-native tone perception disproportionately (Gottfried and Suiter, 1997; Lee et al., 2008, 2009, 2010b). In these studies, fragmented stimuli were constructed by removing various parts of a syllable. For example, in “silent-center” stimuli, the majority of the syllable center was digitally removed such that only the onset and offset of a syllable was present. These stimuli were then presented to native and non-native listeners for tone identification. The results showed that non-native listeners were affected by the incomplete acoustic input to a greater extent than native listeners were. Similarly, the absence of a tonal context also compromised non-native tone perception to a greater extent (Gottfried and Suiter, 1997; Lee et al., 2008, 2009, 2010b). By contrast, neither speaker variability (Lee et al., 2009, 2010a) nor noise (Lee et al., 2010a) affected non-native tone perception disproportionately. The tasks used in these studies were relatively simple and did not involve complex linguistic processing (e.g., tone identification from isolated syllables), yet non-native listeners were affected disproportionately by certain adverse conditions, but not by others.

An obvious contrast between the tone and segmental studies is the nature of the phonetic categories to be processed by the non-native listeners. Lexical tones are not used phonemically in English and are novel suprasegmental contrasts to English-speaking listeners. Consonants and vowels, by contrast, do exist in both tone and non-tone languages. Segmental and suprasegmental contrasts also involve distinct acoustic correlates. Therefore, for non-tone language users, mapping lexical tones onto their native phonological system is not as straightforward as mapping segmental contrasts. Consequently, findings from tone perception by non-native listeners may or may not generalize to segmental perception by non-native listeners. A direct comparison between tone and segmental perception using the same research paradigm should provide an answer to the question of whether acoustic variability affects tone and segmental perception similarly.

In this study, we extended our previous Mandarin tone studies on the effects of speaker variability and noise (Lee et al., 2009, 2010a) to an investigation of Mandarin fricative perception by native and non-native (English) listeners. As in the tone studies, we presented the fricative stimuli in two presentation formats (blocked by speaker and mixed across speakers) to evaluate the effect of speaker variability. The stimuli were mixed with five levels of speech-shaped noise to evaluate the effect of noise. With the same experimental design and target population as in our previous tone studies, but focusing on segmental contrasts, this study was expected to further clarify the role of speaker variability and noise in speech perception by native and non-native listeners.

B. Fricatives in Mandarin and English

Fricative consonants are used in both Mandarin and English. Therefore, how Mandarin fricatives are perceived by English listeners may be characterized by current models of second-language (L2) speech perception such as the Speech Learning Model (SLM) (e.g., Flege, 1995) and the Perceptual Assimilation Model (PAM) (e.g., Best, 1995). Although SLM has focused on L2 speech acquisition by experienced listeners and PAM has focused on L2 speech perception by naïve listeners (Best and Tyler, 2007), both models characterize the formation of L2 phonetic categories in terms of the relationship between L2 and the listeners’ native sound inventories. Therefore, the fricatives in the two languages will be described next. Our discussion will focus on voiceless fricatives and the place of articulation distinction among them. Voiced fricatives were not included in this study because there is only one voiced fricative in Mandarin ([z]), as opposed to four in English) and it is often regarded as an approximant [j] (e.g., Ladefoged and Disner, 2012, p. 190).

There are five voiceless fricatives in Mandarin: Labiodental [f], alveolar [s], palatalized post-alveolar (alveolo-palatal) [ɻ], flat post-alveolar (retroflex) [ɻ], and velar [x]. They are represented by f, s, x, sh, and h in the Pinyin Romanization system for Mandarin and will be referred to as such throughout the text. The place of articulation for the fricatives is specified primarily by where the maximal constriction is made. X-ray tracings and palatograms suggest that tongue shape and the presence of a sublingual cavity are also relevant for place distinctions among the sibilant fricatives (Ladefoged and Maddieson, 1996, p. 151).

English also has five voiceless fricatives: Labiodental [f], dental [θ], alveolar [s], palato-alveolar [ʃ], and glottal [h]. As in Mandarin, the place of articulation is also specified by the location of the maximal constriction in the vocal tract. The status of the [h] sound has been debated. Some authorities (e.g., Ladefoged and Maddieson, 1996, pp. 325–326) do not consider [h] a consonant because it lacks the articulatory and acoustic characteristics of a typical consonant (e.g., definite displacement of formant frequencies). This is perhaps why comprehensive acoustic studies on English fricatives (e.g., Jongman et al. 2000; McMurray and Jongman, 2011) did not include the glottal sound in their analyses. Because
the Mandarin velar fricative [x] is acoustically and perceptually similar to the [h] sound and our acoustic analysis (see Sec. II A) was intended as a direct comparison with Jongman et al. (2000), the Mandarin velar fricative was not included in this study.

