The effect of speech material on the band importance function for Mandarin Chinese

Yufan Du,^{1,a)} Yi Shen,² Xihong Wu,^{1,a)} and Jing Chen^{1,a),b)}

¹Department of Machine Intelligence, Peking University, Beijing, China

²Department of Speech and Hearing Sciences, Indiana University Bloomington, 200 South Jordan Avenue, Bloomington, Indiana 47405, USA

(Received 23 January 2019; revised 23 May 2019; accepted 25 June 2019; published online 25 July 2019)

Speech material influences the relative contributions of different frequency regions to intelligibility for English. In the current study, whether a similar effect of speech material is present for Mandarin Chinese was investigated. Speech recognition was measured using three speech materials in Mandarin, including disyllabic words, nonsense sentences, and meaningful sentences. These materials differed from one another in terms of the amount of contextual information and word frequency. The band importance function (BIF), as defined under the Speech Intelligibility Index (SII) framework, was used to quantify the contributions across frequency regions. The BIFs for the three speech materials were estimated from 16 adults who were native speakers of Mandarin. A Bayesian adaptive procedure was used to efficiently estimate the octave-frequency BIFs for the three materials for each listener. As the amount of contextual information increased, low-frequency bands (e.g., 250 and 500 Hz) became more important for speech recognition, consistent with English. The BIF was flatter for Mandarin than for comparable English speech materials. Introducing the language- and material-specific BIFs to the SII model led to improved predictions of Mandarin speech-recognition performance. Results suggested the necessity of developing material-specific BIFs for Mandarin. © 2019 Acoustical Society of America. <https://doi.org/10.1121/1.5116691>

[AKCL]

Pages: 445–457

I. INTRODUCTION

Performance in recognizing target speech in background noise depends on the type of speech material (e.g., monosyllabic words, spondee words, meaningful sentences, etc.) used. The effect of speech material on speech recognition has been studied extensively for English (Miller *et al.*, 1951; Hirsh *et al.*, 1954), and many speech recognition models include material-specific parameters. Taking the Speech Intelligibility Index (SII; ANSI, 1997) as its theoretical framework, the current study examines the effect of speech material for Mandarin Chinese. Of particular interest is whether the contributions from various frequency regions for speech intelligibility vary as the test material changes. The potential benefits of including material- and language-specific parameters in the SII model are also evaluated.

The SII (ANSI, 1997), formerly known as the Articulation Index (AI; ANSI, 1969), is an established model of speech recognition. The SII takes a value between 0 and 1, representing the proportion of speech sound audible to the listener. It is given by

$$\text{SII} = \sum_i w_i A_i, \quad (1)$$

where A_i indicates speech audibility in the i th frequency band, and w_i indicates the relative weight for the i th band.

The value of A_i is between 0 and 1, and it is determined by whether the speech is presented above the absolute threshold and, in the presence of background noise, the signal-to-noise ratio (SNR). The function that describes how w_i varies across frequency bands is often referred to as the band importance function (BIF).

To predict performance for speech recognition tests commonly used in clinical audiology and hearing research, a transfer function is used to relate the SII to the proportion correct (PC) in word recognition. The transfer function is a monotonically increasing function with a sigmoidal shape. Traditionally, the transfer function is formulated as (Fletcher and Galt, 1950)

$$\text{PC} = [1 - 10^{-\text{SII}/Q}]^N, \quad (2)$$

where N and Q are two fitting parameters, while variations of the formulation have been explored by some researchers (e.g., Ching *et al.*, 1998; Boothroyd, 2008). Regardless of the specific formulations of the transfer function, the shape of the function can be quantified using a threshold and a slope parameter. The threshold of the transfer function is the SII necessary for recognizing 50% of the test words, while the slope indicates how rapidly the performance grows with increasing SII.

When applying the SII model for a speech recognition test, it is often necessary to adjust both the BIF and transfer function depending on the type of test material (e.g., monosyllabic words, disyllabic words, sentences, etc.). For English test materials, the effects of test material on both the

^{a)}Also at: Speech and Hearing Research Center and Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China

^{b)}Electronic mail: chenjing@cis.pku.edu.cn

BIF and transfer function have been extensively studied in the past. The BIF for nonsense syllables typically contains one prominent peak near 2 kHz, and the spectral weight diminishes near the lowest and highest ends of the spectrum (e.g., French and Steinberg, 1947). On the other hand, the BIFs for monosyllabic words and words in low-context sentences are broader compared to those for nonsense syllables with a secondary peak appearing near 400–500 Hz (Studebaker and Sherbecoe, 1991; Bell *et al.*, 1992; Studebaker *et al.*, 1993). Finally, for words in high-context sentences and continuous discourse, the BIFs are even broader with a pronounced bimodal configuration observed in some studies (e.g., Studebaker *et al.*, 1987; Bell *et al.*, 1992; Sherbecoe and Studebaker, 2002; Lee and Mendel, 2017; Yoho *et al.*, 2018). It has been suggested that the broadening and bimodal configuration of the BIF are associated with the increased redundancies and contextual cues in the test materials (Studebaker and Sherbecoe, 1991; DePaolis *et al.*, 1996). To accommodate the effect of speech material, six BIFs for different English speech materials are provided in the ANSI-3.5 (ANSI, 1997) standard, including nonsense syllables (Fletcher and Steinberg, 1929; French and Steinberg, 1947; Pavlovic and Studebaker, 1984), Central Institute for the Deaf (CID) W-22 word lists (Studebaker and Sherbecoe, 1991), monosyllables of the NU-6 test (Studebaker *et al.*, 1993), Diagnostic Rhyme Test (DRT) material (Duggirala *et al.*, 1988), short passage of easy reading material (Studebaker *et al.*, 1987), and monosyllables of the Speech Perception in Noise (SPIN) test (Bell *et al.*, 1992).

Besides the BIF, the effect of speech material on the shape of the transfer function is also a well-known phenomenon (e.g., Fletcher and Steinberg, 1929; Miller *et al.*, 1951; Mueller and Killion, 1990). A consistent finding from numerous studies is that the threshold of the transfer function decreases and the slope of the function becomes steeper as the amount of contextual cues increases in the test material (e.g., Kalikow *et al.*, 1977; Duggirala *et al.*, 1988; van Rooij and Plomp, 1991; Bell *et al.*, 1992; DePaolis *et al.*, 1996; Olsen *et al.*, 1997), and this effect of context has been successfully modeled using principles of probability (Boothroyd, 1978; Boothroyd and Nittroer, 1988; Bronkhorst *et al.*, 1993; Bronkhorst *et al.*, 2002). Other factors, such as word-frequency, may also affect the transfer function (e.g., Hirsh *et al.*, 1954; Howes, 1957; Savin, 1963). The word-frequency effect is related to the context effect described earlier in that it is harder to identify a word based on only a subset of its phonemes if it is an uncommon word. This is likely due to listeners' response biases toward the more frequently occurring alternative words (e.g., Broadbent, 1967; Benkí, 2003).

The SII model has been adopted to predict speech intelligibility for other languages (e.g., (Hochmuth *et al.*, 2015). A number of researchers have suggested the necessity of developing language-specific SII models to better account for the differences in acoustic properties and phoneme distribution across languages (e.g., Wong *et al.*, 2007a; Jin *et al.*, 2017). Consequently, language-specific SII models have been developed for several languages (Wong *et al.*, 2007a; Lee and Kim, 2012; Kuo, 2013; Chen *et al.*, 2016; Jin *et al.*, 2016; Narne *et al.*, 2016). Moreover, the effect of test

material on the BIF and transfer function of the SII model has been extended to other languages. Similar to English, it has been shown that greater redundancies in the test materials (close-set tests or sentences) lead to a steeper rate at which the recognition score increases with speech audibility (i.e., a steeper transfer function) compared to test materials with less contextual cues (e.g., isolated words) for a number of languages, including Dutch (Smits *et al.*, 2013), French (Jansen *et al.*, 2012), Mandarin (Kuo, 2013), Swedish (Magnusson, 1996), etc. Following these previous studies, the current study aims to provide further clarifications on the effect of speech material on the SII parameters, i.e., the BIF and transfer function, for Mandarin Chinese.

