Speech Communication 136 (2022) 63–75

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

Recognition of vocoded speech in English by Mandarin-speaking English-learners

Jing Yang [a,*], Andrew Wagner [b], Yu Zhang [c], Li Xu [b]

[a] Communication Sciences and Disorders, University of Wisconsin-Milwaukee, Milwaukee, United States
[b] Communication Sciences and Disorders, Ohio University, United States
[c] Communication Sciences and Disorders, Oklahoma State University, United States

ARTICLE INFO

Keywords:
Vocoded speech
Non-native listeners
Phoneme recognition
Sentence perception

ABSTRACT

The purpose of this study was to examine the impact of spectral degradation on speech processing in non-native listeners. The participants included 27 native English (L1) listeners and 43 native Mandarin listeners who learned English as a second language (L2). The speech stimuli included 12 English vowels embedded in a /hVd/ context, 20 English consonants embedded in a /Ca/ context, and HINT, CUNY, and R-SPIN sentences. All stimuli were processed using 2-, 4-, 6-, 8-, and 12-channel noise vocoders. The results showed that compared to the L1 listeners, the L2 listeners demonstrated less improvement in phoneme recognition with increasing number of channels, which was associated with the phoneme confusions due to the impact of their native language. Both consonant and vowel recognition made significant contributions to sentence recognition in the L2 listeners. In addition, the L2 listeners were less effective than the L1 listeners in applying contextual information and linguistic knowledge to sentence recognition. However, the facilitating role of contextual cues in sentence recognition was consistently present in the L2 listeners but they required more spectral information to maximize the contextual benefit in comparison to the L1 listeners. The overall perceptual performance of the L2 listeners was positively correlated with and predicted by the length of residence in the U.S.

1. Introduction

Numerous previous studies have reported that speech perception by native listeners remains robust in various adverse listening conditions including noisy environment, temporally-interrupted, or spectrally-degraded speech (e.g., Assmann and Summerfield, 2004; Bashford et al., 1996; Shannon et al., 1995; Warren, 1970). To non-native listeners, understanding speech materials in a new language is much more challenging because their speech perception is constrained by phonemic confusion, limited vocabulary pool, and incomplete linguistic knowledge of the new language (Lecumberri et al., 2010). Additionally, speech input in the real world is always distorted in a variety of ways, which exacerbates the difficulty of non-native speech perception.

According to the most commonly acknowledged frameworks in speech recognition (e.g., Cohort model, Marslen-Wilson, (1987); Shortlist model, Norris, (1994); TRACE model, McClelland and Elman, (1986)), speech processing starts with receiving auditory information in the form of low-level phonetic features of individual phonemes and ends with extracting meaning and constructing high-level representation of utterances. During this process, there are two types of mechanisms that represent different directions of information flow. The bottom-up processing mechanism emphasizes that the information necessary for speech perception resides in the acoustic signal. When this mechanism is applied, listeners analyze low-level acoustic information, and then map or integrate it into a larger picture. Previous studies revealed that consonants and vowels of speech sounds carry distinct acoustic cues that make different contributions to speech intelligibility (Fogerty and Humes, 2010; Fogerty et al., 2012; Kewley-Port et al., 2007). In isolated words, consonants and vowels show relatively equal contribution to the word intelligibility. However, in a sentence, vowels carry unique speech information that play a more determining role in sentence recognition. The top-down processing mechanism emphasizes that listeners apply high-level linguistic context and stored knowledge to exclude plausible alternative speech units and to identify the most appropriate candidates for speech comprehension. In normal-hearing native listeners, the two systems are highly interactive and function in parallel (Dell and Newman, 1980; Tyler et al., 2000). However, the role of bottom-up and top-down mechanisms in processing non-native language materials

* Corresponding author.
E-mail address: jyang888@uwm.edu (J. Yang).

https://doi.org/10.1016/j.specom.2021.11.008
Received 21 April 2020; Received in revised form 8 April 2021; Accepted 29 November 2021
Available online 1 December 2021

remains controversial (Field, 2004). Some researchers suggested that non-native listeners, especially those in the early stage of second language learning, rely heavily on bottom-up information due to the lack of language experience and prior knowledge of the new language (Hansen and Jensen, 1994). Other researchers proposed that non-native listeners, in certain situations, may predominately use top-down instead of bottom-up information to perceive non-native speech (Koster, 1987; Mueller, 1980; Tsui and Fullilove, 1998).

While these debates were derived from the perception of language materials in optimal listening situations, in recent years, an increasing amount of research has been carried out to identify at which level the non-native deficit is manifested in adverse listening conditions (Bradlow and Alexander, 2007; Cutler et al., 2004; Mayo et al., 1997). Among various suboptimal listening conditions, speech perception in noise has been widely examined in both native and non-native listeners (see Lecumberri et al., 2010, Mattys et al., 2012, Wang and Xu, 2021 for a review). Mayo et al. (1997) found that monolingual listeners and early bilingual listeners were better at utilizing contextual information than late bilingual listeners in recognizing sentences. This finding suggested that the magnitude of non-native deficit was manifested in the ability to utilize the high-level top-down information. The findings in Cutler et al. (2004) showed that non-native deficits were manifested at all levels of processing, not just at the phoneme level. Bradlow and Alexander (2007) examined the combined effects of semantic and acoustic enhancement in sentence-in-noise recognition by native and non-native English listeners. They found that native listeners required enhancement in only one source (improved contextual information or clearer acoustic cues) to show improved recognition performance. However, non-native listeners required enhancement in both aspects to achieve comparable recognition improvement as native listeners. The authors concluded that the difference between the native and non-native listeners resided in the "signal clarity required for contextual information to be effective, rather than in an inability of non-native listeners to take advantage of this contextual information per se." (p. 2339).

Different from noise-induced masking effect, vocoded speech presents a simplified acoustic profile in which the speech signal can be degraded in temporal and/or spectral domains. Theoretically, recognizing vocoded speech may involve a perceptual mechanism different from listening in noise. As the former involves how to recover and reconstruct speech information based on simplified input signals, the latter involves how to segregate target speech signals from background noise. Technically, the amount of spectro-temporal information can be controlled by manipulating the number of frequency channels and the low-pass cutoffs (LPFs) of the temporal envelope extractor. By virtue of this nature, vocoded speech has been commonly used to examine how varying amount of low-level acoustic information influences speech intelligibility and how top-down processing interacts with the bottom-up mechanism in speech recognition (Robert et al., 2011; Sohoglu et al., 2014; Xu et al., 2005; Xu and Zheng, 2007). Although normal listeners, either native or non-native, do not listen to vocoded speech in their everyday life, this specialization of speech signals provides a valuable way to examine how listeners organize degraded acoustic information and utilize compensatory mechanisms in speech processing. Practically, vocoded speech provides a simulation of speech signals received by cochlear implant (CI) users. While most current research on CIs focuses on the auditory perceptual ability in CI users' native language (Chen and Wong, 2017; Peterson et al., 2010; Sparreboom et al., 2010), how CI users respond to non-native speech materials remains largely unexplored. Hence, testing the recognition of vocoded speech by non-native listeners not only adds to our existing knowledge about speech processing in normal hearing population, but also casts new light on non-native speech perception in CI recipients.

Compared to unprocessed speech signals, noise-vocoded speech contains coarse spectral and temporal information for both vowels and consonants (Xu and Pfingst, 2008). Yet, researchers reported that this type of degraded speech signals could still yield high intelligibility (Dorman et al., 1997; Friesen et al., 2001; Hill et al., 1968; Shannon et al., 1995; Xu et al., 2005, 2021). An excellent recognition score can be reached with 4 to 16 frequency bands for native listeners, depending on the difficulty of the tasks (Dorman et al., 1997; Kim et al., 2015; Loizou et al., 1999; Shannon et al., 1995, 2004; Xu et al., 2005, 2021; Xu and Zheng, 2007). However, there is scant data on how much acoustic detail non-native listeners need to recover the speech information and achieve reasonably good performance in perceiving the spectrally-degraded speech signals.

Additionally, because vocoded speech is not a commonly encountered speech form, the high intelligibility with a small number of spectral channels, especially for phoneme recognition, requires extensive perceptual learning. Davis et al. (2005) conducted a series of experiments to examine what listeners learned in perceptual adaption of noise-vocoded signals and whether the perceptual improvement was driven by the prelexical level mechanism (i.e., initial stage of speech processing with acoustic signal) or lexical level top-down mechanism. The authors found that perceptual adaption involved a general property of the processed signals instead of the acoustic-auditory form of individual vocoded words. The high-level lexical information and top-down processes were more important in perceptual improvement of noise-vocoded speech. With regard to the form in which high-level semantic cues benefit the processing of degraded speech, some researchers proposed that the high-level lexical information helps in predicting the physical form of speech stream and modulating early acoustic processing (Sohoglu et al., 2014). Other researchers argued that semantic cues beyond the segmental level were involved in predicting and activating the forthcoming words or the distorted input (Corp and Rabagliati, 2020; Signoret et al., 2018).

Unlike native listeners who have well-established phonetic categories and possess intact linguistic knowledge in the perceived language, non-native listeners demonstrate phonetic drift due to the influence of established phonetic system in their native language (Best, 1995; Escudero, 2009; Tobin et al., 2017) and inferior linguistic knowledge in the new language (Clahsen and Felser, 2006). Therefore, it is of interest to us that when noise-vocoded speech signals with varying amounts of acoustic information are presented to non-native listeners, what kind of acoustic-phonetic feature that non-native listeners extract and how well they incorporate the low-level auditory information and high-level contextual information in processing the distorted non-native speech signals. To address these questions, the present study used vocoded speech including both phonemes and sentences as recognition stimuli. Phoneme recognition, with no interference from the lexical or sentence-level cues, was used to address the question of how non-native listeners effectively extract low-level acoustic information in response to spectral degradation of the speech signals. Previous studies suggested that consonants and vowels play different roles in sentence recognition, talker identification, and lexical access (Fogerty et al., 2012; New et al., 2008; Owren and Cardillo, 2006). We wondered for non-native speakers, how spectrally-degraded consonants and vowels would contribute to sentence recognition. The sentence recognition task was used to address the questions of whether and to what extent non-native listeners benefit from linguistic knowledge and context cues in recognizing non-native speech when the spectral information is degraded. In the present study, we adopted various types of sentence tests in which the HINT and CUNY sentences were used to assess the sentence recognition performance and the R-SPIN sentences were used to evaluate the role of high-level contextual information.

To date, very few studies examined the recognition of vocoded speech by non-native listeners (Mack et al., 1990; Nittrouer and Lowenstein, 2010; Padilla and Shannon, 2000, 2002). Mack and colleagues (1990) examined the recognition of natural and LPC-vocoded speech by native English listeners and native German speakers who learned English as a second language. The results demonstrated that the non-native listeners performed worse in both word and sentence recognition with vocoded speech. Meanwhile, the recognition score of semantically

anomalous sentences was lower than that of meaningful sentences and the natural vs. vocoded difference was larger in the semantically anomalous sentences than in the meaningful sentences. That is, the degraded speech exerted a more negative influence on anomalous sentences than on meaningful sentences. Padilla and Shannon (2000, 2002) tested the recognition accuracy of vocoded speech in English in quiet and noise conditions by Spanish-English bilinguals. The authors found that bilinguals performed much worse than English monolinguals and the deficits of non-native speech perception were worsened with the added noise. Among different types of test stimuli, the bilingual listeners showed greater difficulties in recognizing vowels, words, and sentences with spectral degradation but less so in recognizing consonants. Nittrouer and Lowenstein (2010) tested the recognition of sine-wave replicas (Remez et al., 1981) and four-channel noise-vocoded sentences in English by native English-speaking listeners and native Mandarin-speaking listeners. The results showed that the deprivation of basic acoustic information had a greater adverse impact on the native Mandarin listeners than on the native English listeners. In the present study, we extended from those studies and designed various vocoded conditions to examine how non-native listeners utilize the low-level acoustic cues and high-level context information to recognize speech with varying amount of spectral information.

