Relative contributions of acoustic temporal fine structure and envelope cues for lexical tone perception in noise

Beier Qi^{a)}

Department of Otolaryngology—Head and Neck Surgery, Beijing Tongren Hospital, Capital Medical University, Beijing, China

Yitao Mao

Department of Radiology, Xiangya Hospital, Central South University, Changsha, Hunan, China

Jiaxing Liu^{a)} and Bo Liu^{a)}

Department of Otolaryngology—Head and Neck Surgery, Beijing Tongren Hospital, Capital Medical University, Beijing, China

Li Xu^{b)}

Communication Sciences and Disorders, Ohio University, Athens, Ohio 45701, USA

(Received 24 October 2016; revised 21 March 2017; accepted 11 April 2017; published online 3 May 2017)

Previous studies have shown that lexical tone perception in quiet relies on the acoustic temporal fine structure (TFS) but not on the envelope (E) cues. The contributions of TFS to speech recognition in noise are under debate. In the present study, Mandarin tone tokens were mixed with speech-shaped noise (SSN) or two-talker babble (TTB) at five signal-to-noise ratios (SNRs; -18 to $+6$ dB). The TFS and E were then extracted from each of the 30 bands using Hilbert transform. Twenty-five combinations of TFS and E from the sound mixtures of the same tone tokens at various SNRs were created. Twenty normal-hearing, native-Mandarin-speaking listeners participated in the tone-recognition test. Results showed that tone-recognition performance improved as the SNRs in either TFS or E increased. The masking effects on tone perception for the TTB were weaker than those for the SSN. For both types of masker, the perceptual weights of TFS and E in tone perception in noise was nearly equivalent, with E playing a slightly greater role than TFS. Thus, the relative contributions of TFS and E cues to lexical tone perception in noise or in competing-talker maskers differ from those in quiet and those to speech perception of non-tonal languages.

© 2017 Acoustical Society of America. [<http://dx.doi.org/10.1121/1.4982247>]

[MSS]

Pages: 3022–3029

I. INTRODUCTION

The time waveforms of acoustic signals can be mathematically decomposed into a slowly varying envelope (E) and a faster-oscillating carrier of a constant amplitude referred to as the temporal fine structure (TFS) based on Hilbert transform. In the present paper, we use E and TFS to refer to the acoustic E and acoustic TFS as opposed to the neural E and neural TFS that are derived in the auditory pathway (see Shamma and Lorenzi, 2013; Moon *et al.*, 2014). It is worth noting that there is a lack of one-to-one mapping between the acoustic E and the neural representation of E cues and between the acoustic TFS and the neural representation of TFS cues in the early auditory system, as demonstrated by Shamma and Lorenzi (2013) and Moon *et al.* (2014). Despite such limitations in vocoder-based studies, the perceptual dichotomy of the two acoustic cues (E and TFS) has been studied in a number of laboratories since the work by Smith *et al.* (2002; see, e.g., Xu and Pfingst, 2003; Zeng *et al.*, 2004; Lorenzi *et al.*, 2006; Gilbert and

Lorenzi, 2006; Moore, 2008; Fogerty, 2011; Wang *et al.*, 2011a; Wang *et al.*, 2015; Wang *et al.*, 2016; Apoux *et al.*, 2013). A consensus has been reached among researchers that E cues are important for speech perception in quiet conditions whereas TFS cues are important for pitch-related perception, including music perception and lexical tone perception in tone languages. Nonetheless, Lorenzi and colleagues have presented evidence that normal-hearing listeners with sufficient training can use TFS cues to achieve good speech recognition in quiet (Lorenzi *et al.*, 2006; Gilbert and Lorenzi, 2006). However, it is known that E information can be reconstructed from the TFS signals at the output of the cochlear filters (Ghitza, 2001; Zeng *et al.*, 2004). Recent empirical evidence suggests that reconstructed E information may account for the intelligibility of the TFS speech (Swaminathan *et al.*, 2014; Léger *et al.*, 2015).

The relative contributions of the acoustic E and TFS in speech perception in noise have become a matter of debate in recent years. In vocoder processing, the temporal envelopes of a number of frequency bands are preserved and the TFS is replaced by noise bands or tones (Shannon *et al.*, 1995; Dorman *et al.*, 1997; Loizou *et al.*, 1999; Xu and Pfingst, 2008). Vocoder processing of speech materials mixed with noise produced poor speech-recognition performance in

^{a)}Also at: Key Laboratory of Otolaryngology—Head and Neck Surgery (Ministry of Education), Beijing Institute of Otolaryngology, Beijing, China.