The BIFs for Chinese (Mandarin or Cantonese) has some noticeable differences from those for English. Wong *et al.* (2007a) showed that low frequencies (<400 Hz) are more important for Cantonese sentence recognition than English. Similar high low-frequency weights have also been shown in the BIFs for Mandarin materials (Kuo, 2013; Chen *et al.*, 2016). This may be caused by a number of factors. First, Chinese is a tonal language in which the same syllable can carry different meanings by using different lexical tones, and the tone perception primarily depends on the fundamental frequency (F_0) contour (Lin and Wang, 1992). In contrast, for English, the F_0 contour may impact the sentence intonation, but it does not change the meaning of the word. Second, the syllable structures of Chinese and English are different. Chinese syllables usually have a consonant-vowel (CV) structure, while English syllables allow consonant clusters that appear in onset and coda (Chen *et al.*, 2013). It has also been reported that vowels may have a greater contribution to sentence intelligibility in Mandarin Chinese than in English (Chen *et al.*, 2013). Third, in contemporary Mandarin Chinese, most words are two-character compound words in which each of the two characters has its own semantic representation. In English, the proportion of compound words is much less than that in Chinese for normal speech tests, and listeners generally do not have to access the meaning of the individual syllables to recognize the whole word. Hence, the manner in which the meaning of words is accessed may differ substantially across the two languages.

Kuo (2013) measured the BIFs and transfer functions for Mandarin monosyllabic words and meaningful sentences. The results showed that additional contextual cues in sentences led to a downward shift in the spectral weights in the BIF and a steeper slope for the transfer function compared to words, similar to the results for English. The current study extended this previous result to additional types of speech materials, including disyllable words and nonsense sentences. Additionally, the BIFs and transfer functions for various speech materials were estimated for individual listeners. This within-subject experimental design is advantageous over the approach taken by many previous studies, where data from multiple listeners are pooled together, in that it isolated the effect of speech material from that of individual differences.

One factor that hinders the studies of the SII parameters is the time-consuming process involved in estimating the

BIFs. Traditionally, the shape of the BIF, i.e., the relative importance of different frequency regions, is estimated by systematically measuring speech-recognition performance under various low- and high-pass filtering conditions (French and Steinberg, 1947; Studebaker and Sherbecoe, 1991). Chen *et al.* (2016) adopted the traditional procedure in their study of Mandarin monosyllabic words, and it took more than 20 hours of testing for each estimate of the BIF from a single listener. Besides the long testing time required, the traditional procedure also relies on the assumption that the contributions of various speech frequency bands are independent. The validity of this assumption may depend on the speech materials and specific speech-filtering techniques used (Warren *et al.*, 1995; Healy and Warren, 2003; Healy and Bacon, 2007; Humes and Kidd, 2016). To address this issue, a number of recent studies have recommended to present speech stimuli using a “compound” approach (Apoux and Healy, 2012) rather than the traditional low- and high-pass filtering approach. The compound approach involves filtering broadband speech into a number of frequency bands and randomly selecting a subset of the bands to present the speech stimulus on every trial (Healy *et al.*, 2013; Bosen and Chatterjee, 2016; Shen and Kern, 2018).

In the current study, a Bayesian adaptive procedure, namely the quick-band-importance-function (qBIF) procedure (Shen and Kern, 2018), was used to enable efficient estimation of the BIFs and transfer functions. The qBIF procedure presents speech in background noise using the compound approach. The stimulus on a given trial, determined by the SNR and the bands selected to present the stimuli (speech bands), is optimized based on previously collected data so that the expected information gain about the parameters of the SII model (including the transfer function and BIF) is maximized. The efficiency and reliability of the qBIF procedure have been demonstrated for both English (Shen and Kern, 2018) and Mandarin (Du *et al.*, 2018) using monosyllabic words. The testing time required for estimating each octave-frequency BIF is typically less than 30 minutes. In the current study, the qBIF procedure was used to enable the estimation of the BIFs for three different Mandarin speech materials from each listener. Given that the qBIF procedure is a relatively recent method, the current study included repeated qBIF runs to assess the test-retest reliability of the procedure. This was to ensure that any observed differences among the speech materials were not due to measurement errors.

Understanding the effect of speech material for Mandarin not only has theoretical values, it is also likely to impact clinical practice in audiology for the Mandarin-speaking patient population. As one important application of the SII is to evaluate the efficiency of hearing-aid fitting, the need for a version of the SII model that is validated for Mandarin becomes prominent with the increase of hearing-aids users in China (Zhang, 2009). Moreover, standard hearing-aid prescription procedures make use of the SII predictions (Moore and Glasberg, 1998; Keidser *et al.*, 2011), but the BIFs in the standard SII model are based on English and it may not be suitable for Mandarin-speaking listeners. The current study provides estimates of the BIFs for three

Mandarin speech materials that have been widely used in clinical practice. The necessity of using language- and material-specific BIFs to predict speech recognition performance for Mandarin is evaluated.

II. METHODS

A. Participants

Sixteen native Mandarin Chinese speakers (18–27 yr old, mean = 22 yr) participated in this study. Five of them were male. All listeners were naive to the current experiment. They all had pure-tone thresholds ≤ 25 dB hearing level (HL) at octave frequencies from 250 to 8000 Hz in both ears, and all were paid for their participation.

B. Stimuli

Mandarin disyllable words, nonsense sentences, and meaningful sentences were included as the test materials in the current study. The disyllable words and meaningful sentences were from the Mandarin Speech Test Materials (MSTMs) corpus developed at Tong’ren Hospital (Zhang *et al.*, 2008), and they have been widely used for speech audiometry in China. The nonsense sentences were selected from a corpus developed by Yang *et al.* (2007). This speech corpus has been used in many previous works on speech perception of Mandarin Chinese (Huang *et al.*, 2008; Du *et al.*, 2011; Chen *et al.*, 2012; Du *et al.*, 2018).

In Mandarin Chinese, the smallest meaningful unit is syllable, with each syllable corresponding to a Chinese character in text. The disyllabic words are two-character compound words in which each of the two Chinese characters has its own semantic representation. For example, the Chinese word for “Beijing” is a disyllabic (/Bei3/ and /Jing1/) compound word, in which each syllable has a specific referent (the character pronounced as /Bei3/ means “north,” the character pronounced as /Jing1/ means “capital city”). According to the data analysis based on People’s Daily published between 1994 and 2002, there are 31 disyllabic words that begin with “/Bei3/,” and 119 disyllabic words that begin with the syllable “/Bei/,” whose tone can be any one of the four lexical tones. On the other hand, there are 45 disyllable words that end with “/Jing1/,” and 189 disyllable words that end with the syllable “/Jing/,” whose tone can be any one of the four (Yang *et al.*, 2007). Because there are many possible combinations beginning with “/Bei/” or ending in “/Jing/,” a Chinese listener may have to access the meaning of each of the syllables (corresponding to one character) in order to recognize the whole word. Hence, the scoring of Chinese sentences was based on syllables’ recognition in previous studies (Wong *et al.*, 2007a; Kuo, 2013; Chen *et al.*, 2016), and we adopted the same scoring method.

For the disyllable words stimuli, there were eight disyllabic word lists in MSTMs (Zhang *et al.*, 2008), and each list contained 50 disyllabic words. These disyllabic words were spoken by a male talker (mean fundamental frequency = 142.52 Hz). For the meaningful sentences stimuli, there were 15 sentence lists in MSTMs, also spoken by the same

male talker, and each list contained 10 meaningful sentences. Each meaningful sentence had 2–12 characters (i.e., syllables), including 2–8 key characters. The nonsense sentences stimuli, developed by [Yang et al. \(2007\)](#), were spoken by a female talker (mean fundamental frequency = 244.09 Hz). Each nonsense sentence had three key words: subject, predicate, and object, and each word was disyllabic, containing two syllables. Therefore, there were six key characters in a nonsense sentence. Comparing to the meaningful sentences, the nonsense sentences did not provide any semantic context for recognizing the key words. Sixteen lists of nonsense sentences were selected from the original corpus, and each lists contained 18 nonsense sentences. All speech signals were digitalized at a 16-kHz sampling rate.