2. Methods

2.1. Participants

The participants included 27 native English spakers (3 males and 24 females) and 43 native Mandarin speakers (20 males and 23 females) who learned English as a second language. The native English speakers (L1 listeners) were all college students aged between 20 and 32 years old (M = 23.4 yrs, SD = 2.6 yrs). The native Mandarin speakers (L2 listeners) aged between 21 and 52 years old and all spoke Mandarin in their daily life. Most native Mandarin listeners came from northern dialect regions of China but seven of them came from other dialect regions. The seven listeners all learned Mandarin since they enrolled in school before six years of age. The native Mandarin listeners varied in the chronological age (M = 30.4 yrs, SD = 8.6 yrs), age of English learning (M = 10.3 yrs, SD = 3.2 yrs), years of residence in the U.S. (M = 6.7 yrs, SD = 5.7 yrs), and the amount of daily L2 usage. Note that each listener was requested to estimate the amount of English usage on a 3-level percentage scale (30%, 50%, and 70%) that represented low, medium, and high percentage of daily L2 usage. Among the 43 L2 listeners, 22 reported low level of L2 usage, 12 reported medium level of L2 usage, and 9 reported high level of L2 usage. All participants were recruited from the Midwest region of the U.S. through word-of-mouth, and flyers around the campuses. While all L1 listeners, and some L2 listeners were college students studying for their Bachelor's or Master's degrees, many L2 listeners had earned their Master's or Doctoral degrees. None of the participants reported having any cognitive impairments, hearing problems or speech-language problems in their native language. The participants received monetary compensations for their time and effort in the study. The approval of Institutional Review Board and informed consent from all participants have been obtained.

2.2. Test materials

The phoneme recognition task consisted of two tests: consonant recognition and vowel recognition. For the consonant test, the stimuli included 20 English consonant phonemes /p, b, t, d, k, g, f, v, s, z, ʃ, ð, ʧ, ʤ, m, n, l, r, w, j/ embedded in a /Ca/ context produced by one male speaker and one female speaker selected from the Shannon et al. (1999) stimulus set. For the vowel test, the stimuli included 12 English vowel phonemes /i, ɪ, e, ɛ, æ, u, ʊ, o, ʌ, ɔ, ɑ, ɝ/ embedded in a /hVd/ context produced by two male speakers and two female speakers selected from the Hillenbrand et al. (1995) stimulus set.

The sentence recognition used Hearing in Noise Test (HINT) (Nilsson et al., 1994), City University of New York (CUNY) (Boothroyd et al., 1985), and Revised Speech Perception in Noise (R-SPIN) (Bilger et al., 1984) sentence tests. The HINT sentences are composed of 25 phonetically balanced sentence lists, each containing 10 sentences varying from 4 to 7 words in length for individual sentences. The CUNY sentence materials include 40 lists, each containing 12 sentences. The length of individual sentences in each list varies from 3 to 14 words. Since little data has been reported in the direct comparison of different sentence-recognition tests in non-native listeners, the present study adopted both HINT and CUNY sentences to cross-check the perceptual performance and to avoid the potential bias caused by the selection of sentence stimuli. R-SPIN sentences were used to examine the potential benefit of context information in recognizing vocoded speech in non-native listeners. The R-SPIN sentences include 8 lists, each containing 50 sentences. Each sentence contains 5 to 7 words in length. In each list, 25 sentences provide a high-predictability environment (HP) with rich semantic context information and 25 sentences provide a low-predictability environment (LP) with limited semantic context information for the final target word of each sentence. For each sentence list, the recognition accuracy is calculated for LP and HP sentences, respectively. In the present study, each participant was assigned to six randomly selected lists for each type of sentences.

2.3. Vocoder processing

The speech signals were processed through a custom MATLAB program(Xu et al., 2005). The signals were first band-pass filtered into a number of frequency bands using sixth-order Butterworth filters. The number of the frequency bands was set at 2, 4, 6, 8, 12 with the overall bandwidth between 150 and 5500 Hz [refer to Xu et al. (2005) for the corner frequency for each band in each condition]. Then, the temporal envelope of each frequency band was extracted using half-wave rectification and low-pass filters (second-order Butterworth) with a 160-Hz cutoff frequency. The extracted envelope from each band was used to modulate a white noise that was filtered using the same number of analysis bands for the original speech signals. Finally, these modulated noise bands were combined and presented to the participants for identification.

2.4. Procedures

Prior to the perceptual experiment, each participant was requested to complete a questionnaire that was used to collect the participant's demographic information and information regarding their language experience and language use. During the perceptual test that was conducted in a sound booth, the stimuli were delivered to the participant diotically through Sennheiser HD280 Professional headphones at their most comfortable level. A custom MATLAB program was developed to implement the perceptual tests.

For phoneme recognition, each listener was required to complete a training session for consonants and vowels, respectively, prior to the real test sessions. The training session lasted approximately 30 min with feedback provided. The stimuli used in the training sessions included both processed and unprocessed signals produced by different speakers. For each participant, the training included three subsessions: Preview, Play-It, and Practice. Twenty alphabetically-represented /Ca/ syllables or 12 alphabetically-represented /hVd/ words were presented in a grid on a computer screen situated in front of the listener. In Preview session, the unprocessed stimuli were played with the corresponding alphabetically-represented forms highlighted on the screen; In Play-It session, the participant clicked one /Ca/ or /hVd/ syllable shown in the grid. Then, the corresponding processed signals in all five conditions and unprocessed signals were played back. In Practice session, the participant listened to randomly organized processed and unprocessed signals and responded by clicking the button that represented the token

they heard. Feedback was provided during the practice session. During the real test session, the unprocessed signals and five vocoder conditions were presented in a random order and no feedback was provided. Each stimulus was played once and the listener responded by choosing the /Ca/ or /hVd/ syllables shown in the grid. A total number of 240 (20 syllables × 2 talkers × 6 conditions) consonant stimuli and 288 (12 syllables × 4 talkers × 6 conditions) vowel stimuli were presented to each participant.

For sentence recognition, a practice session with feedback provided was completed prior to the actual testing. The practice session included 25 sentences composed of five in each processed condition. In the real testing, participants were tested with HINT sentences first, followed by CUNY, and then R-SPIN sentences. The order of unprocessed and vocoder-processed conditions was randomized. Within each test, the participants were allowed to play each sentence up to 3 times. The participants were required to type in what they had heard in a text box on the computer screen. For HINT and CUNY sentences, the participants were asked to type down the whole sentence heard. For R-SPIN sentences, the participants were asked to type down the final word of each sentence. No feedback was provided. The accuracies for HINT and CUNY sentence tests were calculated based on the correctly recognized number of (key) words out of the total number of (key) words. For R-SPIN sentences, the accuracy was calculated based on the correctly recognized number of final target words out of the total number of 25 words for the LP and HP sentences respectively. For the L2 listeners, obvious spelling errors and predictable errors due to native language interference (e.g., subject-verb agreement, past tense) were accepted and not counted as wrong. The total number of sentences were 60 (10 × 6) for HINT, 72 (12 × 6) for CUNY, and 300 (50 × 6) for R-SPIN. Break time was provided to each participant at least twice during the experiment. Additional break time was allowed upon request. The total amount of time used for the whole experiment was approximately 3 h for each participant.

3. Results

3.1. Phoneme recognition accuracy

Fig. 1 presents the group mean performance for L1 and L2 listeners for consonant and vowel recognition. In consonant recognition (Fig. 1, left panel), when the spectral resolution was low (2 channels), both L1 and L2 listeners showed poor recognition performance. The average recognition score was only 19% correct for both groups. As the spectral resolution increased from 2 to 12, both groups showed improved recognition accuracy. However, the magnitude of improvement was greater in the L1 listeners than in the L2 listeners. Similar to the results of consonant recognition, both groups of listeners showed extremely low recognition accuracy in vowel recognition (approximately 13% correct for the L1 listeners and 11% correct for the L2 listeners) at 2 channels (Fig. 1, right panel). Both L1 and L2 listeners demonstrated improved vowel recognition performance as a function of increased number of channels. However, the L1 listeners showed a greater amount of improvement as the number of spectral channels increased than the L2 listeners. For example, the L1 listeners improved from 13% to 33% correct whereas the L2 listeners improved from 11% to 19% correct from 2 to 4 channels. Unlike the consonant recognition for which neither group showed much improvement from the 12-channel condition to the unprocessed condition, the vowel recognition performance showed further improvement from the 12-channel condition to the unprocessed condition. In addition, the performance difference between the L1 and L2 listeners became larger as the number of spectral channels increased in vowel recognition than in consonant recognition.

A Generalized Linear Mixed-effects Model (GLMM) was applied to consonant and vowel recognition data, respectively. The recognition accuracy data were fitted with a binomial probability distribution with a logit link function. Specifically, the factors of participants' language background, channel condition, interaction between these two factors, and participants' age were defined as the fixed effects whereas the factor of subject was defined as a random effect. The model was constructed with a random intercept for subjects added first and then a random slope for each main effect on subjects was added. The results showed that inclusion of random slopes for the main effects on subjects did not improve the model for consonant or vowel tests. The best fit model was the one with a random intercept for subjects included. Listeners' recognition performance for both consonants and vowels were significantly affected by their language background (consonants: $F(1, 407) = 41.9$, $p < 0.0001$; vowels: $F(1, 401) = 77.7$, $p < 0.0001$), channel condition (consonants: $F(5, 407) = 447.8$, $p < 0.0001$; vowels: $F(5, 401) = 526.2$, $p < 0.0001$), language by channel interactions (consonants: $F(5, 407) = 11.9$, $p < 0.0001$; vowels: $F(5, 401) = 13.9$, $p < 0.0001$), and age (consonant: $F(1, 407) = 6.4$, $p = 0.012$; vowel: $F(1, 401) = 6.5$, $p = 0.011$). The fixed coefficients revealed that the recognition outcomes in 2-, 4-, 6-, and 8-channel conditions were significantly different from the

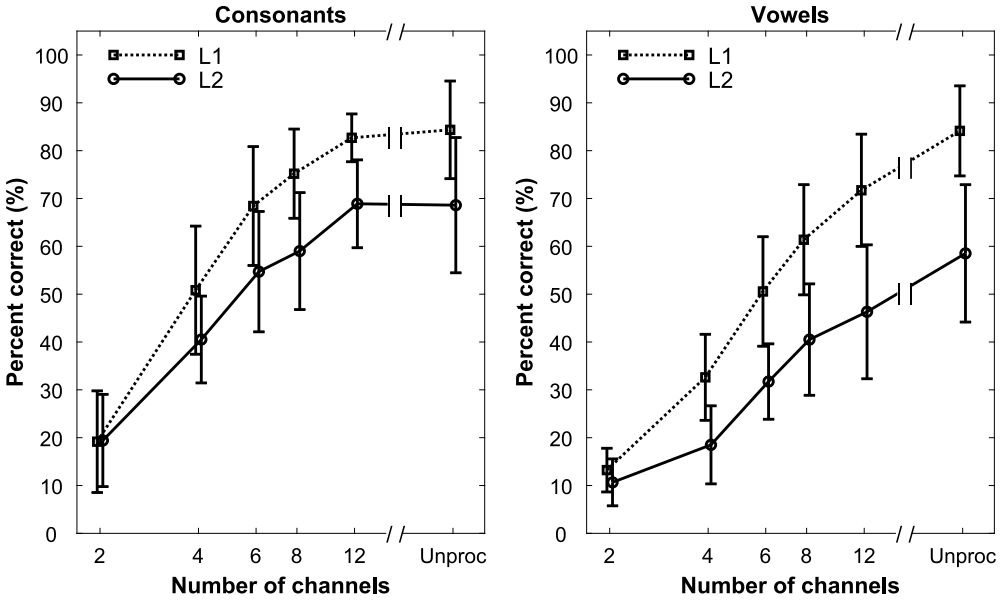

Fig. 1. The group mean percentage correct with standard deviations for L1 and L2 listeners for consonant (left) and vowel (right) recognition in 2-, 4-, 6-, 8-, 12-channel, and unprocessed (labeled as Unproc) conditions.

unprocessed condition (reference level) for consonant recognition (all $p < 0.05$). For vowel recognition, the recognition outcomes in 2-, 4-, 6-, 8-, and 12-channel conditions were all significantly different from the unprocessed condition (all $p < 0.05$).

