Assessing the perceptual contributions of level-dependent segments to sentence intelligibility

Tian Guan and Guang-xing Chu
Research Centre of Biomedical Engineering, Graduate School at Shenzhen, Tsinghua University, Lishui Road, Xili, Nanshan District, Shenzhen 518055, China

Yu Tsao
Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan

Fei Chen[a)]
Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Xueyuan Road 1088#, Xili, Nanshan District, Shenzhen 518055, China

(Received 24 November 2015; revised 17 October 2016; accepted 28 October 2016; published online 16 November 2016)

The present work assessed the contributions of high root-mean-square (RMS) level (H-level, containing primarily vowels) and middle-RMS-level (M-level, with mostly consonants and vowel-consonant transitions) segments to the intelligibility of noise-masked and noise-suppressed sentences. In experiment 1, noise-masked (by speech-spectrum shaped noise and 6-talker babble) Mandarin sentences were edited to preserve only H- or M-level segments, while replacing the non-target segments with silence. In experiment 2, Mandarin sentences were subjected to four commonly-used single-channel noise-suppression algorithms before generating H-level-only and M-level-only noise-suppressed sentences. To test the influence of an effective signal-to-noise ratio (SNR) on intelligibility, both experiments incorporated a condition in which the SNRs of H-level segments and M-level segments were matched. The processed sentences were presented to normal-hearing listeners to recognize. Experimental results showed that (1) H-level-only sentences carried more perceptual information than M-level-only sentences under both noise-masked and noise-suppressed conditions; and (2) this intelligibility advantage of H-level-only sentences over M-level-only sentences persisted even when effective SNR levels were matched, and it might be attributed to the perceptual advantage of vowels in speech intelligibility. In addition, the lesser distortion in H-level segments than in M-level segments following noise-suppression processing suggests that differential processing distortion might contribute to the H-level advantage observed.
 [http://dx.doi.org/10.1121/1.4967453]

[JFL] Pages: 3745–3754

I. INTRODUCTION

To understand the factors that support reliable speech perception, particularly in adverse listening environments, it is of great importance to identify those speech segments that carry the most intelligibility information. Three types of sentence segmentation methods have been in common use for assessment of segmental contributions to speech intelligibility, namely, vowel and consonant boundary (e.g., Cole *et al.*, 1996; Kewley-Port *et al.*, 2007; Fogerty and Kewley-Port, 2009), relative root-mean-square (RMS) level (e.g., Kates and Arehart, 2005; Chen and Loizou, 2012), and cochlea-scaled entropy (CSE; Stilp and Kluender, 2010). Studies investigating the segmental contributions of vowels and consonants to speech intelligibility have indicated that vowel-only sentences (consonants replaced by noise) produced a remarkable 2:1 intelligibility advantage over consonant-only sentences (vowels replaced by noise) in normal-hearing (NH) listeners (e.g., Cole *et al.*, 1996; Kewley-Port *et al.*, 2007; Fogerty and Kewley-Port, 2009). However, the relative segmental contributions of vowels and consonants to speech intelligibility have been controversial. Even with the most sophisticated phoneme detection algorithms, fully isolating the individual contributions of vowels and consonants can be quite challenging because vowels carry co-articulatory information about consonants and vice versa at consonant-vowel boundaries.

The relative RMS-level and CSE based segmentation methods obviate the traditional phonetic distinction between vowels and consonants. In relative RMS-level based segmentation, the speech signal is divided into high (H), meddle (M), and low (L) levels. The H-level consists of segments at or above the overall RMS level of the whole utterance, the M-level consists of segments ranging from 10 dB below the overall RMS level to the overall RMS level, and the L-level consists of segments ranging from 30 dB below to 10 dB below the overall RMS level (see example in Fig. 1). H-level segments include primarily vowels and semivowels, M-level segments include mostly consonants and vowel-consonant transitions, and L-level segments include primarily weak consonants (Kates and Arehart, 2005; Chen and Loizou, 2012). Several studies have assessed the importance of stimulus change (e.g., consonant-vowel boundary and entropy) to speech intelligibility. Fogerty and Kewley-Port (2009)