The target speech was presented with a simultaneous masker. The masker signal was a steady speech-spectrum noise (SSN), which was a broadband noise spectrally shaped to match the long-term average spectrum of the talker. Two SSNs were generated separately for the MSTM materials (disyllabic words and meaningful sentences spoken by the male talker) and the nonsense sentences (spoken by the female talker). The talker-specific SSN noise was used to ensure the SNR was relatively constant across frequency. The target speech and SSN were mixed together with the SSN gated on 0.5 s before the speech and gated off 0.5 s after the speech. The resulting mixture stimulus was passed through a bank of six octave-frequency filters. The center frequencies for the filters were 250, 500, 1000, 2000, 4000, and 8000 Hz. Each of the filters was constructed as a 12th-order Butterworth filter. The mixture stimulus in a subset of the six bands (i.e., the speech bands) was recombined and presented to the listener. The SNR and speech bands for each trial were adaptively determined according to the qBIF procedure.

C. Apparatus

The current experiment was conducted in an anechoic chamber, which was 5.6 m in length, 4 m in width, and 1.93 m in height. Stimuli were presented via a soundcard (RME Fireface UC, Haimhausen, Germany) and a loudspeaker (Dynaudio Acoustics, BM64, Skanderborg, Denmark). During the experiment, the listener was seated in front of the loudspeaker. The distance between the loudspeaker and the center of listener's head was about 2.00 m. The height of the loudspeaker and the listener's ear was approximately the same. The level of the original speech stimuli, before mixing with the SSN and filtering, was calibrated at 65 dBA at the subjects' head position using a sound meter (model 824, Larson Davis, Inc., Depew, NY).

D. Design and procedure

The experiment was divided into three phases, corresponding to the three types of speech materials. For each of the listeners, the disyllabic words were tested in the first phase, the meaningful sentences were tested in the second phase, and the nonsense sentences were tested in the final phase. During the first phase, the estimation of the BIF was repeated using two runs of the qBIF procedure (i.e., a test and a retest). Four of the disyllabic word lists (200 words)

were randomly selected and used for the first qBIF run, while the remaining four lists were used for the second qBIF run. During the second phase, a single qBIF run (150 sentences) was conducted because the corpus for the meaningful sentences contained a relatively small number of unique tokens. The third phase consisted of a test and a retest qBIF run using nonsense sentences. Eight of the nonsense sentence lists (144 sentences) were randomly selected for the first qBIF run, while the remaining 8 lists were used for the second qBIF run. Although each qBIF run included an initial stage that allowed the listener to get familiarized with the stimuli, it was decided that additional practice trials were necessary before the two qBIF runs. These practice trials included nonsense sentences from the same speech corpus as the test sentences but were not used in the qBIF runs. Overall, each qBIF run took approximately 20, 30, and 30 min to complete for the disyllabic words, nonsense sentences, and meaningful sentences, respectively. Therefore, the three phases of the experiment took 40, 30, and 60 min, in that order, leading to a total of 130 min of data collection. The listeners were encouraged to take breaks between qBIF runs.

The qBIF procedure was implemented as custom software in the MATLAB (The MathWorks, Natick, MA) environment, which also controlled stimulus presentation. During the experiment, listeners responded to each stimulus by verbally repeating back the presented speech. The experimenter monitored the listener's response outside the anechoic chamber via a talkback microphone system. The correctness of each character (i.e., syllable) in the target speech was separately scored, and the experimenter entered the total number of characters in the target speech and the number of characters correctly repeated by the listener into the MATLAB software. The software then computed the appropriate stimulus parameters (i.e., SNR and speech bands) for the next trial. The responses were scored online during the experiment because the qBIF procedure, being an adaptive procedure, required the recognition scores from each trial to derive the stimulus parameters for the next trial. To limit potential variations in online scoring, data collection from all listeners and in all conditions was conducted by the same experimenter (i.e., the first author).

Since the algorithm for stimulus optimization has been described by [Shen and Kern \(2018\)](#) and [Du et al. \(2018\)](#) in detail, only a brief description is provided here. In the current implementation of the qBIF procedure, the possible values for SNR were −5, 0, 5, 10, and 15 dB. The speech bands could be any combinations of two, three, four, or five frequency bands (a total of 56 combinations). Each qBIF run had two stages. During the first stage, the number of speech bands was always five, and the SNRs were selected according to a pre-determined schedule that started from high SNRs (5 dB for disyllabic words and meaningful sentences, and 15 dB for nonsense sentences) and gradually increased the test difficulty. The initial SNRs were chosen according to pilot studies. In pilot studies, we found that under the 5-dB SNR and five-speech-bands condition, most subjects can achieve 65% percent correct or better for disyllabic words and meaningful sentences, therefore this condition was considered relatively easy and adequate for familiarizing the listener with the stimuli. On the other hand, the recognition

of nonsense sentences was more difficult, and the performance was consistently lower than 65% correct under the 5-dB SNR and five-speech-bands condition, therefore this condition was not suitable for initial familiarization. Consequently, we set the initial SNR to be 15 dB to ensure sufficiently high performance for nonsense sentences at the beginning of a qBIF run. This initial stage aimed to familiarize the listener with the test material while accumulating sufficient data to provide an initial fit of the SII model. When the overall percentage of key character recognition dropped below 65% or 50 trials were tested (whichever occurred first), the first stage terminated and the qBIF procedure moved on to the second stage. During the second stage, the SNR and speech bands were optimized iteratively according to an entropy-based criterion. This criterion maximized the expected information gain (i.e., entropy loss) about the SII parameters (e.g., the transfer function and BIF) from the next trial (Kontsevich and Tyler, 1999; Shen and Richards, 2013; Shen *et al.*, 2014).

III. RESULTS

A. The effect of speech material on the BIF

Figure 1 shows the average BIFs across the 16 listeners. The average octave-frequency spectral weights in the BIFs are summarized in Table I. For Fig. 1 and Table I, before averaging the spectral weights across listeners, the results from the two qBIF runs were first averaged together for disyllabic words and nonsense sentences. The average BIF for disyllabic words (top left panel) had the highest weight at the 2000-Hz band, which was consistent with previous studies for Mandarin monosyllabic words (see diamonds in Fig. 1; data from Du *et al.*, 2018). The BIFs for nonsense

FIG. 1. The average octave-frequency BIFs across 16 listeners estimated using the qBIF procedure for Mandarin disyllabic words (top left), nonsense sentences (top right), and meaningful sentences (bottom left). Error bars indicate $\pm$ one standard error of the mean. In each panel, the result for Mandarin monosyllabic words estimated by Du *et al.* (2018) using the qBIF procedure is plotted (diamonds) as a reference.

TABLE I. Average BIF values and standard errors (in parentheses) obtained in this study for Mandarin Chinese disyllabic words, nonsense sentences, and meaningful sentences.

Center frequency(Hz)	Band weights		
	Disyllabic words	Nonsense sentences	Meaningful sentences
250	0.1036 (0.0076)	0.1116 (0.0074)	0.1699 (0.0056)
500	0.2202 (0.0058)	0.2975 (0.0084)	0.2170 (0.0067)
1000	0.2188 (0.0067)	0.2230 (0.0080)	0.1869 (0.0075)
2000	0.2625 (0.0055)	0.2040 (0.0084)	0.1954 (0.0060)
4000	0.1379 (0.0101)	0.0986 (0.0093)	0.1390 (0.0070)
8000	0.0570 (0.0091)	0.0653 (0.0087)	0.0918 (0.0011)

sentences (top right panel) and meaningful sentences (bottom left panel) had the highest weight at 500 Hz. The BIF for meaningful sentences exhibited a higher weight than the other two speech materials at 250 Hz, while the BIF for nonsense sentences exhibited a higher weight than the other two materials at 500 Hz. The spectral centroids of the average BIFs, on a logarithmic frequency scale, for disyllabic words, nonsense sentences, and meaningful sentences were 1216, 1054, and 1142 Hz, respectively. These were lower than the spectral centroid of 1310 Hz measured by Du *et al.* (2018) using Mandarin monosyllabic words.