3.2. Phoneme confusion matrix

Confusion matrices were constructed for the perceptual data from the L1 and L2 listeners, respectively, for each channel condition of the vocoded signals and the unprocessed signals. For consonant recognition shown in Fig. 2, when the spectral resolution was as low as 2 frequency bands, the L1 listeners could only roughly differentiate different manners for the obstruents but showed substantial confusion among different places within each manner. In the meantime, they showed very poor recognition performance for the sonorants. When the spectral resolution increased to 4 channels, the L1 listeners demonstrated improved recognition accuracy for individual consonant phonemes but still showed evident confusion among sonorants. As the number of channels increased to 6 or above, the L1 listeners showed further improvement in recognition of sonorants and demonstrated high accuracy rates for all English consonants.

As for the L2 listeners, they roughly recognized the manners for the obstruents but showed very poor recognition for the sonorants and different places within each manner in the 2-channel condition. These findings were similar to those of the L1 listeners. As the spectral resolution increased to 4 channels, the L2 listeners showed further improvement in differentiating the manners but they still showed much confusion among different places in comparison to the L1 listeners. However, it is noteworthy that the L2 listeners could separate the voicing feature relatively well. For example, they perceived /b/ as /b/ or /d/ but not as /p/ or /t/. They perceived /k/ as /p/, /t/, or /k/ but not as /b/, /d/, or /g/. However, the L2 listeners still showed low recognition accuracies for the sonorants and tended to perceive the sonorants as /v/, /m/, or /w/. When the number of channels increased to 6 or above, the L2 listeners demonstrated a similar confusion pattern as the L1 listeners although with a lower accuracy rate.

For vowel recognition shown in Fig. 3, both groups of listeners performed extremely poor in the 2-channel condition and did not show observable patterns of confusion. In the 4-channel condition, the L1 listeners roughly separated English vowels into three large clusters: high-front vowels /i, ɪ, e/, low-front vowels /ɛ, æ/, and high-back vowels /u, ʊ, o/, although they still showed much confusion within each cluster. Further, the L1 listeners did not differentiate the low-back vowels /ʌ, ɔ, ɑ/ and tended to misperceive them as the low-front vowel /æ/. In the 6-channel condition, the L1 listeners showed improved recognition accuracy especially for the low-back vowels /ʌ, ɔ, ɑ/ although some listeners still misperceived them as the low-front vowel /æ/. When the number of channels further increased to 8 and 12, the L1 listeners showed significantly improved recognition accuracy for all vowels. However, certain confusion still existed for /ɔ/ and /ɑ/ even in the unprocessed condition, which was probably affected by the low-back vowel merger occurring in many dialect regions of American English (Clopper et al., 2005; Jacewicz et al., 2011; Labov et al., 2006).

Different from the L1 listeners, the L2 listeners only roughly differentiated the two high front vowels /i, ɪ/ and two high back vowels /u, ʊ/ from the other vowels in the 4-channel condition. In the 6-channel condition, the L2 listeners showed improved recognition accuracy in that they separated the low-front vowels /ɛ, æ/ and low-back vowels /ʌ, ɔ, ɑ/ from other vowel groups even though they still demonstrated much confusion within each group. In the 12-channel and unprocessed conditions, the L2 listeners showed a clearer pattern of perceiving English vowels into four large subgroups: high-front vowels /i, ɪ, e/, low-front vowels /ɛ, æ/, high-back vowels /u, ʊ, o/, and low-back vowels /ʌ, ɔ, ɑ/. However, they still exhibited remarkable confusion within each subgroup.

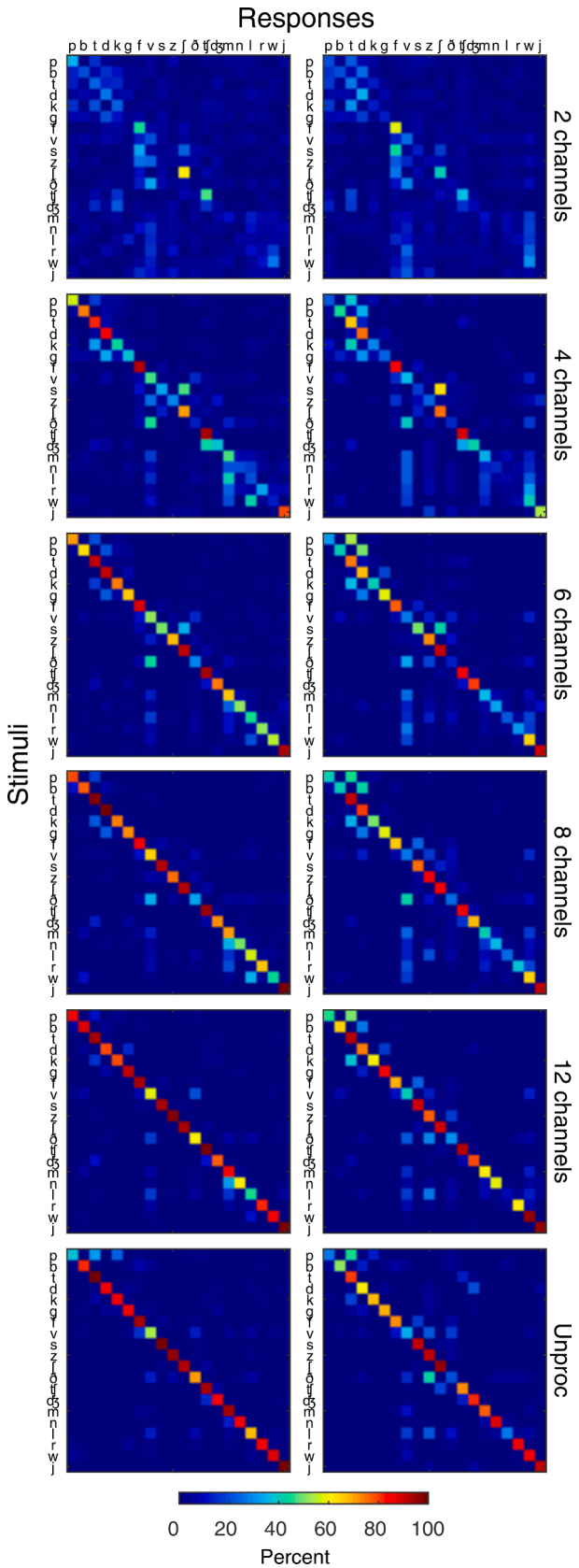

Fig. 2. Confusion matrices for consonant recognition in 2-, 4-, 6-, 8-, 12-channel and unprocessed (labeled as Unproc) conditions for the L1 (left) and L2 (right) listeners. The stimuli are represented by the ordinate and the group-pooled responses are represented by the abscissa. In each small square, the grayscale or color represents the percent of the stimulus-response pair.

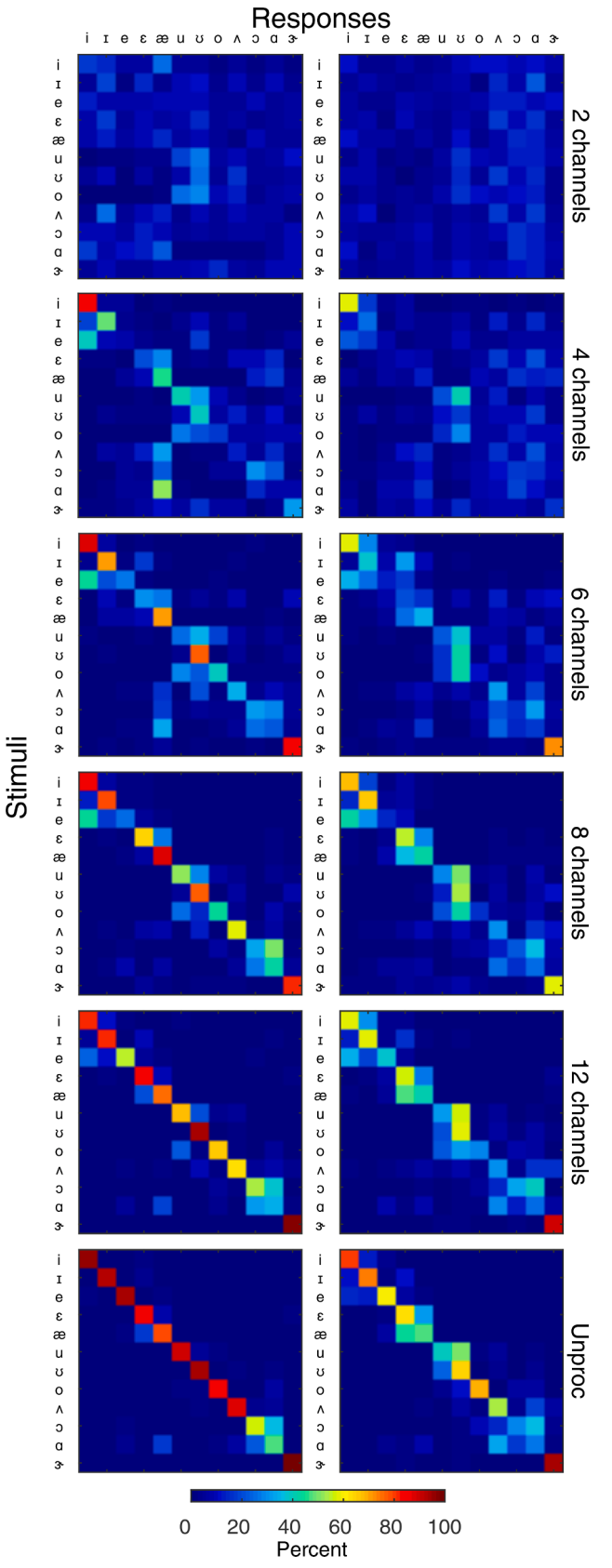

Fig. 3. Confusion matrices for vowel recognition in 2-, 4-, 6-, 8-, 12-channel and unprocessed (labeled as Unproc) conditions for the L1 (left) and L2 (right) listeners. Conventions as in Fig. 2.

3.3. Information transmitted for phonetic features

The confusion matrix data were analyzed for information transmission of phonetic-acoustic features. The features used for consonant analysis included voicing, frication, nasality, place, sibilant, sonorance, and manner as described in Hornsby and Ricketts (2001) and Xu et al. (2005). Each consonant was assigned a binary code for the features of voicing, frication, nasality, sibilant, and sonorance. For the other two features of place and manner, the coding for each feature had five levels. Specifically, the five categories for the place feature included labial, dental, alveolar, palatal, and velar. The five categories for the manner feature included plosive, fricative, affricate, nasal, and approximant. The vowel features included duration, F1, and F2. The coding for vowel duration had two levels: short and long. The F1 and F2 coding included three levels: low, mid, and high (the coding system for consonant and vowel features as well as the cutoffs for categorizing vowel features were the same as described in Xu et al., 2005). For each listening condition in each listener group, the overall confusion matrix collapsed for all participants was used for feature information transmission analysis following the procedure described in Miller and Nicely (1955). The outcome values indexed the amount of information transmitted for each feature in each condition.