A comparison between the two fricative inventories shows that English has a non-sibilant fricative [θ] that Mandarin does not have. On the other hand, Mandarin has two post-alveolar sibilant fricatives ([c] and [ʃ]), whereas English has only one post-alveolar fricative ([ʃ]). For English speakers learning Mandarin, the mapping of the Mandarin non-sibilant [f] to the English counterpart should be relatively straightforward because the L2 category is essentially the same as the L1 category. Therefore, both SLM and PAM should predict that the Mandarin [f] will be assimilated to the English counterpart and no new categories need to be formed.

For the sibilants, however, the learners will have to map the three-way distinction in Mandarin ([s]-[ʃ]-[ʃ]) to existing two-way distinction in English ([s]-[ʃ]). Presumably the Mandarin [s] can be assimilated to the English [s] quite easily. But the post-alveolar space, once occupied by [ʃ] only, will have to accommodate two Mandarin sibilants, meaning an additional category will have to be created in the perceptive space. According to PAM, the [c]-[ʃ] contrast will be particularly difficult to perceive because both sounds will map into an existing category. Although SLM differs from PAM in assuming that the capacity to learn language-specific properties is accessible to L2 learners of all ages (Flege, 2007, p. 366), both models appear to make the same predictions regarding how the Mandarin fricatives will be perceived by English listeners.

C. The present study

Neither SLM nor PAM makes explicit predictions about the effects of speaker variability and noise on L2 speech perception. As in traditional speech perception theories, the implicit assumption is that acoustic variability that does not result in phonemic distinctions is pre-processed to generate abstract phonemic representations. It is this abstract phonemic representations that are the object of discussion in these models. However, processing non-phonemic acoustic variability is an integral part of speech perception. There is evidence that detailed information about speakers and allophonic variations affects speech perception. For example, Creel et al., 2008; McLennan and Luce, 2005). A recent study by McMurray and Jongman (2011) specifically showed that the ability to identify speaker and vowel contributes to the identification of English fricatives. Therefore, investigating the effect of acoustic variability on native and non-native speech perception is expected to contribute to our understanding of speech perception.

In summary, the purpose of the present study was to evaluate whether speaker variability and noise would affect Mandarin fricative identification by native and non-native listeners to different degrees. Speaker variability was manipulated by how the multi-speaker stimuli were presented, i.e., blocked by speaker or mixed across speakers. Previous research showed that speech recognition is more accurate and/or faster when stimuli are presented in the blocked format (e.g., Creelman, 1957). The mixed presentation format makes speaker identity unpredictable from trial to trial, requiring listeners to constantly adapt to different speakers. Because speaker adaptation is an active process that requires cognitive resources, recognition performance is expected to be compromised by the mixed presentation format (Johnson, 2005; Mullenix et al., 1989). The effect of noise was evaluated by mixing the fricative stimuli with speech-shaped noise to create five levels of signal-to-noise ratios (SNR). As would be expected, speech recognition performance usually deteriorates as SNR decreases. The critical question was whether non-native perception would be disproportionately affected by these two sources of acoustic variability.

II. METHOD

A. Materials

Mandarin syllables fa [fa], sa [sa], xia [cu], and sha [sa] with the high-level tone (tone 1) were recorded by six adult Beijing Mandarin speakers (three females and three males; age M = 27 years, SD = 3). The recordings were made in a double-wall sound booth (IAC) with a professional microphone (Shure SM81-LC). The microphone was connected through a preamplifier and analog-to-digital converter (Sound Devices USBPre Microphone Interface) to a personal computer. To control for speaking rate, the Chinese characters and Pinyin Romanization representing the syllables were presented visually to the speakers via Microsoft PowerPoint at 5 s per slide. The speakers were instructed to read the syllables in citation form. The recordings were digitized with the Brown Lab Interactive Speech System (BLISS) (Murtex, 2000) at 20 kHz and 16-bit quantization. Each syllable was identified from the BLISS waveform display, excised from the master recordings, and saved as an audio file. Peak amplitude was normalized across the audio files.

The audio files were then digitally mixed with speech-shaped noise (Nilsson et al., 1994) using MATLAB (The MathWorks, Natick, MA) to generate stimuli at five SNRs: −15 dB, −10 dB, −5 dB, 0 dB, and quiet (i.e., no noise added). These SNRs were selected to allow a direct comparison to our previous study on Mandarin tone identification by native and non-native listeners at the same SNRs (Lee et al., 2010a). The noise duration was identical to the duration of the target syllables. A total of 120 stimuli (4 syllables × 6 speakers × 5 SNRs) were generated. Details about the speech-shaped noise can be found in Nilsson et al. (1994, p. 1087). In brief, a long-term average spectrum was calculated from 336 English sentences spoken by a male speaker, using a Hanning window with continuous root-mean-square averaging. A finite impulse response (FIR) filter was then constructed based on the spectrum levels at 126 frequencies between 0 and 10 kHz. Semi-random white noise was synthesized, filtered with the FIR filter, and scaled to the same amplitude as the speech.