A repeated-measures analysis of variance (ANOVA) was conducted to investigate the effect of frequency band and speech material on the BIF. Kolmogorov-Smirnov tests confirmed the normality of the weights for each of the six frequency bands ($p > 0.1$). Moreover, since the weights across frequency bands sum to unity, it is expected that the effect of speech material on the BIF would be reflected as interactions between frequency band and speech material rather than as a main effect of speech material. To address the inherently inter-related weights across frequencies, weights from the lowest- and highest-frequency bands (250 and 8000 Hz) were excluded from the ANOVA analysis. There were significant main effects of frequency band [$F(3,45) = 109.97$, $p < 0.01$, $\eta^2 = 0.88$] and speech material [$F(2,30) = 21.88$, $p < 0.01$, $\eta^2 = 0.59$]. There was also a significant interaction between frequency band and speech material [$F(6,90) = 18.43$, $p < 0.01$, $\eta^2 = 0.55$]. The significant interaction between frequency band and speech material indicates that the shape of the BIF depends on the type of speech material used.

B. The effect of speech material on the transfer function

The transfer function is a monotonic function describing how speech-recognition performance improves with increasing SII. Following Shen and Kern (2018) and Du *et al.* (2018), the transfer function was modeled using a logistic function

TABLE II. The threshold and slope parameters of the transfer function [see Eq. (1)] averaged across subjects for the three speech materials and their standard errors in parentheses.

	Disyllabic words	Nonsense sentences	Meaningful sentences
Threshold parameter θ	0.2829 (0.0050)	0.4216 (0.0110)	0.2186 (0.0039)
Slope parameter β	2.8047 (0.0877)	1.5968 (0.0521)	4.4152 (0.2252)

$$PC = [1 + e^{-4\beta(SII-\theta)}]^{-1}, \quad (3)$$

where PC is the proportion of correct key character recognition. θ is the threshold parameter of the transfer function, and it corresponds to the SII value associated with 50% key character recognition. A higher threshold indicates that a higher proportion of speech is required to be audible to support successful speech recognition. β is the slope at θ . As the value of β increases, PC increases with SII more rapidly.

Table II lists the threshold and slope parameters of the transfer function averaged across listeners for the three speech materials. The transfer functions for the three materials are also plotted as dashed curves in the three panels of Fig. 2. The transfer function for meaningful sentences was steeper than disyllabic words, which is consistent with previous studies mentioned above (Bell *et al.*, 1992; DePaolis *et al.*, 1996). Two repeated-measures ANOVAs were conducted to investigate the effect of speech material on the threshold and slope parameters. There was a significant main effect of speech material [$F(2,30) = 10.59, p < 0.01, \eta^2 = 0.41$] on the threshold parameter. *Post hoc* Tukey-Kramer comparisons indicated that the threshold parameters of the three speech materials were significantly different from one another ($p < 0.01$). There was also a significant main effect of speech material [$F(2,30) = 84.50, p < 0.01, \eta^2 = 0.85$] on the slope parameter. *Post hoc* Tukey-Kramer comparisons indicated that the slope parameters of the three speech materials were significantly different from one another ($p < 0.01$).

C. The suitability of the simplified SII model in the qBIF procedure

The qBIF procedure was used in the current study to efficiently estimate the SII parameters for multiple speech

materials from the same listener. Since the qBIF procedure was developed relatively recently (Shen and Kern, 2018), extensive validations of the procedure have not yet been available in the literature. There are a number of potential limitations of the qBIF procedure. First, in the current implementation of the qBIF procedure, only octave-BIF can be derived. The coarse spectral representation in this much-simplified model may undermine the ability of the estimated SII model to generate predictions on speech recognition performance. Second, a logistic function [Eq. (3)] is implemented in the qBIF procedure for its mathematical simplicity, however, it deviates from the formulation of the transfer function used traditionally for English [Eq. (2)]. The effects of the specific formulation of the transfer function on the predictive capability of the estimated SII model across a range of speech materials have not been investigated for Chinese.

In order to demonstrate the predictability of the simplified SII model as implemented in the qBIF procedure, the relationship of subjective scores and predicted SII using this simplified SII model was analyzed. Specifically, for each of the 3 speech materials, the BIFs estimated in this study were averaged across the 16 listeners, and the average BIF (Table I) was used to compute the SII values for all possible test stimuli (i.e., all combinations of speech bands and SNR). The predicted scores were then obtained for these test stimuli by passing the computed SII values through the transfer function with the average threshold and slope parameters (Table II). On the other hand, the subjective score for each of the possible test stimuli was calculated by pooling the trial-by-trial data tested using the stimulus across all 16 listeners. In the case that a certain stimulus was not sampled during the experiment for any of the listeners, the stimulus was excluded from further analysis. Finally, the proportion of variance in the subjective scores accounted for by the predictive scores (R^2) was calculated as an indicator of whether the simplified SII model adopted in the current study is suitable to explain the experimental results.

Figure 2 plots the subjective scores against computed SII values as scatter plots, separately for the three speech materials. In each panel, every circle represents a test stimulus (i.e., a combination of speech bands and SNR). The dashed curve represents the transfer function with the average threshold and slope parameters. The R^2 values are 0.63, 0.80, and 0.81 for disyllabic words, nonsense sentences, and

FIG. 2. The left, middle, and right panels show the relationship between the computed SII values and the subjective scores for disyllabic words, nonsense sentences, and meaningful sentences, respectively. The dashed curve in each panel represents the transfer function with the average threshold and slope parameters.

meaningful sentences, respectively. Therefore, the predictive capability of this simplified SII model seemed to be acceptable, and the simplified model provided better fits to the subjective scores for the two sentence materials compared to disyllabic words.

D. The reliability of the estimated BIFs

To demonstrate the test-retest reliability and convergence of the qBIF procedure, the results from the repeated qBIF runs were compared. The top two panels of Fig. 3 show the average BIFs derived from the test and retest for disyllabic words (left) and nonsense sentences (right). Error bars represent $\pm$ one standard error across listeners. The BIFs obtained from the test and retest were highly correlated (for disyllabic words: $r = 0.99$, $p < 0.01$; for nonsense sentences: $r = 0.98$, $p < 0.01$). These results indicate high test-retest reliability at the group level (for a group of 16 subjects). We used the root-mean-square deviation (RMSD) to evaluate the distance between the results of the test and retest. For disyllabic words and nonsense sentences, the RMSD between the average test and average retest BIFs was 0.0133 and 0.0179, respectively. The bottom two panels of Fig. 3 show the individual and average RMSD as functions of the trial number for disyllabic words (left) and nonsense sentences (right). As described in Sec. II, the qBIF procedure has an initial stage during which the estimates of the spectral weights are not available. We set the RMSD to zero for these initial trials. It can be observed that the RMSD decreased as the trial number increased, and it was nearly stable at the

end of the experiment. The average RMSD for nonsense sentences increased at the beginning because subjects did not finish the initial stage at the same trial number. The average test-retest RMSD across all subjects was 0.0613 for disyllabic words and 0.0605 for nonsense sentences at the last trial. These results indicate the appropriate convergence of the qBIF procedure.

E. The benefits of material-specific BIFs

Although the BIFs derived for different speech materials show different shapes (see Fig. 1), it is not immediately clear that these differences in the BIFs would lead to significantly different predictions on speech intelligibility. A cross-validation analysis was conducted to evaluate the potential benefits of material-specific BIFs. In the analysis, the 16 listeners were divided into a training set of 12 listeners and a test set of 4 listeners. The average BIFs from the training set were used to predict the subjective scores from the test set. This procedure was repeated 16 times, corresponding to 16 different ways of dividing the listeners into the training and test sets. Across the 16 repetitions, each listener was assigned to the training set 12 times and the test set 4 times (i.e., the test set included listeners 1–4 during the first repetition, listeners 2–5 during the second repetition, etc.).