[a)]Electronic mail: fchen@sustc.edu.cn

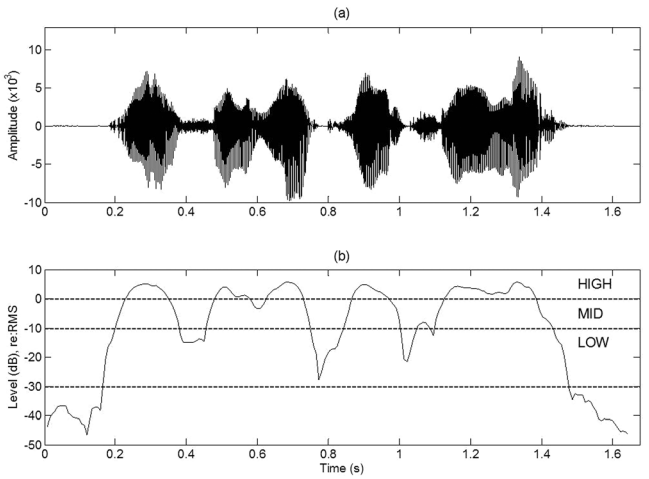

FIG. 1. Example waveforms of (a) a Mandarin sentence and (b) its relative RMS energy expressed in dB relative to the overall RMS level of the whole utterance. Dashed lines in (b) show the boundaries of the high-, middle-, and low-RMS-level regions.

showed that the relative perceptual contributions of vowels and consonants are affected by co-articulation information across consonant-vowel boundaries. Chen and Loizou (2012) found that speech intelligibility is modeled more accurately by M-level segments, which contain most consonant-vowel boundaries, than by H-level segments. Stilp and Kluender (2010) suggested that it is CSE, rather than consonants, vowels, or time, that best predicts speech intelligibility. They found that replacing low-entropy segments yielded a relatively small impact on intelligibility while replacing high-entropy segments significantly reduced sentence intelligibility.

Although the contributions of vowel-consonant and entropy based segmentation to sentence intelligibility have been examined, little attention has been paid to the perceptual contributions of intensity-level-dependent (e.g., H- and M-level) segments to speech understanding, which is the primary aim of this work. Compared with the computation of the consonant-vowel boundary or entropy based speech segmentation, the intensity-level based segmentation is relatively easy to implement (as illustrated by the three-level RMS-based segmentation in Sec. II). Furthermore, examining the perceptual contributions of H- and M-level segments may avoid the effect of segment duration on intelligibility because the H- and M-level segments have almost equal durations (see Sec. II). In contrast, vowel segments are markedly longer than consonant segments. For instance, Chen *et al.* (2013) showed that vowels and consonants occupied about 66.3% and 25.9% of the duration of Mandarin sentences, respectively.

Because H-level segments are characterized primarily by vowel components, and previous studies have demonstrated a vowel over consonant advantage in sentence intelligibility (e.g., Cole *et al.*, 1996; Kewley-Port *et al.*, 2007), the first hypothesis of the present work, examined in experiment 1, is that H-level-only (or H-only) sentences, in which only H-level segments are retained (while the remaining segments are silenced), will be more intelligible than M-level-only (or M-only) sentences, wherein only M-level segments are retained. In addition, segmental contribution to speech intelligibility has been primarily studied in quiet settings. Because noise interference poses a great challenge to speech perception and this interference may affect different speech segments differently, we also hypothesize that H-only sentences are more resistant to noise interference than M-only sentences and that, consequently, their intelligibility advantage over M-only sentences is retained in noisy listening environments. The greater energy of H-level segments relative to M-level segments (by about 10 dB) is expected to produce a relatively better effective signal-to-noise ratio (SNR) when a clean speech signal is mixed with noise. Therefore, we also examine whether the perceptual contribution of H-level segments can be attributed, at least in part, to their larger effective SNR, relative to that of M-level segments, by increasing the effective SNR level of M-level segments to that of H-level segments.

A variety of single-channel noise-suppression algorithms aimed at alleviating background noise interference have been described, including the spectral-subtractive algorithm (Kamath and Loizou 2002), statistical-model-based algorithm (Ephraim and Malah, 1985), and subspace algorithm (Hu and Loizou, 2003). However, most traditional noise-suppression algorithms (e.g., statistically-based Wiener filtering) do not improve speech intelligibility in NH listeners (e.g., Hu and Loizou, 2007; Li *et al.*, 2011). Many studies investigated factors accounting for the performance deficit of traditional noise-suppression processing for speech intelligibility. Most, if not all, noise-suppression algorithms involve a gain reduction stage, in which the mixture spectral envelope is multiplied by a non-linear gain function with the intent of suppressing any background noise that may be present. The shape and choice of the gain function vary across algorithms. However, independent of the function shape, application of the gain function introduces amplification or attenuation distortion into the spectral envelopes. In a study examining the effects of gain-induced non-linear distortions on the intelligibility of noise-suppressed speech, Kim and Loizou (2011) found that amplification, but not attenuation, distortion impaired speech intelligibility. Given the different perceptual cues used by native listeners of different languages, including tonal languages, a comparative evaluation was undertaken wherein four well-established noise-suppression algorithms were applied to noise-masked speech samples in three languages: Mandarin Chinese, Japanese, and English (Li *et al.*, 2011). The majority of tested algorithms did not improve speech intelligibility and significant differences among their performances across the three languages were observed. This indicates that most noise-suppression algorithms were significantly affected by language-specific characteristics.