To provide an acoustic description of the Mandarin fricatives used in this study, we conducted acoustic analysis on the 24 tokens produced by the 6 speakers. Eleven acoustic measures were selected, including six spectral measures.
(spectral peak location, four spectral moments, and F2 frequency at vowel onset), four amplitude measures (normalized amplitude and relative amplitude in the F3, F4, and F5 ranges), and a duration measure (normalized fricative duration). Our selection of the acoustic measures was guided by the literature on English fricative acoustics (Jongman et al., 2000; McMurray and Jongman, 2011) and the acoustic theory of speech production (Fant, 1960; Stevens, 1998). In particular, we selected acoustic measures that have been demonstrated to classify English fricatives that have relatively well-defined relationships to fricative production, and that are relevant to the classification of the fricatives used in this study (i.e., place of articulation). Details on the acoustic analysis and data processing procedure can be found in the Appendix.

A summary of the acoustic analysis results is shown in Table I. English data from Jongman et al. (2000) are also included for comparison. A number of observations can be made. First, none of the acoustic measures single-handedly distinguished all four places of articulation for the Mandarin fricatives. The only acoustic measures that came close were spectral skewness and F2 frequency at vowel onset. Both distinguished five out of the six pair-wise comparisons. Nonetheless, every place distinction is associated with at least two significant acoustic measures. This finding contrasts with Jongman et al. (2000), who showed that all of their acoustic measures successfully distinguished sibilant from non-sibilant fricatives in English.

B. Participants

Ninety-two adults participated in this study. All participants were screened for normal hearing, defined as pure-tone, air-conducted thresholds of ≤20 dB HL at octave frequencies from 1000–4000 Hz. The native listeners

<table>
<thead>
<tr>
<th>Mandarin (This study)</th>
<th>f-s</th>
<th>f-¢</th>
<th>f-§</th>
<th>s-¢</th>
<th>s-§</th>
<th>c-§</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral peak location</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spectral mean</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spectral variance</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spectral skewness</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spectral kurtosis</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2 frequency at vowel onset</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normalized amplitude</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>Relative amplitude ΔA3</td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relative amplitude ΔA4</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relative amplitude ΔA5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Normalized fricative duration</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>English (Jongman et al., 2000)</th>
<th>fv-¢§</th>
<th>fv-sz</th>
<th>fv-j3</th>
<th>0j-sz</th>
<th>0j-j3</th>
<th>sz-j3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectral peak location</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>Spectral mean</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spectral variance</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spectral skewness</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spectral kurtosis</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Locus equations</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F2 frequency at vowel onset</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normalized amplitude</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relative amplitude (ΔA3 or ΔA5)</td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Normalized fricative duration</td>
<td></td>
<td></td>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
included 40 Mandarin speakers recruited from the Ohio University community (26 females and 14 males; age $M = 25$ years, $SD = 4$). All identified Mandarin as their native language. The 52 non-native listeners were Mandarin language students at Ohio University (27 females and 25 males; age $M = 21$ years, $SD = 3$). They included 31 first-year students, 11 second-year students, and 10 third-year students. At the time of the experiment, the first-, second-, and third-year students had received approximately three, six, and nine quarters of Mandarin instruction. None of the non-native participants were heritage students or had substantial exposure to Mandarin prior to enrolling in the Mandarin language courses.

C. Procedure

BLISS was used for stimulus presentation and response data acquisition. Stimuli were delivered in two formats: blocked by speaker and mixed across speakers. In the blocked presentation, the 120 stimuli were assigned to 6 blocks such that each block included 20 stimuli from a particular speaker (4 fricatives $\times 5$ SNRs). Each participant was assigned a uniquely randomized order of stimulus presentation within a block. The order of presentation of the blocks was also randomized. In the mixed presentation, all 120 stimuli were mixed across speakers. As in the blocked condition, each participant received a uniquely randomized order of stimulus presentation. Half of the participants were given the blocked presentation first and the other half of the participants were given the mixed presentation first. In both blocked and mixed presentations, the five SNRs were randomized. A short break was given between the two presentations. In sum, both presentation format and SNR were within-subject factors, i.e., all participants received both presentations at all five SNRs.

Participants were tested individually in an IAC sound booth. The stimuli were calibrated with a sound-level meter (Larson Davis System 824) and presented binaurally via a set of headphones (Koss R80) at $73 \pm 3$ dB SPL. The participants were told that they would be listening to the syllables $fa$, $sa$, $xia$, and $sha$, with the high-level tone, produced by three female and three male speakers. They were also told that some of the syllables would sound quite noisy. The participants were instructed to identify the syllables they heard by pressing computer keys labeled with $fa$, $sa$, $xia$, and $sha$. All participants indicated that they were familiar with the Pinyin Romanization system.