Figure 4 summarizes the computational steps involved in each repetition of the cross-validation analysis. First, for the training set, the average BIFs across the listeners were calculated for the three speech materials. For each of these speech materials (i.e., the training materials), the SII values

FIG. 3. The top two panels show the average BIFs derived from the test and retest for disyllabic words and nonsense sentences. The bottom two panels show the individual and average root-mean-square deviation (RMSD) as functions of trial number for disyllabic words and nonsense sentences experiment.

FIG. 4. The computation procedure within each repetition of the cross-validation analysis to investigate the potential benefit of material-specific BIFs.

for all possible stimuli (i.e., combinations of speech bands and SNR) were calculated. Second, for the test set, the trial-by-trial stimuli and responses from the test set were pooled together for each of the three speech materials (i.e., the test materials). Third, the pooled data from the test set were predicted by the computed SII values using either the matching speech material or mismatch speech materials. This led to a total of nine combinations of the training and test materials. For each of the nine combinations of the training and test materials, a logistic regression was performed. For each key character (i.e., syllable) presented on a given trial in the test set, the predictor in the logistic regression was the computed SII values for the speech bands and SNR on that trial. The dependent variable was the correctness in recognizing the key character. To evaluate the predictive capability of the SII values (hence, the corresponding average BIF), the fitted logistic model was used to generate the predicted scores for all sampled stimuli in the test set. The average recognition scores for these stimuli formed the subjective scores, and the proportion of variance accounted for by the logistic model (R^2) was calculated, similar to the analysis presented in Sec. III C and Fig. 2. The resulting R^2 value reflected how well the model would explain an independent set of speech-recognition data. A higher R^2 value was used as an indicator for greater predictive power. Following the 16 repetitions of the cross-validation procedure, 16 estimates of R^2 were obtained for each combination of the training and test materials.

The mean and standard error of the R^2 value are plotted in Fig. 5. Wilcoxon signed rank tests were conducted to compare the R^2 among different BIFs. When nonsense sentences were the test material, the performance of the SII model with the BIF for the matching training material was better than those for the mismatched training materials (compared to the BIF for disyllabic words: $p < 0.01$, Cohen's $d = 3.02$; compared to the BIF for meaningful sentences: $p < 0.01$, Cohen's $d = 3.22$). When meaningful sentences were the test material, the performance of the SII model with the BIF for the matching training material was also better

than those for the mismatched training materials (compared to the BIF for disyllabic words: $p < 0.01$, Cohen's $d = 5.74$; compared to the BIF for nonsense sentences: $p < 0.01$, Cohen's $d = 4.61$). These results show that using the material-specific BIFs could improve the predictions on speech-recognition performance compared to the BIFs estimated using mismatched speech materials. When disyllabic words were the test material, the advantage of using the matching training material was not always significant (compared to the BIF for nonsense sentences: $p < 0.01$, Cohen's $d = 1.61$; compared to the BIF for meaningful sentences: $p = 0.92$, Cohen's $d = 0.03$). This lack of benefit from material-specific BIFs was in line with the relatively poor predictive power of the

FIG. 5. The average proportion of variance accounted for (R^2) by the SII model with BIFs derived for Mandarin disyllabic words, nonsense sentences, and meaningful sentences. Results are shown for three types of test materials. Single asterisk indicates differences at a significance level of 0.05, while double asterisks indicate differences at a significance level of 0.01.

FIG. 6. The computation procedure within each repetition of the cross-validation analysis to investigate the potential benefit of language-specific BIFs.

simplified SII model for the current recognition data for disyllabic words, as demonstrated in Sec. III C and Fig. 2.

F. The benefits of language-specific BIFs

To evaluate the potential benefit of language-specific BIFs, the Mandarin-based BIFs derived in the current study and the English-based BIFs for similar speech materials were used to predict the speech intelligibility of Mandarin speech. Three cross-validation analyses similar to those presented in Sec. III E were conducted, separately for the three speech materials. For each speech material, a total of 16 repetitions were included in the analysis, so that each listener was assigned to the training set 12 times and the test set 4 times.

Figure 6 illustrates the computational steps involved in each repetition of the cross-validation analysis for one speech material. First, for the training set, the average BIF across the listeners was calculated, which was then used to compute the SII values for all possible stimuli (i.e., combinations of speech bands and SNR). Second, the SII values for the possible stimuli were also computed using the BIFs reported for similar test materials in English. Third, for the test set, the trial-by-trial stimuli and responses from the test set were pooled together. Finally, the pooled data from the test set were predicted by the computed SII values using either the Chinese or English BIFs.

For each BIF, a logistic regression analysis was conducted, treating the computed SII values for all key characters (i.e., syllables) from all trials in the test set as the independent variable and the response correctness as the dependent variable. To evaluate the predictive capability of the SII values (hence, the corresponding BIF), the fitted logistic model was used to generate the predicted scores for all sampled stimuli in the test set. The average recognition scores for these stimuli formed the subjective scores, and the proportion of variance accounted for by the logistic model (R^2) was calculated, similar to the analysis presented in Sec. III C and Fig. 2. Following the 16 repetitions of the cross-validation procedure, 16 estimates of R^2 were obtained for each predicting BIF and each of the 3 speech materials.

Wilcoxon signed rank tests were conducted to compare the R^2 among different BIFs. When Mandarin disyllabic

words were the test material (top panel of Fig. 7), the SII model with the BIF for Mandarin disyllabic words derived in this study performed better than the BIF for English NU6 monosyllabic words using either the qBIF (i.e., “English qBIF NU6 BIF”; Shen and Kern, 2018) or traditional (i.e., “English ANSI NU6 BIF”; ANSI, 1997) method (compared to the English qBIF NU6 BIF: $p < 0.01$, Cohen’s $d = 3.48$; compared to the English ANSI NU6 BIF: $p < 0.01$, Cohen’s

FIG. 7. The average proportion of variance accounted for (R^2) by the SII model with BIFs derived for Mandarin and English speech materials. Results are shown for three types of test materials. Single asterisk indicates differences at a significance level of 0.05, while double asterisks indicate differences at a significance level of 0.01.

$d = 3.72$). When Mandarin nonsense sentences were the test material (middle panel of Fig. 7), a similar English material was the SPIN sentences with low predictability. The BIF for this English material was available in the ANSI standard estimated using the traditional method with high- and low-predictability sentences combined (i.e., “English ANSI SPIN BIF”). Additionally, the BIF for SPIN sentences with low predictability estimated using the compound approach (i.e., “English compound SPIN-Low BIF”) was reported by Healy *et al.* (2013). Compared to the BIFs specified for the SPIN sentences, the BIF for Mandarin nonsense sentences provided better predictions on speech intelligibility (compared to the English ANSI SPIN BIF: $p < 0.01$, Cohen’s $d = 6.71$; compared to the English compound SPIN-Low BIF: $p < 0.01$, Cohen’s $d = 7.81$). When Mandarin meaningful sentences were the test material (bottom panel of Fig. 7), the predictability of the SII model using the BIF for Mandarin meaningful sentences was better than those using the BIFs for English meaningful sentences, including the BIF for the high-predictability SPIN sentences reported by Shen and Kern (2018) estimated using the qBIF procedure (i.e., “English qBIF SPIN-High BIF,” $p < 0.01$, Cohen’s $d = 6.23$), the BIF specified for the SPIN sentences in the ANSI standard (i.e., English ANSI SPIN BIF; $p < 0.01$, Cohen’s $d = 8.40$), and the BIF specified for short passages in the ANSI standard estimated using the traditional method (i.e., “English ANSI short passages BIF,” $p < 0.01$, Cohen’s $d = 7.33$). Therefore, for all three types of Mandarin speech materials, using the language-specific BIFs seems to be advantageous in predicting speech-recognition performance. The possible causes underlying the differences between the Mandarin and English BIFs will be discussed further in Sec. IV C.