The amount of information transmitted for each feature of the consonants and vowels in each condition is shown in Figs. 4 and 5, respectively. Both L1 and L2 listeners demonstrated increased amounts of information transmitted for all tested features from 2 to 12 channels, for both consonant and vowel recognition. For consonant recognition, the L2 listeners showed an overall lower rate for almost all features in all conditions than the L1 listeners. It is noteworthy that the L2 listeners showed compatible patterns of the seven features in each condition as the L1 listeners. Both groups of listeners showed relatively high proportions of information transmitted for voicing, frication, sibilant, and manner. For vowel recognition, the two groups showed distinct patterns on the three features as the number of channels increased. The L1 listeners showed a larger proportion of duration in the 2-channel condition and the amount of duration information showed a continuous increase as the frequency channels increased. However, the L2 listeners showed little information transmitted for duration in all conditions including the unprocessed condition. On the other hand, while the L1 listeners showed relatively equal proportions on F1 and F2 in 6-, 8-, and 12-channel conditions and a larger amount of information on F1 in the unprocessed condition, the L2 listeners consistently showed a greater amount of information on F2 than on F1.

3.4. Sentence recognition

Fig. 6 shows the group mean recognition accuracy for HINT and CUNY sentences as a function of spectral resolution in both L1 and L2 listeners. Both groups of listeners showed improved recognition accuracy as the number of spectral channels increased. However, compared to the L1 listeners, the L2 listeners demonstrated lower recognition accuracies in all conditions. In addition, the two groups of listeners showed different patterns in the recognition improvement as a function of the number of channels. The L1 listeners demonstrated the greatest increase ($>$ 50 percentage points) from the 2- to 4-channel condition but showed no evident increase when the number of channels was $>$ 6. The L2 listeners demonstrated comparable magnitude of improvement from 2- to 4- and 4- to 6-channel conditions. Further, the L2 listeners showed gradual but consistent improvement from 6 channels to the unprocessed condition for both types of sentences. Note that the L1 listeners showed a higher accuracy for CUNY sentences than for HINT sentences when there were four spectral channels. But on other conditions, the L1 listeners showed similar accuracies between the two types of sentences. By contrast, the L2 listeners consistently showed measurably higher recognition accuracies for CUNY sentences than for HINT sentences.

A GLMM fitting was applied to the recognition performance for HINT

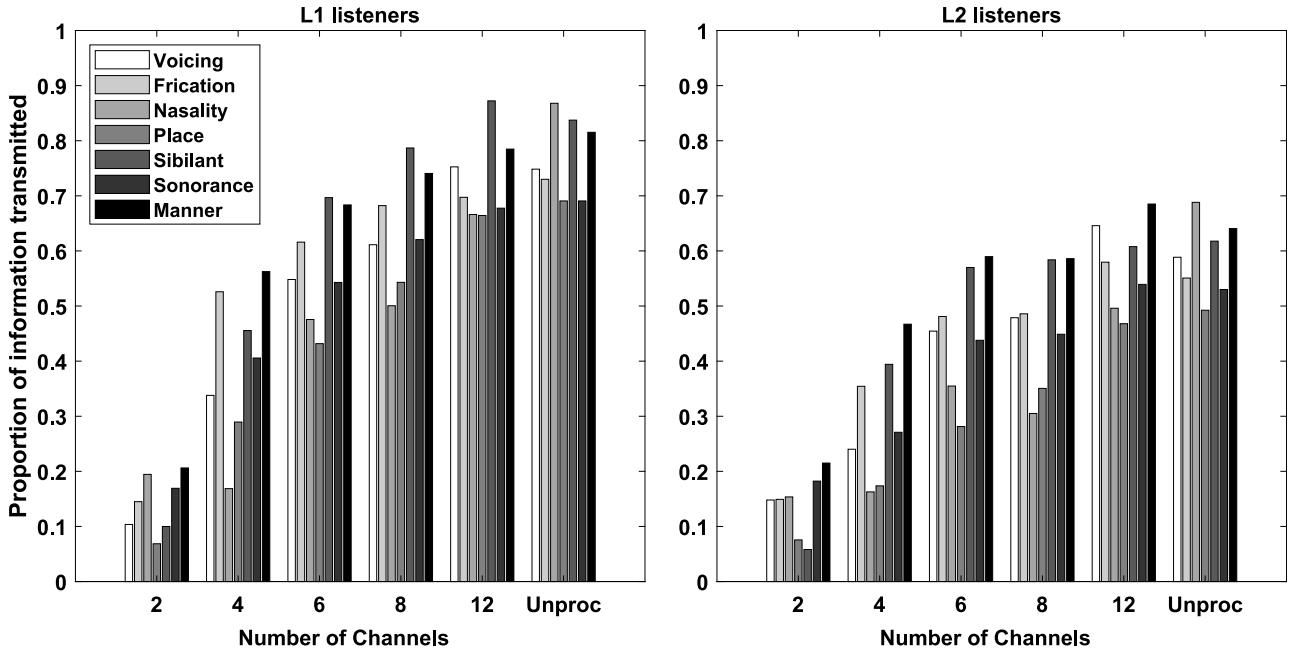

Fig. 4. The amount of information transmitted for the tested phonetic features of consonant recognition in 2-, 4-, 6-, 8-, 12-channel and unprocessed (labeled as Unproc) conditions for the L1 and L2 listeners.

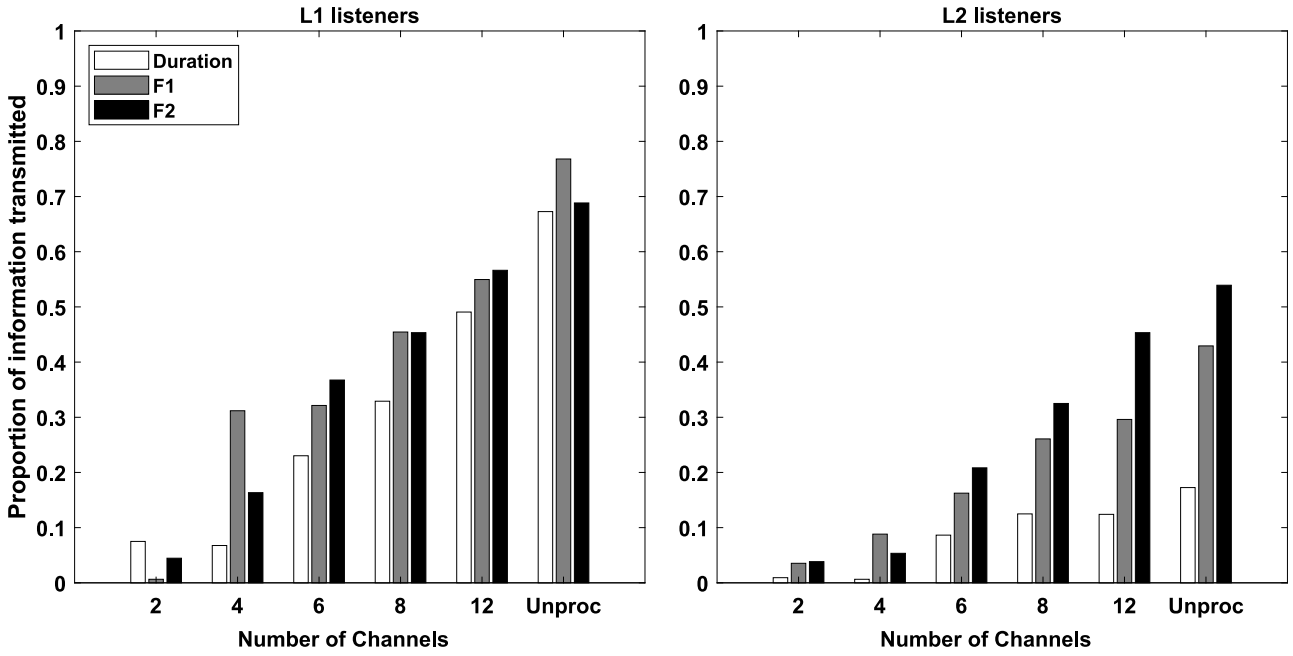

Fig. 5. The amount of information transmitted for the tested phonetic features of vowel recognition in 2-, 4-, 6-, 8-, 12-channel and unprocessed (labeled as Unproc) conditions for the L1 and L2 listeners.

and CUNY sentences, with a binomial probability distribution and a logit link function selected for the model. In particular, the factors of sentence type (HINT or CUNY), number of channels, and language background, interactions between these factors, and participant's age were defined as the fixed effects and the subject effect was defined as a random effect. Different models were tested and the best fit model was the one with a random intercept for subjects and random slopes for the effect of sentence type on subjects and the effect of number of channels on subjects. The results showed significant effects for language background: $F(1, 815) = 151.9$, $p < 0.0001$), number of channels ($F(5, 815) = 443.8$, $p < 0.0001$), and age ($F(1, 815) = 18.4$, $p < 0.0001$). The recognition outcomes of 2-, 4-, 6-, 8-, and 12-channel conditions were significantly different from the reference level of unprocessed condition (all $p < 0.05$). In addition, there was a significant interaction effect for sentence by channel ($F(5, 815) = 45.5$, $p < 0.0001$), language by channel ($F(5, 815) = 6.5$, $p < 0.0001$), language by sentence ($F(1, 815) = 4.2$, $p = 0.04$), and sentence by language by channel ($F(5, 815) = 12.6$, $p < 0.0001$). The significant language by sentence interaction suggested that the L1 and L2 listeners performed differently on these two types of sentences.

Fig. 7 presents the recognition performance of R-SPIN sentences as a function of spectral resolution for the L1 and L2 listeners. As the number

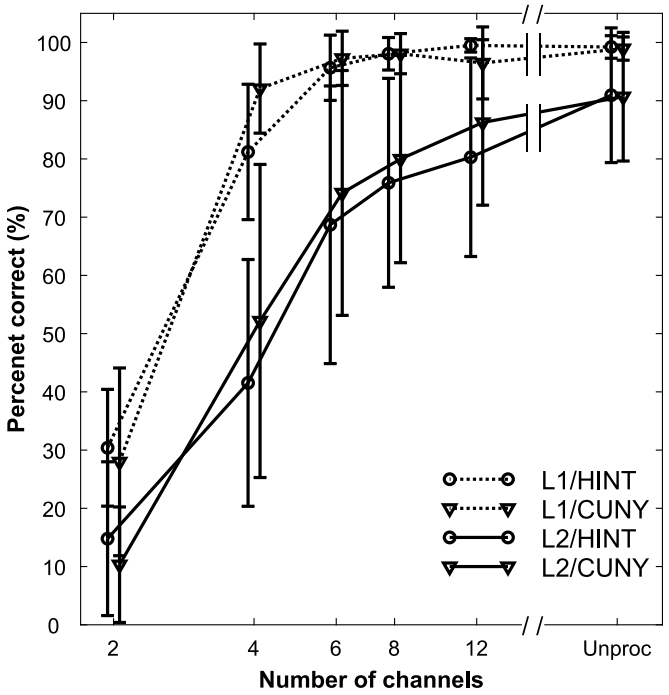

Fig. 6. Recognition performance (group mean and standard deviation) of HINT and CUNY sentences in L1 and L2 listeners in 2-, 4-, 6-, 8-, 12-channel conditions and unprocessed (labeled as Unproc) condition.