We had intended to use Chinese characters as response key labels because a word recognition task is presumably more relevant to spoken language comprehension than the syllable identification task. However, many of the non-native participants had not learned to read the Chinese characters representing the four syllables. Therefore, we decided to use Pinyin labels for all non-native participants and half of the native participants. For the other half of the native participants, the keys were labeled with the Chinese characters \( \text{ü} \) (for “sand”), \( \text{š} \) (for “spread”), \( \text{x} \) (for “shrimp”), and \( \text{š} \) (for “sand”) to evaluate whether the response format would make a difference in identification performance.

The participants were told that they had 5 s to respond to each stimulus and that their responses would be timed; therefore they should respond as quickly as possible. Speeded-response tasks are used extensively in psycholinguistic research to tap into the on-line nature of spoken language comprehension (e.g., Cutler and Chen, 1997). The reaction time data were expected to corroborate data from the accuracy analysis.

III. RESULTS

A. Accuracy

Figure 1 shows the percentage of correct identification by all participants in the two presentation formats and five SNRs. The percentages were arcsine-transformed before parametric tests were conducted. An analysis of variance (ANOVA) was conducted on arcsine-transformed accuracy with a presentation format (blocked and mixed) and SNR ($-15$ dB, $-10$ dB, $-5$ dB, $0$ dB, and quiet) as within-subject factors, Mandarin experience (non-native, native, and native responding to Chinese character labels) as a between-subject factor, and participants as a random factor. When a main effect from the ANOVA was significant, the Bonferroni post hoc test was conducted for pair-wise means comparisons to keep the family-wise Type I error rate at 5%.

The ANOVA revealed a significant main effect of presentation format, $F(1, 89) = 12.69, p < 0.001$. As predicted, identification was more accurate in the blocked presentation than in the mixed presentation. The main effect of SNR was also significant, $F(4, 356) = 372.11, p < 0.001$. As expected, tone identification accuracy decreased as SNR decreased. All pair-wise means comparisons were significant. The main effect of Mandarin experience was also significant, $F(2, 89) = 14.88, p < 0.001$. Not surprisingly, the native listeners were more accurate than the non-native listeners. There were no differences between the two native groups, indicating that the response format did not make a difference in accuracy.

![FIG. 1. Mean accuracy (±SE) of fricative identification by non-native listeners ($N = 52$), native listeners ($N = 20$), and native listeners responding to Chinese character labels ($N = 20$).](image-url)
TABLE II. Proportion of reaction time data (in %) excluded from analysis in the perceptual experiment for non-native listeners (N = 1248), native listeners (N = 480), and native listeners responding to Chinese character labels (N = 480).

<table>
<thead>
<tr>
<th>Presentation format</th>
<th>SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listener group</td>
<td>−15 dB</td>
</tr>
<tr>
<td>Blocked</td>
<td></td>
</tr>
<tr>
<td>Non-native</td>
<td>33</td>
</tr>
<tr>
<td>Native</td>
<td>30</td>
</tr>
<tr>
<td>Native (character)</td>
<td>27</td>
</tr>
<tr>
<td>Mixed</td>
<td></td>
</tr>
<tr>
<td>Non-native</td>
<td>37</td>
</tr>
<tr>
<td>Native</td>
<td>33</td>
</tr>
<tr>
<td>Native (character)</td>
<td>35</td>
</tr>
</tbody>
</table>

There were two significant interactions, including SNR × Mandarin experience, $F(8, 356) = 2.73, p = 0.006$, and SNR × Mandarin experience × presentation format, $F(8, 356) = 2.74, p = 0.006$. The highest-order effect showed that in the blocked presentation, both native and non-native listeners showed a linear decline as SNR decreased. The difference between native and non-native accuracy was particularly prominent at −5 dB SNR. In the mixed presentation, the native listeners responding to character labels still showed a linear decline. However, for native and non-native listeners responding to Pinyin labels, the accuracy was comparable between 0 and −5 dB SNR. In addition, the difference between native and non-native difference was greater at 0, −5, and −10 dB SNR than at quiet and −15 dB SNR. That is, the difference between native and non-native performance was more prominent in the intermediate SNRs than in the easiest and most difficult listening conditions.

B. Reaction time

Reaction time was measured from stimulus onset. Because the duration of individual stimuli varied, we also analyzed reaction time measured from stimulus offset to evaluate whether the stimulus duration difference affected response patterns. Reaction times were log-transformed before parametric tests were conducted. Only correct responses were included in the reaction time analysis. Reaction times exceeding two SDs above or below the means were excluded from the analysis. Table II shows the percentage of data excluded from the reaction time analysis.