The reasons underlying the failure of traditional noise-suppression algorithms to improve the intelligibility of noise-suppressed sentences remains unresolved. More specifically, little has been done to investigate the contributions of RMS-level segmentation to the intelligibility of noise-suppressed speech. Hence, experiment 2 in this work seeks to investigate segmental contribution to the intelligibility of

noise-suppressed sentences. We evaluate the performance of four traditional noise-suppression algorithms for segmentally (i.e., H- and M-only) processed sentences. Because H- and M-level segments contain different aspects of perceptual information and have different signal intensities, we predict that H- and M-level segments respond differently to noise-suppression processing, particularly in terms of distortion contained in noise-suppressed segments. Specifically, for experiment 2, we hypothesize that, due to (1) vowel's perceptual importance (as in Hypothesis 1) and (2) the large effective SNR levels and relatively accurate SNR estimation in H-level segments, H-level segments contain less distortion by noise-suppression processing, and are more intelligible than M-level segments in the context of noise-suppression. Additionally, to diminish the possible influence of differing effective SNR levels between H- and M-level segments in experiment 2, we equalize the effective SNR levels of H- and M-level segments.

II. EXPERIMENT 1: CONTRIBUTIONS OF LEVEL-SEGREGATED SEGMENTS TO THE INTELLIGIBILTY OF NOISE-MASKED SPEECH

The purpose of experiment 1 was to compare the intelligibility of H- and M-only sentences in a noise-masked condition.

A. Methods

1. Subjects

Twelve (six males and six females) native-Mandarin-Chinese listeners (aged from 24 to 26 yrs) participated in the experiment. All participants were undergraduate students from Southern University of Science and Technology, and were paid for their participation in this study. All subjects had NH, as determined by having measured pure-tone thresholds (at 250–8000 Hz) lower than 25 dB hearing level.

2. Materials

The sentence material consisted of sentences taken from the Mandarin Hearing in Noise Test (MHINT) database (Wong *et al.*, 2007). There were in total 24 lists in the MHINT corpus. Each MHINT list had ten sentences, and each sentence contained ten keywords. All the sentences were produced by a male speaker. Two types of maskers were used to corrupt the sentences, which included steady-state speech-spectrum shaped noise (SSN) and multi-talker (i.e., 6-talker) babble. Note that (1) to generate the SSN masker, a finite impulse response filter was designed based on the average spectrum of the MHINT sentences, and a white noise was filtered and scaled to the same long-term average spectrum and level as the sentences; and (2) the 6-talker babble contained six (three male and three female) equal-level interfering talkers. A noise segment of the same length as the clean intact (i.e., full-length) speech signal was randomly cut out of the noise recordings, appropriately scaled to reach the desired input SNR level, and finally added to the speech signals at 5 and 0 dB input SNR levels for each masker condition. The input SNR levels were chosen based on the known performance with full-segment sentences.

3. Signal processing

To create H- and M-only sentences, relative RMS-level based sentence segmentation was implemented by first dividing clean speech signals into short-term (16 ms in this study) segments, with a 25% overlap between adjacent segments, and then classifying each segment as H-, M-, or L-level according to its relative RMS intensity, as exemplified in Fig. 1. We adopted the H-, M-, and L-level segmentation thresholds (0, −10, and −30 dB) proposed by Kates and Arehart (2005) wherein each level includes the segments at or above its level-defining threshold, and in the case of the lower two levels, up to the boundary of the next level above.

Noise-masked sentences were edited such that the target level (H- or M-level) segments were retained and the remaining segments were replaced with silence by setting their amplitudes to zero. Because the purpose of this experiment was to assess which segment (H- or M-level) of the original speech signal contained more intelligibility information when the speech signal was mixed with interfering noise, the noise was added to a clean sentence prior to replacing out-level segments with silence. The waveforms of H- and M-only clean and noise-masked sentences are shown in Fig. 2, and the spectrograms of H- and M-only clean and noise-masked sentences are shown in Fig. 3. Before noise-masking and silence-replacement processing, the third condition, i.e., M-only-adjusted, was first processed by computing the effective SNR level of concatenated H-level segments, and scaling the level of concatenated M-level segments to yield the same effective SNR level as that of concatenated H-level segments.