IV. DISCUSSION

A. Comparisons to previous studies of Mandarin BIFs

Kuo (2013) estimated the BIFs for monosyllabic words from MSTMs (Zhang *et al.*, 2008) and sentences in Mandarin Hearing In Noise Test (MHINT; Wong *et al.*, 2007b), using the traditional low- and high-pass filtering approaches. Their findings showed that the BIFs for Mandarin monosyllabic words and sentences had a similar shape, with a peak in the importance function in the 1413–2239 Hz region and a secondary peak near 125 Hz. The BIFs derived in the current study also exhibited a bimodal configuration (i.e., a low-frequency peak at 500 Hz and a high-frequency peak at 2000 Hz). The low-frequency peak occurred at 500 Hz, higher than those reported by Kuo (2013). For Mandarin meaningful sentences, the current study found that the low-frequency peak carried a higher weight than the high-frequency peak, in contrary to the dominance of the high-frequency peak found by Kuo (2013).

The differences in the BIFs are likely due to the methodological differences between the two studies. Noticeably, the current study used the compound approach to present the speech bands (Apoux and Healy, 2012), which is expected to provide different BIF estimates from the traditional low- and high-pass filtering approach (Healy *et al.*, 2013). Du *et al.*

(2018) compared the BIFs for Mandarin monosyllabic words estimated using both the traditional method (Chen *et al.*, 2016) and the qBIF procedure (using the compound approach). The BIF estimated using the qBIF procedure had a less pronounced peak at 2 kHz and a more evident bimodal configuration than the BIF estimated using the traditional method, which generally parallels with the observed deviations from the current study to that of Kuo (2013).

In addition, the current study and that of Kuo (2013) divided the speech bands differently. The current study used six octave-frequency bands from 250 to 8000 Hz, while Kuo (2013) divided the spectrum into ten bands spanning approximately six octaves (125–7080 Hz). Therefore, the BIF estimated by Kuo (2013) may contain certain spectral detail or features below 250 Hz not captured by the octave-frequency BIF implemented in the current study.

B. The effect of speech material

Previous studies have demonstrated that the BIFs and transfer functions for different English speech materials differ substantially. The results of this study suggested different Mandarin speech materials also led to different BIFs and transfer functions. The effect of speech materials may be caused by a number of factors, and one of the factors is the amount of contextual cues. For English, with greater contextual information, the shape of the BIF tends to flatten, and the low-frequency region becomes more important (Studebaker *et al.*, 1987; Studebaker and Sherbecoe, 1991; DePaolis *et al.*, 1996). The availability of contextual cues also leads to lower thresholds and steeper slopes for the transfer function (Hirsh *et al.*, 1954; Bell *et al.*, 1992; DePaolis *et al.*, 1996). Word frequency may be another factor that underlies the speech material effect. A lower word frequency may be associated with greater difficulties in speech recognition.

Among the three Mandarin speech corpus used in this study, the meaningful sentences carry the greatest contextual information, including both syntactic and semantic cues. As shown in Fig. 1, the spectral weight for Mandarin meaningful sentences was higher than those for the other speech materials at 250 Hz, and the shape of the BIF for meaningful sentences was the flattest among the three test materials. As for the transfer function, it is shown that the transfer function for meaningful sentences was steeper than disyllabic words and nonsense sentences. Paired t -tests showed that the slope of the transfer function for meaningful sentences was significantly steeper than disyllabic words [$t(15) = 5.80$, $p < 0.01$, Cohen’s $d = 1.45$] or nonsense sentences [$t(15) = 11.64$, $p < 0.01$, Cohen’s $d = 2.91$]. The threshold for meaningful sentences was significantly lower than disyllabic words [$t(15) = -12.32$, $p < 0.01$, Cohen’s $d = 3.08$] or nonsense sentences [$t(15) = -18.34$, $p < 0.01$, Cohen’s $d = 4.59$]. These observations are consistent with previous studies for English speech materials.

Among the three Mandarin speech materials included in the current study, nonsense sentences were shown to be the most difficult. As shown in Fig. 2, at a given SII value (e.g., for a SII of 0.3), much lower recognition scores were found

for nonsense sentences than either disyllabic words or meaningful sentences. The nonsense sentences may be different from the other two materials in several ways, making direct comparisons of results (i.e., the shapes of the BIF and the transfer function) across materials difficult.

First, the female talker of the nonsense sentences was different from the male talker for the disyllabic words and meaningful sentences. It is known that the shape of the BIF varies across different talkers, and it is noticeably different between male and female talkers. [Chen et al. \(2016\)](#) measured the BIF for Mandarin monosyllables spoken by both a male and a female talker. The female talker led to higher spectral weights at 160 Hz compared to that for the male talker. [Yoho et al. \(2018\)](#) measured the BIFs for English sentences using both male and female talkers, and the results showed a transposition of the BIF toward lower frequency regions from the female to male talker. On the other hand, [Yoho et al. \(2018\)](#) also noted that the talker effect (i.e., same sentence material but different talkers) on the BIF was relatively weak compared to the effect of speech material (i.e., same talker but different types of materials).

Second, the key words in the nonsense sentences were different from the disyllabic words of MSTMs in terms of word frequency, although they were both disyllabic. The disyllabic words used in MSTMs were much more common than those used in the nonsense sentences. To demonstrate this difference in word frequency, we calculated the average frequency of appearance for the disyllabic words according to a contemporary Chinese database ([Li and Wang, 2008](#)), which lists the frequencies of appearance in daily speech communication for Chinese words. The average frequency of appearance for the disyllabic words used in MSTM is about 0.121‰, and the variance is 0.05‰; on the other hand, the average frequency of appearance for the disyllabic words used in nonsense sentences is only about 0.025‰, and the variance is 0.007‰. It is possible that the lower frequency of appearance makes the recognition of nonsense sentences more difficult than disyllabic words materials.

Third, the recognition of the disyllabic words in nonsense sentences may be more challenging than the recognition of isolated disyllabic words because it requires additional cognitive effort. Because each nonsense sentence contained three key words (six key characters) and the listeners were instructed to recall these key words after the presentation of the sentence. This task is more demanding than recognizing isolated words, since the words presented earlier in the sentence would have to be retained in the short-term memory while later words in the sentence are recognized (e.g., [Stine and Wingfield, 1987](#); [Humes et al., 1993](#); [Riggs et al., 1993](#); [Surprenant, 1999](#)).

Last, the SII model for the nonsense sentences was estimated last for all listeners in the current experiment. This means that the poorer performance associated with the nonsense sentences may have reflected listener fatigue rather than an actual speech-material effect. An effect of fatigue was possible in the current study even though frequent breaks were encouraged during data collection. To investigate whether fatigue dominated the performance toward the end of the test session, the thresholds of the transfer function estimated from the first and second qBIF runs using

nonsense sentences were compared. Across the 16 listeners, there was no significant difference between the thresholds for the test and retest qBIF runs [$t(15) = 1.23$, $p = 0.23$, Cohen's $d = 0.31$]. This test result shows that fatigue may not be a dominating factor affecting the results during the last hour of data collection.

C. Comparison between Mandarin and English BIFs

There are six BIFs specified for different English speech materials in ANSI-3.5 ([ANSI, 1997](#)). The BIFs are also available in the literature for several Mandarin speech materials (e.g., ([Kuo, 2013](#); [Chen et al., 2016](#); [Du et al., 2018](#))). As reviewed in Sec. IV A, methodological differences may lead to differences in the estimated BIFs for both English and Mandarin. Therefore, comparisons between the BIFs for the two languages may not be informative when different estimation methods were used. Since the qBIF procedure was utilized to estimate Mandarin BIFs in the current study, the most comparable results for English BIFs are those reported by [Shen and Kern \(2018\)](#), estimated using the same procedure.

[Shen and Kern \(2018\)](#) estimated the BIF for English sentences in the SPIN ([Kalikow et al., 1977](#)) test materials. They used the high-context (SPIN-High; [Bilger et al., 1984](#)) sentences in the SPIN corpus, which was similar to the meaningful sentences used in the current study in terms of the available syntactic and semantic context. Their results averaged across 30 listeners and the average BIF for Mandarin meaningful sentences from the current study are plotted together in Fig. 8. The BIF for Mandarin sentences exhibited relatively higher weights than those for English sentences in the 250- and 8000-Hz bands.