^{b)}Electronic mail: xul@ohio.edu

normal-hearing listeners (Fu *et al.*, 1998; Dorman *et al.*, 1998; Xu and Zheng, 2007). This observation is reminiscent of what has been reported in cochlear implant users in that they perform well in speech recognition in quiet but show dramatic difficulties in speech recognition in noise (e.g., Fu *et al.*, 1998; Wilson, 2008). In most of the current vocoder-based cochlear implant strategies (Loizou, 2006), constant-rate pulse trains are modulated by the envelopes of the input signals and, thus, the acoustic TFS is not encoded in the cochlear implant stimulations. The lack of TFS information in the cochlear implant stimulations is considered to be one of many potential reasons why cochlear implant users perform poorly in noise conditions. More direct evidence that TFS information contributes to speech recognition in noise comes from the Hopkins *et al.* (2008) study (see also Moore, 2008). In that study, speech signals were divided into 32 frequency bands (channels) and progressively more higher-frequency channels were vocoder processed so as to remove the TFS information in those channels until all channels were vocoder processed. The speech-reception thresholds in a competing single-talker background measured in the normal-hearing listeners increased dramatically as the TFS was removed from progressively more of the channels. Moore and colleagues concluded that TFS information plays an important role for speech recognition in fluctuating background noise (Moore, 2008). More recently, Moon *et al.* (2014) combined psychophysical experiments with simulations of a computational model of neural E and TFS coding in the early auditory system. Their results were consistent with the notion that acoustic TFS cues play a role in speech identification in noise. In particular, with acoustic TFS information, speech identification in fluctuating noise is better than that in steady-state noise, resulting in the so-called “speech masking release” (Moon *et al.*, 2014).

Apoux *et al.* (2013) adopted an auditory chimera technique (Smith *et al.*, 2002; Fogerty, 2011) and investigated the role and relative contributions of E and TFS to sentence recognition in noise. In this modified auditory chimera technique, English speech materials and maskers (i.e., steady-state noise or a single-talker competing speech) were added at different signal-to-noise ratios (SNRs) and filtered into 30 contiguous frequency bands. The E and TFS were extracted from each band using Hilbert transform. The E and TFS from the mixture of maskers to the same sentence but of different SNRs (−18 to +6 dB) were swapped to create an auditory chimera stimulus. Results showed that sentence recognition in either steady-state noise or single-talker competing-speech masker relies almost exclusively on the E cues but very little on the TFS cues. In a follow-up experiment, the stimuli with only the TFS of the masker (i.e., −1000 dB SNR for the target TFS) were found as intelligible as those with only the TFS of the target (i.e., +1000 dB SNR for the target TFS). The authors concluded that acoustic TFS information contributes little if any to speech recognition in noise, but rather serves as a grouping cue to select the time-frequency regions corresponding to the target speech signal (Apoux *et al.*, 2013; Apoux and Healy, 2013).

More recently, Nambi *et al.* (2016) measured speech-recognition performance in four-talker babble. Sentence materials were bandpassed through a 30-band filter. The TFS

was then (1) preserved, (2) time-reversed, or (3) replaced with noise for each of the 30 bands. The sentence recognition performance in the multitalker babble was much better with the preserved TFS than with noise TFS. The performance with the time-reversed TFS was better than that with noise TFS but poorer than that with the original TFS. The authors suggested that both stream segregation and envelope recovery aided TFS-mediated speech recognition in noise.

Lexical tone, used in tone languages, is the pitch contour carried at the syllabic level of a word that defines the meaning of the word. In Mandarin Chinese, there are four lexical tones characterized by their pitch contours: (1) high and flat, (2) rising, (3) falling and then rising, and (4) falling. Lexical tone perception is part of speech perception. However, its perceptual mechanisms are based on those of pitch perception that is shared with music perception (Wang *et al.*, 2011b; see Xu and Zhou, 2011, for a review). Therefore, it is important to understand what acoustic features contribute to tone perception. Xu and Pfingst (2003) have shown that when conflicting TFS and E information about lexical tones is present, as in auditory chimera stimuli, native Mandarin-speaking, normal-hearing listeners perceive the tones based predominantly on the TFS information. Such a finding has been replicated in a number of recent studies (Wang *et al.*, 2011a; Wang *et al.*, 2015; Wang *et al.*, 2016). In studies using vocoder processing, E cues in a number of frequency bands without TFS cues can provide tone recognition of ~75% correct (Xu *et al.*, 2002; also see Xu and Pfingst, 2008, for a review). Thus, without the detailed “spectro-temporal modulation cues” (i.e., spectral and TFS cues) conveying F0 information, lexical tone perception suffers in native tone-language-speaking adult listeners (Cabrera *et al.*, 2014). Kong and Zeng (2006) further showed that tone recognition reached nearly perfect performance with 32-channel noise vocoders. In addition, under noise conditions, tone recognition with E cues (referred to as amplitude-modulation cues by the authors) decreased dramatically. With additional slow-varying frequency-modulation information (referred to as TFS by the authors), tone recognition in noise improved in conditions where the SNR was −5 dB (Kong and Zeng, 2006).