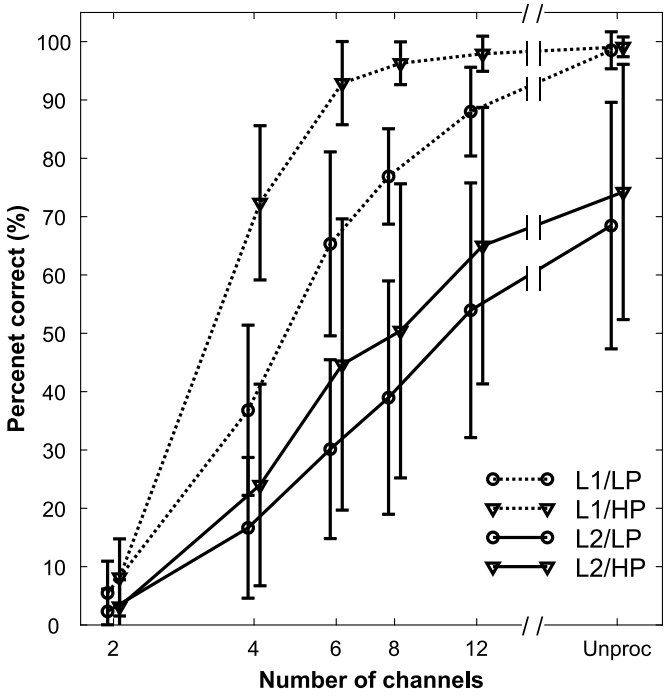

Fig. 7. Recognition performance (group mean and standard deviation) of R-SPIN sentences in L1 and L2 listeners in 2-, 4-, 6-, 8-, 12-channel conditions and unprocessed (labeled as Unproc) condition.

of channels increased, both groups of listeners showed increased recognition accuracies and greater improvement in HP sentences than in LP sentences. The L1 listeners outperformed the L2 listeners in both LP and HP sentences. Even with the unprocessed signals, the recognition accuracy in the L2 listeners only approximated the performance of HP sentences in L1 listeners with just four channels of spectral information. Among the six conditions, the L1 listeners demonstrated a floor effect in the 2-channel condition and a ceiling effect in the unprocessed condition. They showed little performance difference between LP and HP sentences in these two conditions. The average performance difference between LP and HP sentences in the L1 listeners reached approximately 35 percentage points in the 4-channel condition and then showed a constant decrease as the number of channels increased. The reduced performance difference as a function of increasing number of channels was because on one hand, the low-level acoustic-phonetic information became more available as shown in the increased accuracy of LP recognition; and on the other hand, the L1 listeners demonstrated a ceiling effect of HP sentences when the number of channels was ≥ 6. In contrast to the L1 listeners, the L2 listeners showed steady improvement for the target word recognition in both LP and HP sentences and the performance difference between LP and HP sentences was relatively stable as the number of channels increased from 2 to 12.

A GLMM was used to examine the recognition performance of R-SPIN sentences, with a binomial probability distribution and a logit link function selected for the model. In particular, the factors of sentence type (LP or HP), number of channels, language background, interactions involving the three factors, and participants' age were defined as the fixed effects. The subject effect was defined as a random effect. The best fit model was the one with a random intercept for subjects and random slopes for the effect of sentence type on subjects and the effect of channel condition on subjects. The results showed significant effects for all tested factors (language background: $F(1, 815) = 200.5, p < 0.0001$), sentence type ($F(1, 815) = 174.1, p < 0.0001$), number of channels ($F(5, 815) = 470.0, p < 0.0001$), and age ($F(1, 815) = 27.0, p < 0.0001$). The recognition outcomes of 2-, 4-, 6-, 8-, and 12-channel conditions were significantly different from the reference level of unprocessed condition (all $p < 0.0001$). In addition, there was a significant interaction effect for language by sentence ($F(1, 815) = 41.8, p < 0.0001$), sentence by channel ($F(5, 815) = 7.5, p < 0.0001$), language by channel ($F(5, 815) = 13.4, p < 0.0001$), and sentence by language by channel ($F(5, 815) = 2.9, p = 0.013$).

To examine whether the contextual benefit was different between the two listener groups in different conditions, the HP - LP performance difference was fitted with a Linear Mixed-effects Model (LMM). The language background, number of channels, interaction between these two factors, and participants' age were defined as fixed effects and the subject effect was set as a random effect. The results revealed a significant effect for language ($F(1, 386) = 57.3, p < 0.0001$), number of channels ($F(5, 386) = 35.5, p < 0.0001$), language by channel interaction ($F(5, 386) = 17.5, p < 0.0001$), and age ($F(22, 386) = 2.1, p = 0.003$). The significant language effect and language by channel interaction suggested that the L1 listeners received greater contextual benefit than the L2 listeners and the group difference on contextual benefit was mainly reflected on certain channels.

3.5. Correlation and regression analysis

A correlation analysis was conducted to examine the relationship between phoneme recognition and sentence recognition in the L2 listeners (shown in Fig. 8). For each listener, an overall accuracy rate was calculated for consonants and vowels, respectively, across the five vocoder conditions. An overall accuracy rate was also calculated for sentence recognition across the five vocoder conditions and all three types of sentences. Both consonants and vowels showed significant positive correlation with sentence recognition in the L2 listeners (consonants: $r = 0.489, p = 0.001$; vowels: $r = 0.622, p < 0.0001$). A linear regression analysis was conducted to examine whether vowel and/or consonant recognition accuracy predicted the sentence recognition performance. The results confirmed that the L2 listeners' sentence recognition were significantly predicted by both vowel and consonant recognition performance ($r^2 = 0.456, F(2, 40)=16.7, p < 0.0001$).

As shown in the phoneme and sentence recognition performance

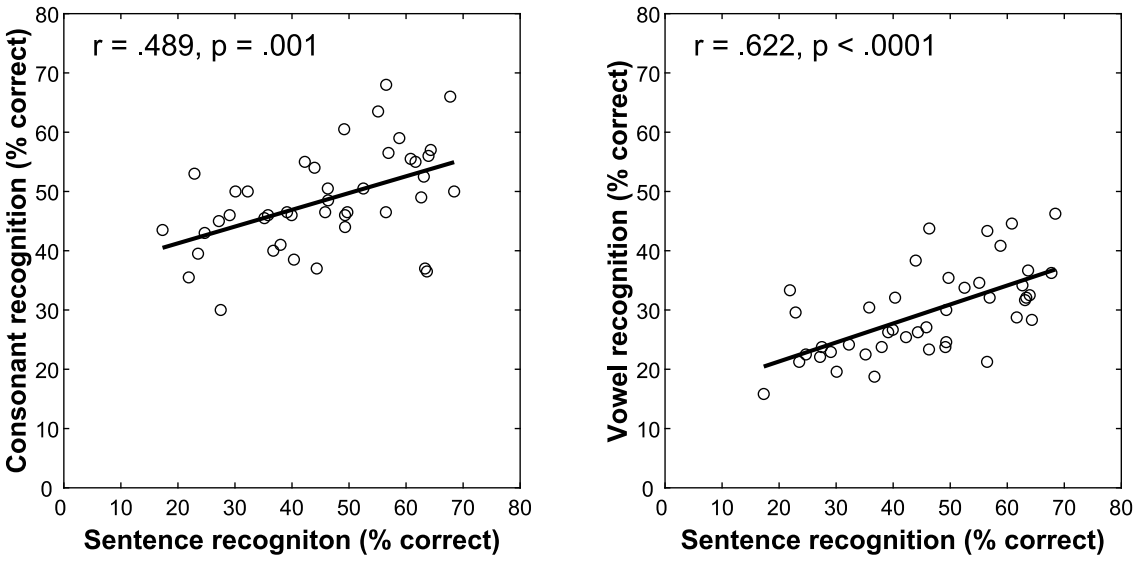

Fig. 8. Correlation between the overall performance of consonants and sentence recognition (left) as well as vowel and sentence recognition (right) in the L2 listeners with vocoded speech signals. For each L2 participant, the percent-correct score for consonant and vowel, respectively, was averaged across the five vocoder conditions. The percent-correct scores for sentence recognition was averaged across three types of sentences in all five vocoder conditions.

(Figs. 1, 6, and 7), the L2 listeners demonstrated greater variability than the L1 listeners. To further examine how the recognition performance was associated with language learning factors, a correlation analysis was conducted between the overall recognition accuracy and each of the language-learning factors including age of L2 learning, length of residence in the U.S., and percentage of daily L2 use (shown in Fig. 9), which was followed by a regression analysis. Because the L2 listeners' chronological age was highly positively correlated to their length of residence in the U.S. ($r = 0.893$, $p < 0.0001$), age factor was not included in the correlation and regression analysis. For each L2 listener, an overall accuracy rate was obtained across consonants, vowels, and three types of sentences in all five vocoder conditions. Note that the amount of L2 usage of each participant was determined based on the participant's self-report as one of three rankings (e.g., 30%, 50%, 70%). Nonparametric analysis was used for the correlation between this factor and the overall recognition accuracy. The results revealed a significant positive relationship between the length of residence in the U.S. and overall recognition performance ($r = 0.484$, $p = 0.001$) as well as between the percentage of daily L2 use and overall recognition performance (Spearman's rho $= 0.464$, $p = 0.002$). A stepwise regression analysis was implemented to examine whether these language learning factors predicted the overall recognition accuracy in the L2 listeners. The results revealed that only the factor of length of residence in the U.S. was a significant predictor ($r^2 = 0.234$, $F(1, 41) = 12.5$, $p = 0.001$). The factor of the amount of daily L2 usage did not account for the variance of the overall recognition accuracy.

4. Discussion

In the present study, we examined the recognition of vocoded English phonemes and sentences by native and non-native English listeners. At issue was how the magnitude of spectral degradation impacts the extraction of low-level acoustic information and the application of high-level contextual information in speech processing in non-native listeners. A group of 43 Mandarin-speaking English learners (L2 listeners) and a control group of 27 native English listeners (L1 listeners) were recruited to recognize English phonemes and sentences that were processed with varying amount of spectral information through vocoder processing. Our results showed that the L2 listeners performed worse than the L1 listeners for both phoneme and sentence recognition in all test conditions except for the 2-channel condition. The L2 listeners were not as effective as the L1 listeners in recovering the segment features as the spectral information increased, which could be largely explained by the phonemic confusions due to the native language impact. Moreover, the L2 listeners received less contextual benefit in sentence recognition as the number of frequency channels increased than the L1 listenes did.

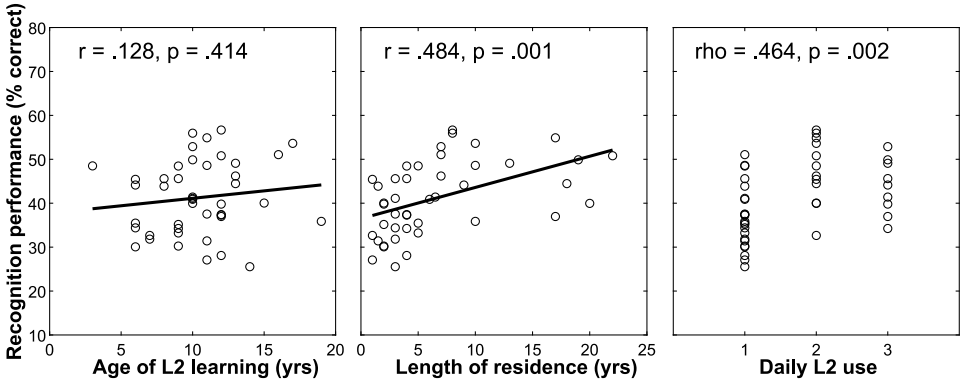

Fig. 9. Correlation between the overall recognition performance and language learning factors including age of learning, length of residence, and the amount of daily-based L2 usage. For each L2 participant, the overall percentage was the average percent accuracy collapsed across consonants, vowels, and three types of sentences in all five vocoded conditions. The amount of L2 usage was coded as 1 for 30%, 2 for 50%, and 3 for 70%.