This difference between the Mandarin and English BIFs may be due to a number of factors. Mandarin Chinese is a tonal language, and the tone perception mainly depends on fundamental frequency (F_0), which could have led to the relatively higher importance for the 250-Hz band. Moreover, the shape of the BIF may also be influenced by the differences

FIG. 8. Comparison of the BIFs for Chinese meaningful sentences derived in the current study and for a similar material in English. Both BIFs were derived using the qBIF procedure.

in the syllable structure and grammar between English and Mandarin.

It is worth pointing out that each sentence in the SPIN corpus only has one key word at the end of the sentence, while every meaningful Mandarin sentence used in the current study has several key words. Yoho *et al.* (2018) found that the BIFs for SPIN (only one key word at the end of sentence) and IEEE sentences (five key words; Rothauser, 1969) did not differ significantly. Therefore, it is unlikely that the difference between the Mandarin and English BIFs was dominated by the number and position of key words in the test sentences.

V. SUMMARY

In this study, the octave-band BIFs for Mandarin Chinese disyllabic words, meaningful sentences, and nonsense sentences were estimated using an adaptive procedure, namely the qBIF procedure. The results suggest that the low-frequency bands became more important with the increase of contextual cues in Mandarin speech materials, which is consistent with the results for English. The BIFs for Mandarin and English estimated using similar speech materials and the same method were quite different. The BIF was flatter for Mandarin than for English. Using the BIF specific to the test speech material and language could significantly improve the predictive power of the SII model. Therefore, it is important to develop separate BIFs for different languages and speech material types. The BIFs derived in this study could provide a reference for hearing aid fitting for the Mandarin-speaking population. However, further work is needed to establish the efficacy of applying the material- and language-specific BIFs for hearing-aid fitting.

ACKNOWLEDGMENTS

The work was supported by the National Natural Science Foundation of China (Grant Nos. 61771023, 11590773, and 61473008), and National Institutes of Health (NIH) Grant No. R21 DC013406.

- ANSI (1969). S3.5, *Methods for the Calculation of the Articulation Index* (American National Standards Institute, New York).
- ANSI (1997). S3.5, *Methods for the Calculation of the Speech Intelligibility Index* (American National Standards Institute, New York), pp. 90–119.
- Apoux, F., and Healy, E. W. (2012). “Use of a compound approach to derive auditory-filter-wide frequency-importance functions for vowels and consonants,” *J. Acoust. Soc. Am.* **132**, 1078–1087.
- Bell, T. S., Dirks, D. D., and Trine, T. D. (1992). “Frequency-importance functions for words in high- and low-context sentences,” *J. Speech Lang. Hear. Res.* **35**, 950–959.
- Benkí, J. R. (2003). “Quantitative evaluation of lexical status, word frequency, and neighborhood density as context effects in spoken word recognition,” *J. Acoust. Soc. Am.* **113**, 1689–1705.
- Bilger, R. C., Nuetzel, J., Rabinowitz, W., and Rzeczkowski, C. (1984). “Standardization of a test of speech perception in noise,” *J. Speech Lang. Hear. Res.* **27**, 32–48.
- Boothroyd, A. (1978). “Speech perception and sensorineural hearing loss,” in *Auditory Management of the Hearing Impaired Child* (University Park Press, Baltimore, MD), pp. 117–144.
- Boothroyd, A. (2008). “The performance/intensity function: An underused resource,” *Ear Hear.* **29**, 479–491.
- Boothroyd, A., and Nittrouer, S. (1988). “Mathematical treatment of context effects in phoneme and word recognition,” *J. Acoust. Soc. Am.* **84**, 101–114.
- Bosen, A. K., and Chatterjee, M. (2016). “Band importance functions of listeners with cochlear implants using clinical maps,” *J. Acoust. Soc. Am.* **140**, 3718–3727.
- Broadbent, D. E. (1967). “Word-frequency effect and response bias,” *Psychol. Rev.* **74**, 1–15.
- Bronkhorst, A. W., Bosman, A. J., and Smoorenburg, G. F. (1993). “A model for context effects in speech recognition,” *J. Acoust. Soc. Am.* **93**, 499–509.
- Bronkhorst, A. W., Brand, T., and Wagener, K. (2002). “Evaluation of context effects in sentence recognition,” *J. Acoust. Soc. Am.* **111**, 2874–2886.
- Chen, F., Wong, L. L., and Wong, E. Y. (2013). “Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility,” *J. Acoust. Soc. Am.* **134**, EL178–EL184.
- Chen, J., Huang, Q., and Wu, X. (2016). “Frequency importance function of the speech intelligibility index for Mandarin Chinese,” *Speech Commun.* **83**, 94–103.
- Chen, J., Li, H., Li, L., Wu, X., and Moore, B. C. (2012). “Informational masking of speech produced by speech-like sounds without linguistic content,” *J. Acoust. Soc. Am.* **131**, 2914–2926.
- Ching, T. Y., Dillon, H., and Byrne, D. (1998). “Speech recognition of hearing-impaired listeners: Predictions from audibility and the limited role of high-frequency amplification,” *J. Acoust. Soc. Am.* **103**, 1128–1140.
- DePaolis, R. A., Janota, C. P., and Frank, T. (1996). “Frequency importance functions for words, sentences, and continuous discourse,” *J. Speech Lang. Hear. Res.* **39**, 714–723.
- Du, Y., Kong, L., Wang, Q., Wu, X., and Li, L. (2011). “Auditory frequency-following response: A neurophysiological measure for studying the ‘cocktail-party problem,’” *Neurosci. Biobehav. Rev.* **35**, 2046–2057.
- Du, Y., Shen, Y., Yang, H., Wu, X., and Chen, J. (2018). “Measuring the band importance function for Mandarin Chinese with a Bayesian adaptive procedure,” in *Proc. Interspeech 2018*, pp. 961–965.
- Duggirala, V., Studebaker, G. A., Pavlovic, C. V., and Sherbecoe, R. L. (1988). “Frequency importance functions for a feature recognition test material,” *J. Acoust. Soc. Am.* **83**, 2372–2382.
- Fletcher, H., and Galt, R. H. (1950). “The perception of speech and its relation to telephony,” *J. Acoust. Soc. Am.* **22**, 89–151.
- Fletcher, H., and Steinberg, J. (1929). “Articulation testing methods,” *Bell Labs Tech. J.* **8**, 806–854.
- French, N. R., and Steinberg, J. C. (1947). “Factors governing the intelligibility of speech sounds,” *J. Acoust. Soc. Am.* **19**, 90–119.
- Healy, E. W., and Bacon, S. P. (2007). “Effect of spectral frequency range and separation on the perception of asynchronous speech,” *J. Acoust. Soc. Am.* **121**, 1691–1700.
- Healy, E. W., and Warren, R. M. (2003). “The role of contrasting temporal amplitude patterns in the perception of speech,” *J. Acoust. Soc. Am.* **113**, 1676–1688.
- Healy, E. W., Yoho, S. E., and Apoux, F. (2013). “Band importance for sentences and words reexamined,” *J. Acoust. Soc. Am.* **133**, 463–473.
- Hirsh, I. J., Reynolds, E. G., and Joseph, M. (1954). “Intelligibility of different speech materials,” *J. Acoust. Soc. Am.* **26**, 530–538.
- Hochmuth, S., Kollmeier, B., Brand, T., and Jürgens, T. (2015). “Influence of noise type on speech reception thresholds across four languages measured with matrix sentence tests,” *Int. J. Audiol.* **54**, 62–70.
- Howes, D. (1957). “On the relation between the intelligibility and frequency of occurrence of English words,” *J. Acoust. Soc. Am.* **29**, 296–305.
- Huang, Y., Huang, Q., Chen, X., Qu, T., Wu, X., and Li, L. (2008). “Perceptual integration between target speech and target-speech reflection reduces masking for target-speech recognition in younger adults and older adults,” *Hear. Res.* **244**, 51–65.
- Humes, L. E., and Kidd, G. R. (2016). “Speech recognition for multiple bands: Implications for the Speech Intelligibility Index,” *J. Acoust. Soc. Am.* **140**, 2019–2026.
- Humes, L. E., Nelson, K. J., Pisoni, D. B., and Lively, S. E. (1993). “Effects of age on serial recall of natural and synthetic speech,” *J. Speech Lang. Hear. Res.* **36**, 634–639.
- Jansen, S., Luts, H., Wagener, K. C., Kollmeier, B., Del Rio, M., Dauman, R., James, C., Fraysse, B., Vormès, E., and Frachet, B. (2012). “Comparison of three types of French speech-in-noise tests: A multicenter study,” *Int. J. Audiol.* **51**, 164–173.
- Jin, I.-K., Kates, J. M., and Arehart, K. H. (2017). “Does language matter when using a graphical method for calculating the Speech Intelligibility Index?,” *J. Am. Acad. Audiol.* **28**, 119–126.
- Jin, I.-K., Lee, J., Lee, K., Kim, J., Kim, D., Sohn, J., and Lee, K.-J. (2016). “The band-importance function for the Korean standard sentence lists for adults,” *J. Audiol. Otol.* **20**, 80.