As in the accuracy analysis, an ANOVA was conducted with presentation format and SNR as within-subject factors, Mandarin experience (non-native, native, and native responding to Chinese character labels) as a between-subject factor, and participants a random factor. When a main effect from the ANOVAs was significant, the Bonferroni post hoc test was conducted for pair-wise means comparisons to keep the family-wise Type I error rate at 5%.

Figure 2 shows the reaction time of fricative identification responses. Because the onset and offset analysis yielded the same statistical conclusions, only the reaction time measured from stimulus onset is shown in Fig. 2. Overall, the reaction time results were consistent with the accuracy data in that higher accuracy was associated with shorter reaction time. The ANOVA revealed a significant main effect of presentation format [onset: $F(1, 89) = 15.54, p < 0.001$; offset: $F(1, 89) = 13.28, p < 0.001$]. As would be expected from the accuracy results, identification was faster in the blocked presentation than in the mixed presentation. The main effect of SNR was also significant [onset: $F(4, 356) = 159.94, p < 0.001$; offset: $F(4, 356) = 168.51, p < 0.001$]. All pair-wise means comparisons were significant, indicating that reaction time increased as SNR decreased. The main effect of Mandarin experience was significant [onset: $F(2, 89) = 18.32, p < 0.001$; offset: $F(2, 89) = 21.57, p < 0.001$]. As expected, the two native groups responded faster than the non-native listeners. In the offset analysis, the difference between the two native groups ($p = 0.0198$) narrowly missed statistical significance in the Bonferroni test (critical $p = 0.0167$) (cf. $p = 0.03$ in the onset analysis). That is, although the response format did not make a difference in response accuracy for the native listeners, there was some indication that the labels of the response keys affected the speed of identification. This finding underscores the potential value of the reaction time measure in revealing processing differences. There were no other effects.

C. Confusion patterns

To evaluate the types of identification errors made by the participants, contingency tables were generated to examine fricative confusion patterns. Because the two native groups (i.e., those responding to Pinyin labels and those responding to Chinese character labels) did not show substantially different response accuracy patterns, data from the two native groups were combined in the confusion analysis. Figures 3 and 4 show the confusion patterns for the native and non-native listeners, respectively.

Numbers shown in these $4 \times 4$ contingency tables are percentages. In a given cell, the actual counts were summed...
up across all members of a group before being converted to percentages.

For the native listeners, the confusion patterns were similar between the blocked and mixed presentations. Identification of f and x was quite accurate (mostly over 80%) throughout the five SNRs. Identification of s dropped below 80% at −10 dB SNR and below 70% at −15 dB SNR. As noise increased, s was frequently misidentified as sh, except in the mixed presentation at −15 dB SNR, where s was misidentified most frequently as f. For sh, identification accuracy dropped below 80% at −10 dB SNR and below 70% at −15 dB SNR. The sh stimuli were commonly misidentified as x, except in the mixed presentation at −10 dB SNR, where sh was misidentified frequently as x and s.

The confusion patterns for the non-native listeners were similar to the native patterns in several ways: (1) The confusion patterns were similar between the blocked and mixed presentations; (2) identification of f and x was quite accurate (over 80%) throughout the five SNRs; (3) identification of s dropped below 80% at −10 dB SNR and below 70% at −15 dB SNR; (4) the sh stimuli were misidentified as x most frequently. By contrast, there were two notable differences from the native data: (1) The s stimuli were commonly misidentified as x and f (instead of being misidentified as sh by the native listeners); (2) identification of sh dropped below 80% as soon as noise was added (whereas native accuracy did not drop below 80% until −10 dB SNR).

To evaluate the association between the fricative stimuli and responses shown in the contingency tables, $\chi^2$ tests were conducted on the confusion data. The null hypothesis was that the stimuli were not associated with the responses. The results are shown in Table III. All 20 tests indicated that the null hypothesis should be rejected, i.e., all listeners were able to identify the stimuli above chance in both presentation formats at all SNRs.

IV. DISCUSSION

The research question asked in this study was whether speaker variability and noise would affect non-native
fricative perception disproportionately. As would be expected, speaker variability and noise adversely affected fricative identification. Fricative identification responses were less accurate and slower in the mixed presentation format. Fricative identification performance also deteriorated as more noise was mixed with the signals. However, we did not find evidence that the effects of speaker variability and Mandarin background interacted, suggesting that the decline in performance due to speaker variability was comparable between the native and non-native listeners. In other words, our strategy of changing test materials from Mandarin tones to fricatives did not reveal any disproportionately negative effect of speaker variability on non-native perception. This result is consistent with previous research on English word recognition (Bradlow and Pisoni, 1999; Takayanagi et al., 2002) and Mandarin tone identification (Lee et al., 2009, 2010a). Taken together, there is no evidence that speaker variability affects non-native speech perception disproportionately.