However, the contextual cues played a consistent facilitating role as the number of frequency bands increased, which suggested that L2 listeners required more spectral information to maximize the contextual benefit relative to L1 listeners.

In terms of phoneme recognition, we noted that the L1 and L2 listeners both showed extremely low accuracy in the 2-channel condition due to the floor effects. As the number of frequency bands increased, the L2 disadvantage exacerbated. This result suggested that L1 listeners could better extract the acoustic information presented in vocoded speech as the number of spectral channels increased in comparison to L2 listeners. In previous studies that examined non-native speech perception in noise, many researchers reported that non-native listeners performed similarly to the native listeners in quiet but suffered more from the presence of various types of background noise (Cutler et al., 2008; Lecumberri and Cooke, 2006; Rogers et al., 2006). Similar to these studies, we found that the L2 listeners showed much lower recognition accuracy for vocoded signals than the L1 listeners. However, different from the finding of greater L2 disadvantage as a function of increased noise levels, the present study showed less L2 disadvantage as a function of increased spectral degradation (decreased number of frequency bands).

One possible explanation could be different perceptual mechanisms involved in speech recognition in these two types of adverse conditions. For speech perception in noise, the target speech signal is presented with a full range of spectral and temporal information. The background noise in forms of stationary noise, modulated noise, multi-talker babble or competing speaker, introduces energetic and/or informational masking effects by "rendering unavailable potential cues to the identity of segments and their boundaries as well as interfering with access to prosodic cues" or introducing meaningful linguistic information that "has the capacity to interfere with decisions at higher levels of processing" (p. 872, Lecumberri et al., 2010). Listeners need to segregate the masker from the target signal. Researchers found that certain acoustic features are more robust in presence of noise (Wright, 2004; Lovitt and Allen, 2006; Hazen and Simpson, 1998). Also, the masking release effect accompanying certain types of noise is manifested differently in native and non-native listeners (Brouwer et al., 2012; Calandruccio et al., 2013; Guan et al., 2015). The perceptual cue weighting in noise may also differ between native and non-native listeners (Cutler et al., 2007; Heinrich et al., 2010). Meanwhile, masking noise only obscures low-amplitude sounds whereas the temporal and spectral fine structure of the speech segments that are of relatively high-amplitude would be maintained. By contrast, for vocoded speech signals, the acoustic profile of speech segments is obscured due to the vocoder processing. The spectral and temporal fine structure is eliminated but the low-level amplitudes within each spectral band are still preserved in the form of temporal envelopes. Listeners need to "recover" the acoustic profile based on the degraded spectral information. When there were only a small number of frequency bands, a limited amount of acoustic features were coarsely presented. It is likely that the more simplified signal would restrict listeners' recovery of the target sounds to a greater extent, regardless of the listeners' language experience. Therefore, even though the L1 listeners generally outperformed the L2 listeners in recognizing vocoded speech in their native language, they did not show as great of an advantage in phoneme recognition in vocoded conditions as they did in unprocessed condition in comparison to the L2 listeners.

The subsequent analysis of the confusion matrix and information transmitted for phonetic features indicated that the reason L2 listeners were less able to accurately identify low-level acoustic information was largely associated with the phoneme confusions caused by the native language effect. According to the cross-language perception theories such as NLM (Kuhl, 1993), PAM (Best, 1995), and L2LP (Escudora, 2009), L2 learners, especially at the initial stage of L2 learning, tend to fully transfer the L1 system into a new language and associate the speech sounds of the new language with existing sounds in their own language system. The native language effect on the recognition performance was clearly shown on the vowel confusion pattern and the information transmission pattern for vowel recognition (Figs. 3 and 5). Mandarin has only five basic vowel phonemes /a, i, u, y, ɤ/, much fewer than the vowel inventory in English. While both Mandarin and English have front-back and high-low distinctions, Mandarin does not have vowels located at different levels of tongue height, especially for the front vowels. In addition, Mandarin lacks tense-lax distinction. Given that Mandarin has a simpler vowel system than English, the L2 listeners demonstrated evident confusion for those vowels that are not phonetically distinctive in their native language. Therefore, even in the 12-channel and unprocessed conditions, the L2 listeners could only roughly separate the high-front, low-front, high-back, and low-back vowel subgroups from each other but still showed considerable confusion within each subgroup. The information transmission analysis provided further evidence regarding the L1-L2 differences and native language effect on L2 listeners' vowel recognition. The L1 listeners showed relatively similar proportions of information on F1 and F2 for the 6-, 8-, and 12-channel conditions and a larger amount of information on F1 in the 4-channel and unprocessed conditions. However, the L2 listeners showed an increasingly larger proportion of F2 as the number of frequency channels increased. This might reflect the lack of diversity in Mandarin vowels on tongue height. Additionally, Mandarin vowels do not show durational differences as the short vs. long distinction for tense-lax vowel contrasts in English. Meanwhile, as Mandarin is a tonal language and the lexical tones carried by vowels vary in the duration (Yang et al., 2017), this might act as a confusing cue for vowel identity. Therefore, the L2 listeners did not show much information transmitted for vowel duration.

As for the consonant recognition, although these two languages have different consonant inventories, they have a similar number of consonant phonemes. In addition, both languages have obstruents that include stops, fricatives, affricates, and sonorants that include nasals, liquids, and glides. The two languages also share common places for many consonants. Given these similarities, the L2 listeners showed higher recognition accuracies and less confusion for consonant recognition relative to vowel recognition. These findings were consistent with Padilla and Shannon (2002) who reported more difficulties in vowel recognition than in consonant recognition of vocoded speech in non-native listeners.

We noticed that the recognition accuracy of phoneme test for the L1 listeners in the unprocessed condition was lower than what we expected, which might be accounted for by several reasons: the stimuli used for consonant recognition were adopted from Shannon et al. (1999) and those for vowel recognition were adopted from Hillenbrand et al. (1995). The listeners of the present study might come from dialect regions different from the speakers who produced the two stimuli sets. Additionally, the stimuli for both consonant and vowel recognition tests were produced by multiple speakers including both males and females. The stimuli were randomized across speakers and across the unprocessed and processed conditions. Moreover, the /Ca/ and /hVd/ syllables were not meaningful words and the written forms were located next to each other on the computer screen. All of these factors might result in procedure errors and random errors when the listeners click the buttons. The last potential reason could be the engagement of the listeners. When analyzing the performance for each participant, we noticed that there were four L1 listeners who had lower than expected performance for the phoneme tests but their sentence recognition results were above the average. This suggested that the four listeners might not be fully engaged in the phoneme recognition tasks.

While vocoded speech provides a limited amount of acoustic information at the segmental level, contextual information plays a vital role in facilitating the recognition of degraded speech for both groups. Kong et al. (2015) compared the recognition of noise-vocoded CUNY and IEEE sentences by normal-hearing native English listeners. The authors found that the listeners showed extremely low recognition accuracy for both types of sentences with 2 spectral channels and remarkable

improvement from 2-channel to 4-channel conditions for both CUNY and IEEE sentences. Meanwhile, the improvement of CUNY sentences, which contained richer contextual cues, was greater than that of IEEE sentences which provided less sentence semantics. Patro and Mendel (2016) reported that when native English listeners were presented with extremely degraded speech signals (periodically interrupted together with 4-channel noise vocoded), the average recognition performance was close to 0% for both LP and HP R-SPIN sentences. As the number of channels increased to 8 or 16, the facilitating role of contextual cues in HP relative to LP sentences became more prominent. Consistent with these findings, our results revealed that when the speech signals were severely degraded with only 2 spectral channels, both L1 and L2 listeners showed extremely low accuracies on both LP and HP sentences. This result suggested that the implementation of top-down processing required access to sufficient low-level acoustic-phonetic input. As the number of channels increased, both L1 and L2 listeners showed improved recognition accuracies and the HP sentences showed a greater improvement than LP sentences in both groups of listeners.

Although both L1 and L2 listeners benefit from the sentence context, the L2 listeners were less able to effectively apply the contextual information as the L1 listeners did. This was reflected as the consistently worse performance for all types of tested sentences and the smaller performance differences between the HP and LP R-SPIN sentences in the L2 listeners than in the L1 listeners. These results were consistent with the findings of significantly worse recognition accuracy of non-native listeners reported in previous studies (Mack et al., 1990; Padilla and Shannon, 2000, 2002). While both L1 and L2 listeners showed very low recognition accuracy in the 2-channel condition for HINT and CUNY sentences, when the number of spectral channels increased to 4, the L1 listeners experienced approximately 50 percentage points increase in recognition accuracy for HINT sentences and 60 percentage points increase in accuracy for CUNY sentences. By contrast, the L2 listeners experienced less than 40 percentage points increase in both types of sentences. The difference in contextual benefit was also evident in the performance difference between the HP and LP sentences in these two groups. The greatest performance difference between the HP and LP sentences was 35 percentage points that occurred in the 4-channel condition in the L1 listeners, while the greatest difference in the L2 listeners was only 12 percentage points that occurred in the 6-channel condition. These differences suggested that the L2 listeners were less able to effectively utilize the sentence context and linguistic knowledge in recognizing degraded speech. One possible explanation was that the L2 listeners might allocate the attentional resources differently from the L1 listeners and attend to different properties of the speech signals in comparison to the L1 listeners (Astheimer et al., 2016; Strange, 2011). In this case, the L2 listeners might assign more attentional and cognitive resources at the acoustic-phonetic cues and be left with limited resources for the application of contextual cues. This might be especially true when the speech signals are degraded and additional cognitive and attentional resources are required in general (Rönnberg et al., 2013; Wild et al., 2012).

When observing the HP and LP sentence performance in these two groups, the L1 listeners showed the greatest contextual benefit in the 4-channel condition and the magnitude of the context benefit became smaller as the number of spectral channels increased due to the improved accuracy for LP sentences and the ceiling effect of HP sentence performance. However, for the L2 listeners, the greatest benefit, although only 12 percentage points, occurred in the 6-channel condition and maintained relatively stable as the number of spectral channels increased. Even in the unprocessed condition, the L2 listeners showed measurable higher recognition accuracy in the HP sentences than in the LP sentences. These results suggested that the contextual information played a consistent facilitating role in sentence recognition for the L2 listeners but they required more acoustic cues to maximize the benefit of sentence context. Another finding related to the importance of sentence context in sentence recognition in non-native listeners was the performance difference between HINT and CUNY sentences in our L2 listeners. The L2 listeners showed consistently higher recognition accuracy for CUNY sentences than for HINT sentences except for the 2-channel condition. HINT sentences have fewer words in each sentence and a fewer number of total words in each list, while CUNY sentences have a larger number of words in most sentences and varying numbers of words in individual sentences of each list. Although HINT sentences seem shorter and easier than CUNY sentences, there is probably less context information in HINT sentences than in CUNY sentences. The consistently higher recognition accuracy for CUNY than for HINT in the L2 listeners confirmed the facilitating role of sentence context in this group. The magnitude of the contextual benefit for the CUNY sentences in the L1 listeners above 4 channels was obscured due to the ceiling effects.

Although sentence context played a consistently facilitating role in speech understanding in L2 listeners, the low-level acoustic information provides the auditory-acoustic foundation to sentence recognition. Previous studies reported that consonants and vowels weigh disproportionally in sentence recognition in native English listeners (Fogerty et al., 2012; New et al., 2008; Owren and Cardillo, 2006). Generally, consonants carry more information for word meaning and lexical access, while vowels carry more information for auditory contextual cues for sentence processing. In the present study, we found that both consonants and vowels provided a significant contribution to sentence recognition in the L2 listeners. Note that the speech signals in the present study were spectrally degraded and the L2 listeners were native Mandarin speakers who showed less confusion and higher recognition accuracy for English consonants than for English vowels due to the native language effect. Therefore, when these L2 listeners were presented with degraded speech that contains a limited amount of acoustic information, they tended to utilize all information they could obtain from both consonants and vowels to recognize sentences.