- Kalikow, D. N., Stevens, K. N., and Elliott, L. L. (1977). "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *J. Acoust. Soc. Am.* **61**, 1337–1351.
- Keidser, G., Dillon, H., Flax, M., Ching, T., and Brewer, S. (2011). "The NAL-NL2 prescription procedure," *Audiol. Res.* **1**, 88–90.
- Kontsevich, L. L., and Tyler, C. W. (1999). "Bayesian adaptive estimation of psychometric slope and threshold," *Vision Res.* **39**, 2729–2737.
- Kuo, M.-W. (2013). "Frequency importance functions for words and sentences in Mandarin Chinese: Implications for hearing aid prescriptions in tonal languages," Ph.D. thesis, University of Western Sydney, Australia.
- Lee, S., and Mendel, L. L. (2017). "Derivation of frequency importance functions for the AzBio sentences," *J. Acoust. Soc. Am.* **142**, 3416–3427.
- Lee, W. K., and Kim, S. J. (2012). "The study of frequency importance function of the Korean monosyllabic words," *Audiol. Speech Res.* **8**, 24–33.
- Li, X., and Wang, T. (2008). *Lexicon of Common Words in Contemporary Chinese* (The Commercial Press, Beijing).
- Lin, T., and Wang, L. (1992). *Yuyinxue jiaocheng (Introduction to Phonetics)* (Peking University Press, Peking) (in Chinese).
- Magnusson, L. (1996). "Speech intelligibility index transfer functions and speech spectra for two Swedish speech recognition tests," *Scand. Audiol.* **25**, 59–67.
- Miller, G. A., Heise, G. A., and Lichten, W. (1951). "The intelligibility of speech as a function of the context of the test materials," *J. Exp. Psychol.* **41**, 329.
- Moore, B. C., and Glasberg, B. R. (1998). "Use of a loudness model for hearing-aid fitting. I. Linear hearing aids," *Br. J. Audiol.* **32**, 317–335.
- Mueller, H. G., and Killion, M. C. (1990). "An easy method for calculating the articulation index," *Hear. J.* **43**, 14–17.
- Narne, V. K., Prabhu, P., Thuvassery, P., Ramachandran, R., Kumar, A., Raveendran, R., and Gafoor, S. A. (2016). "Frequency importance function for monosyllables in Malayalam," *Hear. Balance Commun.* **14**, 201–206.
- Olsen, W. O., Van Tasell, D. J., and Speaks, C. E. (1997). "Phoneme and word recognition for words in isolation and in sentences," *Ear Hear.* **18**, 175–188.
- Pavlovic, C. V., and Studebaker, G. A. (1984). "An evaluation of some assumptions underlying the articulation index," *J. Acoust. Soc. Am.* **75**, 1606–1612.
- Riggs, K. M., Wingfield, A., and Tun, P. A. (1993). "Passage difficulty, speech rate, and age differences in memory for spoken text: Speech recall and the complexity hypothesis," *Exp. Aging Res.* **19**, 111–128.
- Rothausen, E. (1969). "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.* **17**, 225–246.
- Savin, H. B. (1963). "Word-frequency effect and errors in the perception of speech," *J. Acoust. Soc. Am.* **35**, 200–206.
- Shen, Y., and Kern, A. B. (2018). "An analysis of individual differences in recognizing monosyllabic words under the Speech Intelligibility Index framework," *Trends Hear.* **22**, 2331216518761773.
- Shen, Y., and Richards, V. M. (2013). "Bayesian adaptive estimation of the auditory filter," *J. Acoust. Soc. Am.* **134**, 1134–1145.
- Shen, Y., Sivakumar, R., and Richards, V. M. (2014). "Rapid estimation of high-parameter auditory-filter shapes," *J. Acoust. Soc. Am.* **136**, 1857–1868.
- Sherbecoe, R. L., and Studebaker, G. A. (2002). "Audibility-index functions for the connected speech test," *Ear Hear.* **23**, 385–398.
- Smits, C., Theo Goverts, S., and Festen, J. M. (2013). "The digits-in-noise test: Assessing auditory speech recognition abilities in noise," *J. Acoust. Soc. Am.* **133**, 1693–1706.
- Stine, E. L., and Wingfield, A. (1987). "Process and strategy in memory for speech among younger and older adults," *Psychol. Aging* **2**, 272.
- Studebaker, G. A., Pavlovic, C. V., and Sherbecoe, R. L. (1987). "A frequency importance function for continuous discourse," *J. Acoust. Soc. Am.* **81**, 1130–1138.
- Studebaker, G. A., and Sherbecoe, R. L. (1991). "Frequency-importance and transfer functions for recorded CID W-22 word lists," *J. Speech Lang. Hear. Res.* **34**, 427–438.
- Studebaker, G. A., Sherbecoe, R. L., and Gilmore, C. (1993). "Frequency-importance and transfer functions for the Auditec of St. Louis recordings of the NU-6 word test," *J. Speech Lang. Hear. Res.* **36**, 799–807.
- Surprenant, A. M. (1999). "The effect of noise on memory for spoken syllables," *Int. J. Psychol.* **34**, 328–333.
- van Rooij, J. C., and Plomp, R. (1991). "The effect of linguistic entropy on speech perception in noise in young and elderly listeners," *J. Acoust. Soc. Am.* **90**, 2985–2991.
- Warren, R. M., Riener, K. R., Bashford, J. A., and Brubaker, B. S. (1995). "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," *Percept. Psychophys.* **57**, 175–182.
- Wong, L. L., Ho, A. H., Chua, E. W., and Soli, S. D. (2007a). "Development of the Cantonese speech intelligibility index," *J. Acoust. Soc. Am.* **121**, 2350–2361.
- Wong, L. L., Soli, S. D., Liu, S., Han, N., and Huang, M.-W. (2007b). "Development of the Mandarin hearing in noise test (MHINT)," *Ear Hear.* **28**, 70S–74S.
- Yang, Z., Chen, J., Huang, Q., Wu, X., Wu, Y., Schneider, B. A., and Li, L. (2007). "The effect of voice cuing on releasing Chinese speech from informational masking," *Speech Commun.* **49**, 892–904.
- Yoho, S. E., Healy, E. W., Youngdahl, C. L., Barrett, T. S., and Apoux, F. (2018). "Speech-material and talker effects in speech band importance," *J. Acoust. Soc. Am.* **143**, 1417–1426.
- Zhang, H. (2009). "Rethinking the hearing aid market development in China," *Zhongguo Yixue Wenzhai Yanhou Kexue News Rev.* **24**, 70–72.
- Zhang, H., Wang, S., Chen, J., Wang, L., Shao, G.-y., Li, Y.-l., and Jiang, W.-b. (2008). "The Mandarin speech test materials (MSTMs): Development and application," *Chinese Sci. J. Hear. Speech Rehabil.* **6**, 16–17.