What are the relative contributions of TFS and E (as defined by Hilbert transform) to lexical tone perception in noise? Previous studies have shown that tone perception in noise is fairly robust. At an SNR of −5 dB, normal-hearing, native Mandarin-speaking listeners can achieve >90% correct tone recognition on average (Kong and Zeng, 2006; Krenmayr *et al.*, 2011; Lee *et al.*, 2013). Does TFS dominate tone perception in noise or competing-speech maskers as it does in quiet? Or, is the contribution of TFS in noise limited as being demonstrated for English speech recognition in noise (Apoux *et al.*, 2013)? The present study was designed to address the above research questions. We adopted the same technique as in Fogerty (2011) and Apoux *et al.* (2013) to study the relative contributions of TFS and E information to lexical tone perception in noise. Two types of maskers, speech-shaped noise (SSN) and two-talker babble (TTB), were used in the present study. Furthermore, correlation analysis of the E and TFS between the original tone tokens

and the masker mixtures was performed in order to reveal the similarities of the two cues in the two types of maskers to those in the original speech signals.

II. MATERIALS AND METHODS

A. Subjects

Twenty normal-hearing, native-Mandarin-speaking listeners (ten females and ten males) participated in this study. Their ages ranged from 19 to 30 years old [mean $\pm$ standard deviation (SD): 24.2 ± 3.2]. Their normal hearing was verified with pure-tone air-conduction thresholds of 20 dB hearing level (HL) or better at octave frequencies from 250 to 8000 Hz. No participant reported any history of speech or hearing disorders. The use of human participants in this study was reviewed and approved by the Institutional Review Boards of Ohio University and Beijing Institute of Otolaryngology.

B. Speech materials and signal processing

The tone stimuli included ten Chinese monosyllables (i.e., /fu/, /ji/, /ma/, /qi/, /wan/, /xi/, /xian/, /yan/, /yang/, and /yi/), each with four lexical tones. The tone tokens were recorded with a male and a female speaker, both being native-Mandarin-speaking adults. For each monosyllable, the duration of the four tones was equalized using a script written in PRAAT (Boersma and Weenink, 2016) that was provided by Darwin (2005). The mean duration of the four tones of a monosyllable was chosen as the target for equalization. The PRAAT script changed the syllable duration without changing the F0.

The duration-equalized tone tokens were processed using a signal processing scheme similar to that of Apoux *et al.* (2013). We used two types of masker, SSN (Nilsson *et al.*, 1994) and TTB. The latter consisted of root-mean-square (RMS)-equalized speech segments from a male and a female narrative speech in Mandarin Chinese. These two speakers were not the same speakers who produced the target tone tokens. First, the tone tokens were mixed with the masker at five SNRs (i.e., -18, -12, -6, 0, and +6 dB). Then, the mixtures of target and masker in the overall bandwidth of 80–7563 Hz were filtered into 30 frequency bands with a third-order elliptic bandpass filter. The bandwidth of each band was 1 ERB (equivalent rectangular bandwidth, Glasberg and Moore, 1990) wide. Hilbert transform was used to extract E and TFS from each frequency band. The E extracted from a given band at a certain SNR was then combined with the TFS of the same tone token filtered through the same band, but at a different SNR level. After passed through the initial analysis filters, the signals from the 30 bands were summed to generate the chimeric tone tokens (see Apoux *et al.*, 2013, for details). The stimuli of a total of 50 chimera conditions (5 SNRs for E $\times$ 5 SNRs for TFS $\times$ 2 types of maskers) were processed and stored in a computer hard disk for presentations.

C. Procedures

All participants completed the tone recognition test carried out in a sound-treated room. During the test, the participants wore a pair of Sennheiser HD 280 Pro circum-aural headphones (Sennheiser, Wedemark, Germany). The tested ear was selected at random and the stimulus level with equalized RMS for all stimuli was set at the most comfortable level for each individual, which was ~ 65 dB sound pressure level (SPL; A weighted). The stimuli were presented using a custom graphical user interface (GUI) programmed in MATLAB. For each stimulus presentation, the GUI displayed the syllable (e.g., /fu/), the tones (i.e., 1, 2, 3, and 4), and the Chinese characters associated with them (e.g., 夫, 服, 府, and 父). Participants were asked to identify the Chinese word by clicking the corresponding button on the GUI. The responses were collected from the GUI for later analysis.

For the two types of maskers, experiments with the SSN were run first and those with the TTB were run several months later for each participant. For each type of noise, a brief training session was provided to the participants so that they became familiarized with the task and the processed sounds before the real test. The training session included 80 processed tone tokens that were drawn from 4 processed conditions [i.e., (1) E at 0 dB and TFS at 0 dB, (2) E at -12 dB and TFS at 0 dB, (3) E at 0 dB and TFS at -12 dB, and (4) E at -12 dB and TFS at -12 dB SNRs]. Feedback was provided during the training session. In the real test, the 25 processed conditions [i.e., 5 SNRs (-18, -12, -6, 0, and +6 dB) for E combined with 5 SNRs (-18, -12, -6, 0, and +6 dB) for TFS] were randomized. For each condition, there were 80 tone tokens (i.e., 10 monosyllables $\times$ 4 tones $\times$ 2 speakers). The entire experiment took ~ 7 h for each participant who completed it in multiple sessions.