By contrast, a significant SNR × Mandarin experience interaction was found in the accuracy measure, indicating that noise affected native and non-native listeners differently. The SNR × Mandarin experience interaction arose from the fact that non-native listeners were affected to a greater extent by noise at intermediate (0, −5, and −10 dB) SNRs, but not under the easiest (quiet) or most difficult (−15 dB SNR) listening conditions. In other words, there is a range of SNRs in which the processing difference between native and non-native listeners was most prominent. The smaller difference between the native and non-native accuracy under the quiet condition may reflect a ceiling effect. In particular, the overall accuracy for the non-native listeners in this condition was 96% (averaged across blocked and mixed presentations), indicating that the non-native listeners had little difficulty in identifying the fricatives when the speech signals were clear. As soon as noise was added, however, non-native performance was affected to a greater degree than native performance. As the noise level increased, the difference between the native and non-native listeners was gradually neutralized. Our speculation is that the speech signals became too heavily masked by noise at the low SNRs for the difference between native and non-native perception to surface. In other words, the smaller difference in the easiest and most difficult noise conditions may be due to two distinct reasons.

The emergence of the significant SNR × Mandarin experience interaction contrasted with the tone study by Lee et al. (2010a). It should be noted, however, that an SNR × Mandarin experience interaction did surface when Lee et al. (2010a) divided their non-native participants into subgroups based on baseline performance (i.e., tone identification accuracy in the quiet, blocked presentation). In particular, their native listeners and high-performance non-native listeners showed a steady decline in tone identification performance, whereas the mid- and low-performance non-native listeners showed more variable response patterns. Moreover, a follow-up study in our lab with the same experimental design also showed a reliable SNR × Mandarin experience interaction. Taken together, these findings suggest that non-native listeners are affected disproportionately by noise, whether they are perceiving tones or fricatives.

The disproportionate noise effect contrasts with Cutler et al. (2004) finding on English phoneme identification in noise by English and Dutch listeners. Consequently, the disproportionate noise effect found in this study does not support the idea that non-native listeners’ disproportionate difficulty in adverse conditions arises from difficulties at levels other than phoneme identification (Cutler et al., 2004; Bradlow and Alexander, 2007). A comparison between this study and Cutler et al. (2004) shows that the discrepancy could have come from the following sources: First, the type and level of noise: Cutler et al., 2004 used multi-speaker babble as the masker presented at 0, 8, and 16 dB SNR, whereas the present study used speech-shaped noise presented at −15, −10, −5, and 0 dB SNR. The relatively heavy noise and/or the speech-shaped noise used in the present study could have allowed the disproportionate noise effect to surface. Second, the scope of stimuli: Cutler et al. (2004) used a fairly comprehensive set of English syllables encompassing 24 consonants and 15 vowels, whereas the present study focused on four Mandarin fricative consonants presented in a single vowel context. Given the considerable number of phonemes involved in the identification task, one would think that the task of Cutler et al. (2004) would be more challenging and more likely to reveal a disproportionate noise effect. However, it was the present study that showed the effect. Third, the level of experience with the non-native language: The non-native (Dutch) listeners in Cutler et al. (2004) were all fluent in the target language (English), whereas the non-native (English) participants in the present study were relatively inexperienced in the target language (Mandarin). A better command of the non-native language in Cutler et al. (2004) could have neutralized the disproportionate noise effect.

Given the multiple contrasts between the two studies, it is not possible to pinpoint the exact source of the opposing conclusions. However, a direct comparison could be made between the present study and our previous tone study.
(Lee et al. 2010a) because both studies used the same experimental design and the participants were drawn from the same population (Mandarin language students at Ohio University). Non-native listeners were substantially better at identifying fricatives than tones. Overall, response accuracy was higher and reaction time was shorter in the present study than in the tone study. Take, for example, the baseline performance, defined as identification accuracy in the quiet, blocked condition. Average tone identification accuracy in this condition was 75% (SD = 22) (Lee et al., 2010a), whereas average fricative identification accuracy reached 97% (SD = 4) (the present study). It is clear that tone identification was a more challenging task for the non-native listeners, who also showed much more individual variability in their ability to identify tones. As noted, lexical tones are not used phonemically in English, and are novel contrasts for the English-speaking listeners. Therefore, tonal contrasts are more difficult to process.