Due to the relatively small number of the L2 participants and heterogeneous nature of this group in terms of the language learning factors, we did not divide the L2 listeners into subgroups. Instead, correlation and regression analyses were conducted to examine the relationship between the language learning factors and their overall perceptual performance in spectrally degraded speech in English. Consistent with previous studies which showed more native-like production and perception of phonetic contrasts with increased experience in L2 (Flege et al., 1997; Ingvalson et al., 2011), our results revealed that the L2 listeners who had resided in the U.S. for a longer time, recognized vocoded English phonemes and sentences with higher accuracies. While the correlation analysis revealed that those L2 listeners who had more language use in L2 tended to show a better performance in recognizing vocoded speech in English, the regression results indicated that the amount of L2 use did not account for variance in overall recognition accuracy beyond what was predicted by the factor of length of residence. Further, inconsistent with previous studies that showed a strong positive relationship between the age of L2 learning and speech-language abilities in L2 (Flege, 1991; Guion, 2003; Mayo et al., 1997; Weiss and Dempsey, 2008), our results showed no correlation between the age of English learning and the recognition accuracy. Our L2 listeners who had an early starting age of English learning were mainly young adults who just arrived at the U.S. for a shorter time. However, those L2 listeners who started English learning at a later age came to the U.S. at an early time and lived in this country for a longer time. Therefore, the effect of onset age of L2 learning might be confounded by the factor of length of residence. Moreover, although the L2 listeners in the present study started to learn English through classroom instruction in China at varying ages, they were immersed in the natural setting of an English environment at similar ages after adolescence. Previous studies reported the fundamental differences between classroom instruction and natural setting language learning (Freed et al., 2004) and the importance of in-country immersion in foreign language learning (Miller and Ginsberg, 1995; Naysmith and Corcoran, 2001). Therefore, the reported

onset age of English learning of the L2 learners did not reflect the initiation of the most effective language learning in these L2 learners. For future studies, we should recruit more L2 listeners with a better control for the language learning variables.

Based on the findings obtained from our listeners, it is reasonable to predict that when CI recipients are presented with speech stimuli in a second language, they may perform worse than they do in their first language. We acknowledge that noise-vocoded simulation differs from the actual signals received by CI users in several ways. On one hand, most CI users can only effectively utilize information delivered through seven or eight channels (Berg et al., 2020; Friesen et al., 2001). On the other hand, it is hard to achieve the precise frequency mapping between the allocated frequency bands and the actual frequency regions in the cochlea due to the physical difficulty of inserting the electrodes to the target positions (Davis et al., 2005; Dorman et al., 1997; Zhou et al., 2010). Therefore, CI users may show more difficulty in extracting and mapping the low-level acoustic information. In this case, their recognition of L2 speech may rely more heavily on the contextual information and overall linguistic knowledge.

A methodological limitation of the current study, necessitated by the linguistic status of the L2 listeners, lies in that the L2 listeners showed a wide age range, which served to examine how L2 experience impacted non-native listeners' perception of degraded speech. As the L2 listeners recruited in the present study all came to the U.S. in adulthood, the longer residence in the L2 environment meant that they were also older in the chronological age. The age gap between the L2 and L1 listeners resulted in a significant age effect on both phoneme and sentence recognition. However, we believe that the age effect was mainly caused by the L2 experience rather than the potential decline in cognitive and auditory processing that has been documented in previous studies (e.g., Atcherson et al., 2015; Humes and Dubno, 2010). Of the 43 L2 listeners, only four of them were older than 45 that included two in their 50′s. Since remarkable decline in auditory performance does not usually start until 50 years of age (Atcherson et al., 2015), the potential negative impact of older age in the L2 listeners should be limited in the present study. More importantly, as shown in Fig. 9, the recognition performance of our L2 listeners improved as their age increased, which was due to the increased experience in L2 as a function of increased length of residence in the U.S. For future studies, L1 listeners should be recruited from a wider age range to match with the L2 listeners. Another limitation of the present study was the lack of language ability test to quantify the L2 listeners' proficiency in English. While the survey questions addressed different aspects of L2 learning, none of them alone can sufficiently represent their proficiency or ability in the L2. Future studies should seek to adopt a vocabulary test or quick language test to determine the level of L2 proficiency, which can also be used to group L2 listeners. Finally, the phoneme and sentence tests were implemented in the same order to all participants even though the channel conditions and items in each test were randomized. For future studies, a fully randomized research design should be used.

5. Conclusion

In sum, our research provided further evidence showing the non-native deficits in recognizing spectrally-degraded (i.e., vocoded) speech. The deficits were reflected in the extraction of low-level acoustic information and application of high-level contextual cues. Compared to the L1 listeners, the L2 listeners showed greater phonemic confusions originating from the listeners' native language. The facilitating role of contextual cues in sentence recognition was consistently present in the L2 listeners but they required more spectral information to maximize the contextual benefit than the L1 listeners did. Finally, among the tested language learning factors, the overall perceptual performance of the L2 listeners was positively correlated with the amount of experience in English that was mainly represented by the length of residence in the U.S.

CRediT authorship contribution statement

Jing Yang: Conceptualization, Methodology, Data curation, Formal analysis, Writing – original draft, Writing – review & editing, Supervision. **Andrew Wagner:** Data curation. **Yu Zhang:** Data curation. **Li Xu:** Conceptualization, Methodology, Formal analysis, Writing – original draft, Writing – review & editing, Supervision.

Declaration of Competing Interest

There are no conflicts of interest to disclose.

References

Assmann, P., Summerfield, Q., 2004. The perception of speech under adverse conditions. Speech Processing in the Auditory System. Springer, New York, NY, pp. 231–308.

Astheimer, L.B., Berkes, M., Bialystok, E., 2016. Differential allocation of attention during speech perception in monolingual and bilingual listeners. Lang. Cogn. Neurosci. 31 (2), 196–205.

Atcherson, S.R., Nagaraj, N.K., Kennett, S.E., Levisee, M., 2015. Overview of central auditory processing deficits in older adults. Semin. Hear. 36 (3), 150–161.

Bashford, J.A., Warren, R.M., Brown, C.A., 1996. Use of speech-modulated noise adds strong "bottom-up" cues for phonemic restoration. Percept. Psychophys. 58 (3), 342–350.

Berg, K.A., Noble, J.H., Dawant, B.M., Dwyer, R.T., Labadie, R.F., Gifford, R.H., 2020. Speech recognition with cochlear implants as a function of the number of channels–Effects of electrode placement. J. Acoust. Soc. Am. 147 (5), 3646–3656.

Best, C.T., 1995. A direct realist view of cross-language speech perception. In: Strange, W (Ed.), Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research. York Press, Timonium, MD, pp. 171–203.

Bilger, R.C., Nuetzel, J.M., Rabinowitz, W.M., Rzeczkowski, C., 1984. Standardization of a test of speech perception in noise. J. Speech Lang. Hear. Res. 27 (1), 32–48.

Boothroyd, A., Hanin, L., Hnath, T., 1985. A Sentence Test of Speech Perception: Reliability, Set Equivalence, and Short Term Learning. NY Speech and Hearing Sciences Research Center. City University of New York City, New York.

Bradlow, A.R., Alexander, J.A., 2007. Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. J. Acoust. Soc. Am. 121 (4), 2339–2349.

Brouwer, S., Van Engen, K.J., Calandruccio, L., Bradlow, A.R., 2012. Linguistic contributions to speech-on-speech masking for native and non-native listeners–Language familiarity and semantic content. J. Acoust. Soc. Am. 131 (2), 1449–1464.

Calandruccio, L., Brouwer, S., Van Engen, K.J., Dhar, S., Bradlow, A.R., 2013. Masking release due to linguistic and phonetic dissimilarity between the target and masker speech. Am. J. Audiol. 22, 157–164.

Chen, Y., Wong, L.L., 2017. Speech perception in Mandarin-speaking children with cochlear implants–A systematic review. Int. J. Audiol. 56 (sup2), S7–S16.

Clahsen, H., Felser, C., 2006. How native-like is non-native language processing? Trends Cogn. Sci. (Regul. Ed.) 10 (12), 564–570.

Clopper, C.G., Pisoni, D.B., De Jong, K, 2005. Acoustic characteristics of the vowel systems of six regional varieties of American English. J. Acoust. Soc. Am. 118 (3), 1661–1676.

Corps, R.E., Rabagliati, H., 2020. How top-down processing enhances comprehension of noise-vocoded speech–Predictions about meaning are more important than predictions about form. J. Mem. Lang. 113, 104114.

Cutler, A., Weber, A., Smits, R., Cooper, N., 2004. Patterns of English phoneme confusions by native and non-native listeners. J. Acoust. Soc. Am. 116 (6), 3668–3678.

Cutler, A., Cooke, M., Lecumberri, M.L.G., Pasveer, D., 2007. L2 consonant identification in noise–Cross-language comparisons. In: Eighth Annual Conference of the International Speech Communication Association (Interspeech 2007), pp. 1585–1588.

Cutler, A., Garcia Lecumberri, M.L., Cooke, M., 2008. Consonant identification in noise by native and non-native listeners–Effects of local context. J. Acoust. Soc. Am. 124 (2), 1264–1268.

Davis, M.H., Johnsrude, I.S., Hervais-Adelman, A., Taylor, K., McGettigan, C., 2005. Lexical information drives perceptual learning of distorted speech–Evidence from the comprehension of noise-vocoded sentences. J. Exp. Psychol. 134 (2), 222–241.

Dell, G.S., Newman, J.E., 1980. Detecting phonemes in fluent speech. J. Verbal Learn. Verbal Behav. 19 (5), 608–623.

Dorman, M.F., Loizou, P.C., Rainey, D., 1997. Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. J. Acoust. Soc. Am. 102 (4), 2403–2411.

Escudero, P., 2009. Linguistic perception of "similar" L2 sounds. Phonol. Percept. 15, 152–190.

Field, J., 2004. An insight into listeners' problems–Too much bottom-up or too much top-down? System 32 (3), 363–377.

Flege, J.E., 1991. Age of learning affects the authenticity of voice-onset time (VOT) in stop consonants produced in a second language. J. Acoust. Soc. Am. 89 (1), 395–411.

Flege, J.E., Bohn, O.S., Jang, S., 1997. Effects of experience on non-native speakers' production and perception of English vowels. J. Phon. 25 (4), 437–470.