D. Acoustic analysis of E and TFS

In previous work, correlation analyses were performed to evaluate the amount of E recovery from TFS speech at the output of cochlear filtering (Zeng *et al.*, 2004; Gilbert and Lorenzi, 2006; Li *et al.*, 2015) or to compare the similarities of the acoustic E or TFS of two type of speech materials (Apoux *et al.*, 2013; Xu, 2016). The latter approach was adopted in the present study to allow us to examine the similarities of the acoustic E or TFS between the original tone tokens and the masker-corrupted tone tokens. In these analyses, each of the 80 tone tokens was mixed with either SSN or TTB at an SNR between -18 and +18 dB in 3-dB steps. The signals were passed through a bank of 30 contiguous filters [same as above (see Sec. II B) in the chimera processing]. The E and TFS of each of the 30 frequency bands were extracted using Hilbert transform. The same operation was performed on the original tone tokens without adding any noise. A Pearson correlation of the E or TFS of each of the 30 frequency bands was performed between the masker-mixed tone tokens and the original tone tokens. The 30 correlation coefficients of one tone token were averaged to represent the correlation of that particular token. The mean and SD of the correlation coefficients across all 80 tone tokens

FIG. 1. Average tone-recognition scores across the 20 normal-hearing listeners in the SSN conditions. (Left) Tone-recognition scores as a function of SNR in TFS. Each line represents the data for SNR in E. (Middle) Tone-recognition scores as a function of SNR in E. Each line represents the data for SNR in TFS. (Right) Contour plot of the average tone-recognition scores. The abscissa and the ordinate represent SNR in E and in TFS, respectively. Color represents the tone-recognition performance as indicated by the color bar on the right.

were then computed to represent the correlation of E or TFS at a particular SNR of a certain type of masker.

E. Data analysis

Data analysis was performed in MATLAB with the Statistics Toolbox. Percent-correct scores of tone recognition were first computed for each test condition for each participant. The percent-correct scores were treated as binomial data. Following a logit transformation of the percent-correct data, a generalized linear model (GLM; [Warton and Hui, 2011](#)) was used to examine the effects of the SNR in TFS and the SNR in E on tone-recognition performance, as well as those of type of masker on tone-recognition scores.

III. RESULTS

The group average tone-recognition scores for the SSN and TTB conditions are shown in Figs. 1 and 2, respectively. The SDs, not shown in Figs. 1 and 2, across all 25 processing conditions ranged from 3.9% to 9.5% (mean=6.0%) in the SSN conditions and from 4.0% to 7.8% (mean=5.5%) in the TTB conditions. In Figs. 1 and

2, the left panel shows the tone-recognition scores as a function of SNR in TFS, with each line representing the data for SNR in E. The middle panel shows the same data. Each line plots data from a particular SNR in TFS as a function of SNR in E. In both the left and middle panels, tone-recognition scores increased as SNR, in either TFS or E, increased. When the SNR in both TFS and E was equal, a baseline situation equivalent to tone recognition in noise, the average tone-recognition scores were 27.6%, 60.2%, 82.1%, 93.9%, and 94.7% correct in the SSN conditions at SNRs of -18, -12, -6, 0, and +6 dB, respectively (Fig. 1) and were 53.5%, 72.0%, 86.4%, 92.7%, and 95.0% correct in the TTB conditions at the same SNRs (Fig. 2). A GLM analysis revealed that the average tone-recognition scores between the two types of maskers were significantly different at SNRs of -18, -12, and -6 dB (all $p < 0.001$) but not at SNRs of 0 and +6 dB (both $p > 0.05$). In the right panel of Figs. 1 and 2, to assist better visualization of the data, a contour plot shows the performance scores in color with SNRs in E and TFS as the abscissa and the ordinate, respectively. The contours are roughly parallel to the main diagonal of the plot.

FIG. 2. Average tone-recognition scores across the 20 normal-hearing listeners in the TTB conditions. Conventions same as Fig. 1.