By contrast, the fricative identification task involved mapping the Mandarin fricatives onto existing English fricative categories. Descriptively, the two languages share two common fricatives (f and s). According to SLM and PAM, the task for the non-native listeners was simply to interpret the remaining two fricatives (x and sh) with reference to existing English categories. One expected consequence of the mapping is that the common fricatives (f and s) would be identified more accurately than the “new” fricatives (x and sh). The confusion data provided only partial support for this prediction. As expected, f was identified quite well. However, s was not. Furthermore, x was identified well too, which is contrary to the hypothesized relative ease of identification. It is likely that the mapping is not just based on how well the absolute match is between fricatives in the two languages. Rather, the confusion pattern suggests that the internal category structure of the fricatives (i.e., the relationship among the fricatives) played a role in how the Mandarin fricatives were perceived.

Our acoustic analysis provided a basis for interpreting the confusion patterns. Because the acoustic analysis was limited to the 24 tokens produced in the clear, our interpretation of the noise results should be considered preliminary. First, the fact that both groups of listeners identified f very well (over 80% accurate even in the most challenging condition) makes sense because both languages have the f sound. Furthermore, it is the only non-sibilant in Mandarin and would not be confused with other non-sibilants. The acoustic finding that f was distinguished well from all other fricatives is consistent with this observation.

Second, the perceptual results showed that identification of s was not as robust as that of f. For both groups of listeners, identification of s dropped below 80% at −10 dB SNR and below 70% at −15 dB SNR. At first sight, the relatively low accuracy for s is rather counterintuitive because s is a common fricative that exists in both languages. In addition, s as a sibilant is acoustically stronger than the f sound. Therefore, the relatively low accuracy most likely arose from confusion with other fricatives. The s stimuli were commonly misidentified as sh by the native listeners, whereas they were misidentified as x and f by the non-native listeners. As summarized in Table 1, s and sh were distinguished by six acoustic measures, whereas s and x were distinguished by only two measures. Based on the number of acoustic measures associated with the distinctions, one would expect x to be misidentified more frequently as s. The non-native listeners did show this pattern, but the native listeners did not, suggesting that the two groups of listeners attended to different acoustic cues. In particular, for the native listeners, they did not misidentify s as x as often, suggesting that they did take advantage of the acoustic measures that distinguished s and x. On the other hand, they misidentified s as sh, suggesting that they did not take advantage of the acoustic measures that distinguished s and sh. The only acoustic measure that distinguished s-x, but not s-sh, was F2 frequency at vowel onset. That is, the native listeners appeared to have assigned more weight to F2 frequency at vowel onset than the non-native listeners did. Note also that this is one of the acoustic measures that distinguished the largest number of Mandarin fricative contrasts.

Third, the perceptual results showed that both groups of listeners identified x very well (over 80% accurate even in the most challenging condition). In addition, the confusion data showed that when x was misidentified, the errors were quite evenly distributed between s and sh, indicating that both native and non-native listeners were using a similar set of acoustic cues to identify the distinctions between x and the other two sibilants. It is not clear, however, why x was identified with a higher accuracy than the other two sibilants. Our speculation is that the articulatory gesture of palatalization in x carried over to the following vowel, as indicated by the high value of F2 at vowel onset, and became an “enhancing” feature of the sound (Stevens et al., 2004). When the “defining” feature of fricatives (i.e., friction noise characteristics) was masked by noise, listeners were able to identify the x sound by taking advantage of the enhancing feature (high F2 frequency at vowel onset) that might be less susceptible to the masking noise. This proposal could be tested in future studies by examining friction noise identification without the following vowel. In particular, when information about F2 frequency at vowel onset becomes unavailable, the advantage of x may disappear. Furthermore, McMurray and Jongman’s (2011) study on English fricative identification also suggests that vowel context is useful because listeners may obtain speaker information from the vowel, based on which fricative identity may be estimated.

Finally, the perceptual results for sh showed that native accuracy did not drop below 80% until −10 dB SNR, whereas non-native accuracy dropped below 80% as soon as noise was added, indicating that non-native sh perception is particularly error-prone under noisy conditions. For both groups of listeners, sh was most commonly misidentified as x, although the confusion pattern was much more prominent for the non-native participants. Note, however, that the confusion was not symmetrical, i.e., the x sound was not misidentified as sh as often. This result suggests that when the non-native listeners encountered a Mandarin post-alveolar fricative, they had a tendency to identify it as x instead of sh. It is commonly reported by Mandarin instructors that x is harder for students to learn than the other fricatives. The bias
toward x responses may reflect an over-precaution of the non-native listeners as they attempted to distinguish between x and sh.