Fogerty, D., Humes, L.E., 2010. Perceptual contributions to monosyllabic word intelligibility–Segmental, lexical, and noise replacement factors. J. Acoust. Soc. Am. 128, 3114–3125.
Fogerty, D., Kewley-Port, D., Humes, L.E., 2012. The relative importance of consonant and vowel segments to the recognition of words and sentences–Effects of age and hearing loss. J. Acoust. Soc. Am. 132, 1667–1678.
Freed, B., Segalowitz, N., Dewey, D., 2004. Context of learning and second language fluency in French–Comparing regular classroom, study abroad, and intensive domestic immersion programs. Stud. Second Lang. Acquis. 26, 275–301.
Friesen, L.M., Shannon, R.V., Baskent, D., Wang, X., 2001. Speech recognition in noise as a function of the number of spectral channels–Comparison of acoustic hearing and cochlear implants. J. Acoust. Soc. Am. 110 (2), 1150–1163.
Guan, J., Liu, C., Tao, S., Mi, L., Wang, W., Dong, Q., 2015. Vowel identification in temporal-modulated noise for native and non-native listeners–Effect of language experience. J. Acoust. Soc. Am. 138 (3), 1670–1677.
Guion, S., 2003. The vowel systems of Quichua–Spanish bilinguals–An investigation into age of acquisition effects on the mutual influence of the first and second languages. Phonetica 60, 98–128.
Hansen, C., Jensen, C., 1994. Evaluating lecture comprehension. In: Flowerdew, J. (Ed.), Academic Listening. Cambridge University Press, Cambridge, pp. 241–268.
Hazan, V., Simpson, A., 1998. The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. Speech Commun. 24 (3), 211–226.
Heinrich, A., Flory, Y., Hawkins, S., 2010. Influence of English r-resonances on intelligibility of speech in noise for native English and German listeners. Speech Commun. 52 (11–12), 1038–1055.
Hill, F.J., McRae, L.P., McClellan, R.P., 1968. Speech recognition as a function of channel capacity in a discrete set of channels. J. Acoust. Soc. Am. 44 (1), 13–18.
Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K., 1995. Acoustic characteristics of American English vowels. J. Acoust. Soc. Am. 97 (5), 3099–3111.
Hornsby, B.W., Ricketts, T.A., 2001. The effects of compression ratio, signal-to-noise ratio, and level on speech recognition in normal-hearing listeners. J. Acoust. Soc. Am. 109 (6), 2964–2973.
Humes, L.E., Dubno, J.R., 2010. Factors Affecting Speech Understanding in Older Adults. Springer, New York, pp. 211–257.
Ingvalson, E.M., McClelland, J.L., Holt, L.L., 2011. Predicting native English-like performance by native Japanese speakers. J. Phon. 39 (4), 571–584.
Jacewicz, E., Fox, R.A., Salmons, J., 2011. Cross-generational vowel change in American English. Lang. Var. Change 23 (1), 45–86.
Kewley-Port, D., Burkle, T.Z., Lee, J.H., 2007. Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. J. Acoust. Soc. Am. 122 (4), 2365–2375.
Kim, B.J., Chang, S.A., Yang, J., Oh, S.H., Xu, L., 2015. Relative contributions of spectral and temporal cues to Korean phoneme recognition. PLoS One 10, e0131807.
Kong, Y.Y., Donaldson, G., Somarowthu, A., 2015. Effects of contextual cues on speech recognition in simulated electric-acoustic stimulation. J. Acoust. Soc. Am. 137 (5), 2846–2857.
Koster, C.J., 1987. Word Recognition in Foreign and Native Language: Effects of Context and Assimilation. Foris Publications, Dordrecht.
Kuhl P.K. (1993) Innate predispositions and the effects of experience in speech perception–The native language magnet theory. In: de Boysson-Bardies B., de Schonen S., Jusczyk P., McNeilage P., Morton J. (eds) Developmental Neurocognition – Speech and Face Processing in the First Year of Life. NATO ASI Series (Series D – Behavioural and Social Sciences), vol 69. Dordrecht: Springer.
Labov, w., Ash, S., Boberg, C., 2006. Atlas of North American English: Phonetics, Phonology, and Sound Change. Mouton de Gruyter, Berlin.
Lecumberri, M.L.G., Cooke, M., Cutler, A., 2010. Non-native speech perception in adverse conditions–A review. Speech Commun. 52 (11–12), 864–886.
Lecumberri, M.L.G, Cooke, M., 2006. Effect of masker type on native and non-native consonant perception in noise. J. Acoust. Soc. Am. 119 (4), 2445–2454.
Loizou, P.C., Dorman, M., Tu, Z., 1999. On the number of channels needed to understand speech. J. Acoust. Soc. Am. 106 (4), 2097–2103.
Lovitt, A., Allen, J.B., 2006. 50 years late–Repeating miller-nicely 1955. In: The Ninth International Conference on Spoken Language Processing.
Mack, M., Tierney, J., Boyle, M.E., 1990. The Intelligibility of Natural and LPC-vocoded Words and Sentences Presented to Native and Non-Native Speakers of English (No. TR-869). MIT Lexington Lincoln Lab.
Marslen-Wilson, W.D., 1987. Functional parallelism in spoken word-re cognition. Cognition 25 (1–2), 71–102.
Mattys, S.L., Davis, M.H., Bradlow, A.R., Scott, S.K., 2012. Speech recognition in adverse conditions–A review. Lang. Cogn. Process. 27 (7–8), 953–978.
Mayo, L.H., Florentine, M., Buus, S., 1997. Age of second-language acquisition and perception of speech in noise. J. Speech Lang. Hear. Res. 40 (3), 686–693.
McClelland, J.L., Elman, J.L., 1986. The TRACE model of speech perception. Cogn. Psychol. 18 (1), 1–86.
Miller, G.A., Nicely, P.E., 1955. An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Am. 27 (2), 338–352.
Miller, L., & Ginsberg, R. (1995). Folklinguistic theories of language learning. In B. F. Freed (Ed.), Second Language Acquisition in a Study Abroad Context (pp. 293- 316). Philadelphia: John Benjamins.
Mueller, G., 1980. Visual contextual cues and listening comprehension–An experiment. Mod. Lang. J. 64, 335–340.
Naysmith, J., Corcoran, S. (2001). Culture shocks–Immersion education at the University College Chichester. In P. Bodycott & V. Crew (Eds.), Language and Cultural Immersion – Perspectives on Short Term Study and Residence Abroad (pp. 81–89). Hong Kong: The Hong Kong Institute of Education.
New, B., Araújo, V., Nazzi, T., 2008. Differential processing of consonants and vowels in lexical access through reading. Psychol. Sci. 19, 1223–1227.
Nilsson, M., Soli, S.D., Sullivan, J.A., 1994. Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. J. Acoust. Soc. Am. 95 (2), 1085–1099.
Nittrouer, S., Lowenstein, J.H., 2010. Learning to perceptually organize speech signals in native fashion. J. Acoust. Soc. Am. 127 (3), 1624–1635.
Norris, D., 1994. Shortlist–A connectionist model of continuous speech recognition. Cognition 52 (3), 189–234.
Owren, M.J., Cardillo, G.C., 2006. The relative roles of vowels and consonants in discriminating talker identity versus word meaning. J. Acoust. Soc. Am. 119, 1727–1739.
Padilla, M., Shannon, R.V., 2000. English phoneme and word recognition by non-native English speakers as a function of spectral resolution and English experience. J. Acoust. Soc. Am. 108 (5), 2651–2652.
Padilla, M., Shannon, R.V., 2002. Could lack of experience with a second language be modeled as a hearing loss? J. Acoust. Soc. Am. 112 (5), 2385–2385.
Peterson, N.R., Pisoni, D.B., Miyamoto, R.T., 2010. Cochlear implants and spoken language processing abilities–Review and assessment of the literature. Restor. Neurol. Neurosci. 28 (2), 237–250.
Patro, C., Mendel, L.L., 2016. Role of contextual cues on the perception of spectrally reduced interrupted speech. J. Acoust. Soc. Am. 140 (2), 1336–1345.
Remez, R.E., Rubin, P.E., Pisoni, D.B., Carrell, T.D., 1981. Speech perception without traditional speech cues. Science 212, 947–949.
Roberts, B., Summers, R.J., Bailey, P.J., 2011. The intelligibility of noise-vocoded speech–Spectral information available from across-channel comparison of amplitude envelopes. Proc. R. Soc. B: Biol. Sci. 278, 1595–1600.
Rogers, C.L., Lister, J.J., Febo, D.M., Besing, J.M., Abrams, H.B., 2006. Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. Appl. Psycholinguist. 27 (3), 465–485.
Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Rudner, M., 2013. The Ease of Language Understanding (ELU) model–Theoretical, empirical, and clinical advances. Front. Syst. Neurosci. 7, 31.
Shannon, R.V., Fu, Q.J., Galvin Iii, J., 2004. The number of spectral channels required for speech recognition depends on the difficulty of the listening situation. Acta Otolaryngol. 124 (0), 50–54.
Shannon, R.V., Jensvold, A., Padilla, M., Robert, M.E., Wang, X., 1999. Consonant recordings for speech testing. J. Acoust. Soc. Am. 106 (6), L71–L74.
Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270 (5234), 303–304.
Signoret, C., Johnsrude, I., Classon, E., Rudner, M., 2018. Combined effects of form-and meaning-based predictability on perceived clarity of speech. J. Exp. Psychol.: Hum. Percept. Perform. 44, 277–285.
Sohoglu, E., Peelle, J.E., Carlyon, R.P., Davis, M.H., 2014. Top-down influences of written text on perceived clarity of degraded speech. J. Exp. Psychol.: Hum. Percept. Perform. 40 (1), 186.
Sparreboom, M., van Schoonhoven, J., van Zanten, B.G., Scholten, R.J., Mylanus, E.A., Grolman, W., Maat, B., 2010. The effectiveness of bilateral cochlear implants for severe-to-profound deafness in children–A systematic review. Otology & Neurotology 31 (7), 1062–1071.
Strange, W., 2011. Automatic selective perception (ASP) of first and second language speech–A working model. J. Phon. 39 (4), 456–466.
Tobin, S.J., Nam, H., Fowler, C.A., 2017. Phonetic drift in Spanish-English bilinguals–Experiment and a self-organizing model. J. Phon. 65, 45–59.
Tsui, A., Fullilove, J., 1998. Bottom-up or top-down processing as a discriminator of L2 listening performance. Appl. Linguist. 19, 432–451.
Tyler, L.K., Voice, J.K., Moss, H.E., 2000. The interaction of meaning and sound in spoken word recognition. Psychon. Bull. Rev. 7 (2), 320–326.
Wang, X., Xu, L., 2021. Speech perception in noise–Masking and unmasking. J. Otol. 16 (2), 109–119.
Warren, R.M., 1970. Perceptual restoration of missing speech sounds. Science 167 (3917), 392–393.
Weiss, D., Dempsey, J.J., 2008. Performance of bilingual speakers on the English and Spanish versions of the Hearing in Noise Test (HINT). J. Am. Acad. Audiol. 19 (1), 5–17.
Wild, C.J., Yusuf, A., Wilson, D.E., Peelle, J.E., Davis, M.H., Johnsrude, I.S., 2012. Effortful listening–The processing of degraded speech depends critically on attention. J. Neurosci. 32 (40), 14010–14021.
Wright, R. (2004). A review of perceptual cues and cue robustness. In Hayes, B., Kirchner, R., and Steriade, D. (Eds.), Phonetically-Based Phonology, Cambridge University Press.
Xu, L., Pfingst, B.E., 2008. Spectral and temporal cues for speech recognition–Implications for auditory prostheses. Hear. Res. 242 (1–2), 132–140.
Xu, L., Thompson, C.S., Pfingst, B.E., 2005. Relative contributions of spectral and temporal cues for phoneme recognition. J. Acoust. Soc. Am. 117 (5), 3255–3267.
Xu, L., Xi, X., Patton, A., Wang, X., Qi, B., Johnson, L., 2021. A cross-language comparison of sentence recognition using American English and Mandarin Chinese HINT and AzBio sentences. Ear Hear. 42 (2), 405–413.
Xu, L., Zheng, Y., 2007. Spectral and temporal cues for phoneme recognition in noise. J. Acoust. Soc. Am. 122, 1758–1764.
Yang, J., Zhang, Y., Li, A., Xu, L., 2017. On the duration of Mandarin tones. In: Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech 2017). Stockholm, Sweden, pp. 1407–1411.
Zhou, N., Xu, L., Lee, C.-Y., 2010. The effects of frequency-place shift on consonant confusion in cochlear implant simulations. J. Acoust. Soc. Am. 128, 401–409.