A GLM analysis was performed to examine the effects of TFS and E on tone-recognition scores. The SNR in TFS and the SNR in E were the two main factors in the GLM fitting. In the SSN conditions, the coefficients (β) of the GLM fitting for SNR in TFS, SNR in E, and the interaction between SNRs in TFS and E were 0.070 ($t=26.0$, $p<0.0001$), 0.095 ($t=36.7$, $p<0.0001$), and -0.002 ($t=-8.8$, $p<0.0001$), respectively. In the TTB conditions, the coefficients (β) for SNR in TFS, SNR in E, and the interaction between SNRs in TFS and E were 0.052 ($t=19.6$, $p<0.0001$), 0.073 ($t=28.8$, $p<0.0001$), and -0.0003 ($t=-1.5$, $p=0.13$), respectively. These results indicated that both TFS and E contributed significantly to tone recognition in various SNRs. The interaction between SNRs in TFS and E was statistically significant in the SSN conditions but not in the TTB conditions. Such an interaction and the lack of it could be visualized in the line plots of the data shown in Fig. 1 (nonparallel lines for the SSN conditions) and Fig. 2 (more or less parallel lines for the TTB conditions), respectively.

The Pearson correlation coefficient of each tone token was computed for the E or TFS between the original signal and the noise mixture at 13 SNRs (-18 to $+18$ dB in 3-dB steps). Figure 3(top) shows the mean data across the 80 tone tokens. The SD of the correlation coefficients, not shown in Fig. 3, was, on average, 0.064 across the 80 tone tokens. The correlation coefficients were greater in TTB than in SSN conditions. In either masker condition, the correlation increased for both E and TFS as the SNR increased. The coefficients for the E were greater than those for the TFS in the positive SNRs. For each type of masker (i.e., SSN or TTB), we computed the arithmetic mean of the correlation coefficients of E and TFS at any of the combinations of SNR in E and SNR in TFS. The two resultant 13×13 matrices of mean correlation coefficients were plotted in the image plots shown in the middle and bottom panels of Fig. 3 for the SSN and TTB conditions, respectively.

IV. DISCUSSION

Using the same auditory-chimera processing scheme as that devised in Apoux *et al.* (2013), we showed that lexical tone recognition in noise or competing-speech maskers relies on both acoustic TFS and E information of the syllables (Figs. 1 and 2). These results are different from those reported with English speech recognition in noise (Apoux *et al.*, 2013). In the Apoux *et al.* (2013) study, the authors found that English speech-recognition performance depended mainly or almost exclusively on the SNR of the E in the presence of the steady-state noise or a competing-talker masker. The perception mechanism for tone recognition is different from that of speech recognition in English. Tone recognition relies on the coding of the F0. In quiet listening conditions, English speech recognition relies on the E cues (Smith *et al.*, 2002). On the other hand, lexical tone recognition is known to be dependent on the TFS cues in quiet (Xu and Pfingst, 2003; Wang *et al.*, 2011a; Wang *et al.*, 2015; Wang *et al.*, 2016). The dominant contribution of TFS to tone recognition in quiet is similar to that found in music

FIG. 3. Correlation of E and TFS between the original and noise-mixed tone tokens. (Top) Correlation coefficients (r) of E and TFS between the original tone tokens and those mixed with two types of noise. The solid red lines and the dotted blue lines represent correlation coefficients for the E and TFS, respectively. The circle and triangle symbols represent SSN and TTB conditions, respectively. (Middle) Image plot of the mean correlation coefficients between those of E and TFS for the SSN conditions. The color of each element of the image represents the mean of the correlation coefficient of E at a particular SNR and that of TFS at a particular SNR. The white contour lines demarcate regions of similar values of the mean correlation coefficients. (Bottom) Same as the middle panel, but for the TTB conditions.

perception (Smith *et al.*, 2002). In the masker conditions, although the TFS cues are not the dominant cues for lexical tone recognition, it can still contribute significantly to tone recognition. Importantly, there is a trade-off in relative

contributions to tone recognition in noise between the TFS and E cues. When the SNR in TFS is low (e.g., ≤ -6 dB), a favorable SNR in the E (e.g., ≥ 0 dB) can provide good tone recognition. Likewise, when the SNR in E is low (e.g., ≤ -6 dB), a favorable SNR in TFS (e.g., ≥ 0 dB) can also provide good tone recognition to the normal-hearing listeners. Quantitatively, the contribution of E to tone recognition in noise is slightly greater than that of TFS. The contours shown in Figs. 1 and 2 (right panels), are more vertical-going than horizontal-going. These observations have been confirmed by the GLM fitting results in which the coefficients (β) of the GLM fitting of the percent-correct data as a function of SNR in the E were slightly larger than those for SNR in the TFS.

Correlation analyses have been used in several studies either to quantify the amount of E recovered from the TFS speech signal as compared to the original speech signals (e.g., Zeng *et al.*, 2004; Gilbert and Lorenzi, 2006; Li *et al.*, 2015) or to depict the similarities of the E or TFS between two types of signals (Apoux *et al.*, 2013; Xu, 2016). In the present study, we adopted the latter approach. Here, we computed the correlation coefficients of E or TFS between the original tone tokens and those mixed with maskers of various SNRs. We found that both E and TFS in the masker mixtures bore tremendous resemblance to those of the original tone tokens especially at the positive SNRs (Fig. 3, upper panel). The mean correlation coefficients of E and TFS at various SNRs (Fig. 3, middle and bottom panels) show a pattern that is similar to that of the tone recognition performance (Figs. 1 and 2, right panels). Here, we did not adjust the weights of the E or TFS but used the arithmetic mean of the correlation coefficients. If we assume that perception is based on the correlation of the acoustic signals, the similar patterns between the mean correlation coefficients and the recognition data would suggest that the weights of E and TFS are equivalent for lexical tone perception in noise or competing-speech maskers.