V. CONCLUSION

Speaker variability and noise adversely affected Mandarin fricative identification by native and non-native listeners. Identification was less accurate and took longer when the stimuli were presented mixed across speakers than blocked by speaker. Similarly, identification was less accurate and took longer as more noise was added. Whereas noise affected non-native identification disproportionately, we did not find evidence that speaker variability affected native and non-native identification differently. Compared to our previous Mandarin tone identification studies that employed the same experimental design, the present results showed that fricative identification was less challenging for the non-native listeners. The confusion pattern, with reference to acoustic analysis results, showed that native and non-native listeners attended to different acoustic cues in identifying the fricatives. Results from this series of studies on tones (Lee et al., 2008, 2009, 2010a, b) and fricatives (the present study) reaffirmed the observation that not all sources of acoustic variability are treated equally in speech perception. Whereas noise affected non-native perception disproportionately, speaker variability did not seem to pose a special challenge to non-native listeners.

ACKNOWLEDGMENTS

This research was partially supported by Research Challenge Award RC-11-002 from Ohio University. We thank two anonymous reviewers for their helpful comments.

APPENDIX

1. Acoustic analysis procedure and data processing

Acoustic analysis was conducted with the BLISS (Mertus, 2000) and MATLAB. BLISS was used to measure spectral peak location, F2 frequency at vowel onset, relative amplitude in specific frequency ranges, and normalized fricative duration. MATLAB was used to calculate spectral moments and normalized amplitude.

To obtain spectral peak location, a 40-ms full Hamming window was placed at the center of the frication noise. The discrete Fourier transform (DFT)-linear predictive coding (LPC) analysis in BLISS was then used to generate spectra with 98% pre-emphasis and 24 poles. The spectral peak with the highest amplitude on the DFT spectrum was identified and verified by consulting the LPC spectrum. The frequency of the peak was measured as the spectral peak location. In terms of the articulatory-acoustic relationship, the spectral peak location is associated with the lowest resonant frequency of the front cavity. Therefore, this frequency is expected to be higher for fricatives with a more front place of articulation.

Spectral moments were calculated with MATLAB. Fast Fourier transforms (FFTs) were conducted with a 40-ms full Hamming window at three locations: Onset, middle, and offset of the fricative. The four moments (mean, variance, skewness, and kurtosis) were calculated from the FFT spectra. Because spectral moments are statistical measures characterizing the overall shape of the spectrum, their relationship to articulation is not as direct as the relationship between spectral peak location and front cavity resonance (for a summary, see Jongman et al., 2000, p. 1253).

F2 frequency at vowel onset was obtained by placing a 23.3-ms full Hamming window at the beginning of the first glottal pulse. The frequency of F2 was identified from the DFT spectrum and verified by consulting the LPC spectrum. In terms of the articulatory-acoustic relationship, F2 is associated with the resonant frequency of the cavity posterior to the constriction. Measuring F2 frequency at vowel onset therefore provides information about relative tongue body position at the moment of releasing the fricative constriction.

Normalized amplitude was calculated with MATLAB. Root-mean-square amplitude was computed for the entire frication noise and the entire vowel portion of the syllable. The difference between the two portions was calculated as the normalized amplitude (cf. Jongman et al., 2000, where vowel amplitude was measured by averaging over three consecutive glottal periods at the point of maximum vowel amplitude).Normalized amplitude provides information about the relative strength of the noise source and the glottal source during fricative and vowel production. Therefore, one would expect that non-sibilant fricatives have a lower normalized amplitude than sibilant fricatives.

To measure relative amplitude in specific frequency ranges, a 23.3-ms full Hamming window was placed at the center of the frication noise; another window of the same size was placed at vowel onset. The DFT-LPC analysis in BLISS was used to generate a spectrum for the fricative and another spectrum for the vowel. The two spectra were compared to identify spectral peaks that were considered to represent natural frequencies of the vocal tract in the F3, F4, and F5 ranges. The amplitude of the spectral peaks was measured. The amplitude difference between the fricative and vowel in those specific frequency ranges was calculated as the relative amplitude. The differences among fricatives in this measure can be predicted from the front cavity resonance excited by the noise source. For example, [ʃ] is associated with acoustic energy in the F3 range, whereas [s] is associated with energy in the F5 or F6 range. Consequently, the amplitude of [s] relative to the vowel in the F3 range would be smaller than the relative amplitude of [ʃ] in that frequency range. The perceptual relevance of the relative amplitude difference has been demonstrated by Hedrick and Ohde (1993) and Stevens (1985).

Finally, for normalized fricative duration, the fricative-vowel boundary for each syllable was identified from the BLISS waveform display. Because all four fricatives were voiceless, the boundary was set at the beginning of the first glottal pulse. The fricative noise duration and syllable duration were measured. Normalized fricative duration was calculated by dividing the fricative noise duration by the syllable duration.

For each acoustic measure, an ANOVA was conducted with place of articulation (f, s, x, and sh) and speaker gender (female and male) as within-subject factors, and speakers as
a random factor. When a main effect from the ANOVAs was significant, the Bonferroni post hoc test was used for pairwise means comparisons to keep the family-wise Type I error rate at 5%.