Our results in the present study are also different from those found in quiet listening conditions wherein the acoustic TFS information plays a dominant role in lexical tone recognition (Xu and Pfingst, 2003; Wang *et al.*, 2011a; Wang *et al.*, 2015; Wang *et al.*, 2016). Therefore, noise or competing-speech maskers can disrupt the TFS information and limit its contribution to tone recognition. On the other hand, acoustic E cues play an increasingly important role in noise or competing-speech maskers for tone recognition despite that they play a minimum role for tone recognition in quiet. These results were reminiscent of the findings in hearing-impaired listeners. As the degree of hearing loss increases, listeners rely more on the E cues and less on the TFS cues for tone recognition (Wang *et al.*, 2011a). The mechanisms for hearing-impaired listeners performing tone recognition in quiet and those for normal-hearing listeners performing tone recognition in maskers are different. One is probably related to a widened frequency tuning and/or impaired phase-locking ability in the hearing-impaired listeners. The other is related to a smearing of the TFS information. Based on the correlation analysis on the original speech and the masker-corrupted signals, it appears that the E is

more resilient to the smearing effects of the noise (Apoux *et al.*, 2013). Our correlation analyses were consistent with that in Apoux *et al.* (2013) in that the correlation coefficients were greater for the E than for the TFS when the SNRs were ≥ 0 dB (Fig. 3). Thus, when the TFS information is limited in masker conditions, listeners resort to the use of the E information. Tone-recognition performance maintains fairly high in noise as long as the SNR in either E or TFS is ≥ 0 dB.

Even in the adverse noise conditions, tone recognition is quite robust (see Kong and Zeng, 2006; Krenmayr *et al.*, 2011; Lee *et al.*, 2013). In the comparable SNR conditions for both steady-state and fluctuating maskers, tone recognition performance in the present study appeared to be higher than English sentence recognition as reported in Apoux *et al.* (2013). However, it is unfair to directly compare Mandarin tone recognition and English sentence recognition scores. Note that tone recognition, being a four-alternative forced-choice task, is a simpler test than sentence recognition. Also, listeners might utilize both TFS and E information for tone recognition, whereas for speech recognition listeners appear to use only the E cues in noise. Tone-recognition performance in TTB was better than that in SSN (Figs. 1 and 2). A previous study examined tone recognition in six-talker Mandarin babbles and compared the performance in white noise (Dees *et al.*, 2007). They showed that the six-talker babbles exerted greater masking effects on tone recognition than the white noise. White noise has less masking to speech signals due to its flat spectrum that is not concentrated in the speech spectrum range. It is not known whether the six-talker babble would have greater masking on tones than the SSN. For the two talker-babble, it seems that our listeners could take advantage of the dips in the spectrotemporal structure of the maskers and glimpse the tone patterns of the targets. Our correlation analyses revealed that the correlation of both E and TFS to the original tone tokens in the TTB conditions were stronger than that in the SSN conditions (Fig. 3; see also Fig. 2 of Apoux *et al.*, 2013).

Our results further shed light on the challenges that tone-language-speaking cochlear implant users face in everyday situations. Current vocoder-based speech processing strategies in multichannel cochlear implants convey the temporal envelope information through the electric pulse trains of a constant rate, whereas the TFS information is discarded. We have shown that Mandarin-speaking children with cochlear implants perform poorly in tone perception with moderate levels of noise (e.g., SNR at 0 dB; Mao and Xu, 2017). The present study demonstrated that the acoustic TFS cues are as important as the E cues for lexical tone perception in noise and competing-speech maskers. However, efforts to deliver TFS information through electric stimulations will still encounter many obstacles such as broad electric field, mismatch of frequency allocation, distant electrode-nerve interface, and uneven spiral ganglion cell distribution and survival (see Zeng *et al.*, 2014).

V. CONCLUSIONS

In summary, in contrast to the results reported for lexical tone perception in quiet and those for English speech

perception in noise, the present study showed that both TFS and E cues contributed to lexical tone perception in noise and competing-speech maskers. The perceptual weights on acoustic TFS and E that the normal-hearing listeners used for lexical tone perception in maskers were by and large equivalent although E cues appeared to exert slightly stronger contributions to tone perception in maskers. Future studies will examine the relative contributions of TFS and E to speech perception (e.g., sentence recognition) in various maskers of tonal languages and compare them to those of non-tonal languages.

ACKNOWLEDGMENTS

Emily Hahn, Zhiwei Mou, Lauren Muscari, and Alexa Patton provided technical assistance. This study was partially supported by grants from the National Institutes of Health National Institute on Deafness and Other Communication Disorders (Grant No. R15-DC014587), National Natural Science Foundation of China (Grant Nos. 81200753 and 81371103), Beijing Natural Science Foundation (Grant No. 7123210), Beijing Nova Program (Grant No. Z151110000315064), and Beijing Health System Healthcare Professionals Career Development Grant (Grant No. 2015-3-019).

- Apoux, F., and Healy, E. W. (2013). "A glimpsing account of the role of temporal fine structure information in speech recognition," *Adv. Exp. Med. Biol.* **787**, 119–126.
- Apoux, F., Yoho, S. E., Youngdahl, C. L., and Healy, E. W. (2013). "Role and relative contribution of temporal envelope and fine structure cues in sentence recognition by normal-hearing listeners," *J. Acoust. Soc. Am.* **134**, 2205–2212.
- Boersma, P., and Weenink, D. (2016). "PRAAT: Doing phonetics by computer (version 6.0.21) [computer program]," <http://www.praat.org> (Last viewed 7 October 2016).
- Cabrera, L., Tsao, F.-M., Gnansia, D., Bertoncini, J., and Lorenzi, C. (2014). "The role of spectro-temporal fine structure cues in lexical-tone discrimination for French and Mandarin listeners," *J. Acoust. Soc. Am.* **136**, 877–882.
- Darwin, C. (2005). "Globally change the pitch and duration using a script for the PRAAT program," http://www.lifesci.sussex.ac.uk/home/Chris_Darwin/Praatscripts/Fodurndynamic (Last viewed 23 April 2017).
- Dees, T. M., Bradlow, A. R., Dhar, S., and Wong, P. C. (2007). "Effects of noise on lexical tone perception by native and non-native listeners," in *Proceedings of 16th ICPHS*, pp. 817–820.
- Dorman, M. F., Loizou, P. C., Fitzke, J., and Tu, Z. (1998). "The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels," *J. Acoust. Soc. Am.* **104**, 3583–3585.
- Dorman, M. F., Loizou, P. C., and 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**, 2403–2411.
- Fogerty, D. (2011). "Perceptual weighting of the envelope and fine structure across frequency bands for sentence intelligibility: Effect of interruption at the syllabic-rate and periodic-rate of speech," *J. Acoust. Soc. Am.* **130**, 489–500.
- Fu, Q.-J., Zeng, F.-G., Shannon, R. V., and Soli, S. D. (1998). "Importance of tonal envelope cues in Chinese speech recognition," *J. Acoust. Soc. Am.* **104**, 505–510.
- Ghitza, O. (2001). "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception," *J. Acoust. Soc. Am.* **110**, 1628–1640.
- Gilbert, G., and Lorenzi, C. (2006). "The ability of listeners to use recovered envelope cues from speech fine structure," *J. Acoust. Soc. Am.* **119**, 2438–2444.
- Glasberg, B. R., and Moore, B. C. J. (1990). "Derivation of auditory filter shapes from notched-noise data," *Hear. Res.* **47**, 103–138.
- Hopkins, K., Moore, B. C. J., and Stone, M. A. (2008). "Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech," *J. Acoust. Soc. Am.* **123**, 1140–1153.
- Kong, Y.-Y., and Zeng, F.-G. (2006). "Temporal and spectral cues in Mandarin tone recognition," *J. Acoust. Soc. Am.* **120**, 2830–2840.
- Krenmayr, A., Qi, B., Liu, B., Liu, H., Chen, X., Han, D., and Zierhofer, C. M. (2011). "Development of a Mandarin tone identification test: Sensitivity index d' as a performance measure for individual tones," *Int. J. Audiol.* **50**, 155–163.
- Lee, C.-Y., Tao, L., and Bond, Z. S. (2013). "Effects of speaker variability and noise on Mandarin tone identification by native and non-native listeners," *Speech Lang. Hear.* **16**(1), 46–54.
- Léger, A. C., Desloge, J. G., Braidia, L. D., and Swaminathan, J. (2015). "The role of recovered envelope cues in the identification of temporal-fine-structure speech for hearing-impaired listeners," *J. Acoust. Soc. Am.* **137**(1), 505–508.
- Li, B., Hou, L., Xu, L., Wang, H., Yang, G., Yin, S., and Feng, Y. (2015). "The effect of steep high-frequency hearing loss on speech recognition using temporal fine structure in the low frequency regions," *Hear. Res.* **326**, 66–74.
- Loizou, P. C. (2006). "Speech processing in vocoder-centric cochlear implants," *Adv. Otorhinolaryngol.* **64**, 109–143.
- Loizou, P. C., Dorman, M., and Tu, Z. (1999). "On the number of channels needed to understand speech," *J. Acoust. Soc. Am.* **106**, 2097–2103.
- Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. J. (2006). "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," *Proc. Natl. Acad. Sci. U.S.A.* **103**, 18866–18869.
- Mao, Y., and Xu, L. (2017). "Lexical tone recognition in noise in normal-hearing children and prelingually deafened children with cochlear implants," *Int. J. Audiol.*, in press.
- Moon, I. J., Won, J. H., Park, M.-H., Ives, D. T., Nie, K., Heinz, M. G., Lorenzi, C., and Rubinstein, J. T. (2014). "Optimal combination of neural temporal envelope and fine structure cues to explain speech identification in background noise," *J. Neurosci.* **34**(36), 12145–12154.
- Moore, B. C. J. (2008). "The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people," *J. Assoc. Res. Otolaryngol.* **9**, 399–406.
- Nambi, P. M. A., Mahajan, Y., Francis, N., and Bhat, J. S. (2016). "Temporal fine structure mediated recognition of speech in the presence of multitalker babble," *J. Acoust. Soc. Am.* **140**, EL296–EL301.
- Nilsson, M., Soli, S. D., and 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**, 1085–1099.
- Shamma, S., and Lorenzi, C. (2013). "On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system," *J. Acoust. Soc. Am.* **133**, 2818–2833.
- Shannon, R. V., Zeng, F.-G., Kamath, V., Wyganski, J., and Ekelid, M. (1995). "Speech recognition with primarily temporal cues," *Science* **270**, 303–304.
- Smith, Z. M., Delgutte, B., and Oxenham, A. J. (2002). "Chimaeric sounds reveal dichotomies in auditory perception," *Nature* **416**, 87–90.
- Swaminathan, J., Reed, C. M., Desloge, J. G., Braidia, L. D., and Delhorne, L. A. (2014). "Consonant identification using temporal fine structure and recovered envelope cues," *J. Acoust. Soc. Am.* **135**(4), 2078–2090.
- Wang, S., Dong, R., Liu, D., Wang, Y., Liu, B., Zhang, L., and Xu, L. (2015). "The role of temporal envelope and fine structure in Mandarin lexical tone perception in auditory neuropathy spectrum disorder," *PLoS One* **10**(6), e0129710.
- Wang, S., Dong, R., Liu, D., Wang, Y., Mao, Y., Zhang, H., and Xu, L. (2016). "Perceptual separation of sensorineural hearing loss and auditory neuropathy spectrum disorder," *Laryngoscope* **126**, 1420–1425.
- Wang, S., Xu, L., and Mannell, R. (2011a). "Relative contributions of temporal envelope and fine structure cues to lexical tone recognition in hearing-impaired listeners," *J. Assoc. Res. Otolaryngol.* **12**, 783–794.
- Wang, W., Zhou, N., and Xu, L. (2011b). "Musical pitch and lexical tone perception with cochlear implants," *Int. J. Audiol.* **50**, 270–278.
- Warton, D. I., and Hui, F. K. (2011). "The arcsine is asinine: The analysis of proportions in ecology," *Ecology* **92**, 3–10.
- Wilson, B. S., and Dorman, M. F. (2008). "Cochlear implants: A remarkable past and a brilliant future," *Hear. Res.* **242**, 3–21.

- Xu, L. (2016). "Temporal envelopes in sine-wave speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech 2016)*, San Francisco, CA, pp. 1682–1686.
- Xu, L., and Pfingst, B. E. (2003). "Relative importance of temporal envelope and fine structure in lexical-tone perception (L)," *J. Acoust. Soc. Am.* **114**, 3024–3027.
- Xu, L., and Pfingst, B. E. (2008). "Spectral and temporal cues for speech recognition: Implications for auditory prostheses," *Hear. Res.* **242**, 132–140.
- Xu, L., Tsai, Y., and Pfingst, B. E. (2002). "Features of stimulation affecting tonal-speech perception: Implications for cochlear prostheses," *J. Acoust. Soc. Am.* **112**, 247–258.
- Xu, L., and Zheng, Y. (2007). "Spectral and temporal cues for phoneme recognition in noise," *J. Acoust. Soc. Am.* **122**, 1758–1764.
- Xu, L., and Zhou, N. (2011). "Tonal languages and cochlear implants," in *Auditory Prostheses: New Horizons*, edited by F.-G. Zeng, A. N. Popper, and R. R. Fay (Springer Science and Business Media, New York), pp. 341–364.
- Zeng, F.-G., Nie, K., Liu, S., Stickney, G. S., Del Rio, E., Kong, Y.-Y., and Chen, H. (2004). "On the dichotomy in auditory perception between temporal envelope and fine structure cues," *J. Acoust. Soc. Am.* **116**, 1351–1354.
- Zeng, F.-G., Tang, Q., and Lu, T. (2014). "Abnormal pitch perception produced by cochlear implant stimulation," *PLoS One* **9**, e88662.