Paradigms in which speech is replaced with silence have been shown to interrupt speech intelligibility more than

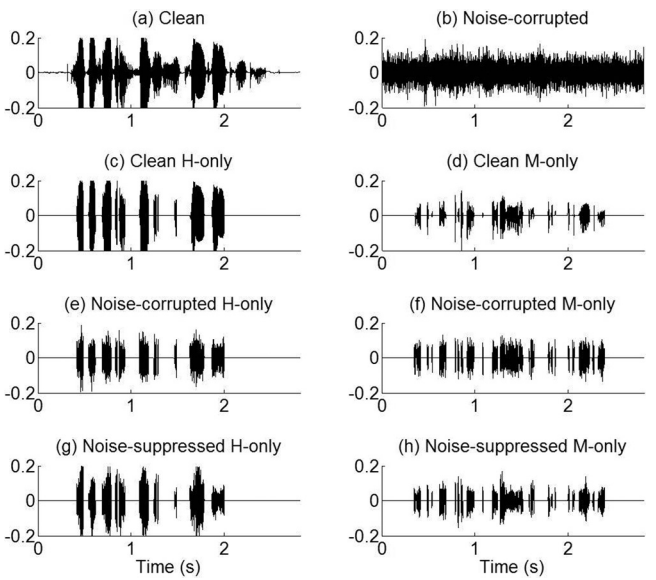

FIG. 2. The waveforms of (a) clean speech, (b) noise-masked speech (by SSN masker at −10 dB SNR), (c) and (d) H- and M-only clean sentences, (e) and (f) H- and M-only noise-masked sentences, and (g) and (h) H- and M-only noise-suppressed (by Wiener filtering) sentences.

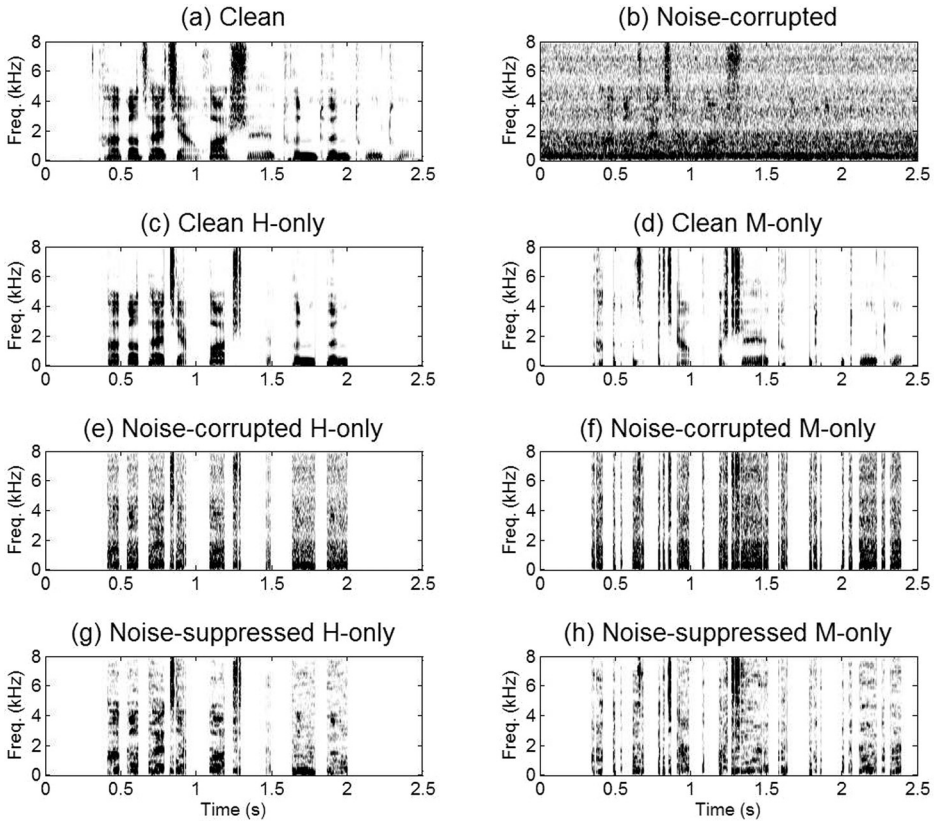

FIG. 3. The spectrograms of (a) clean speech, (b) noise-masked speech (by SSN masker at −10 dB SNR), (c) and (d) H- and M-only clean sentences, (e) and (f) H- and M-only noise-masked sentences, and (g) and (h) H- and M-only noise-suppressed (by Wiener filtering) sentences.

paradigms in which speech is replaced with noise (Cole *et al.*, 1996; Powers and Wilcox, 1977; Benard and Başkent, 2014). The present work chose to replace selected segments with silence to avoid the usage of two noise sources (i.e., one for noise masking to the desired input SNR level, and the other for segmental replacement in noise-replaced paradigm).

This study assessed intelligibility under three segmental conditions: H-only, M-only-adjusted, and M-only. Statistical analysis with all 240 MHINT sentences showed that H-, M-, and L-level segments occupied 27.6%, 23.2%, and 25.0% of the whole sentence duration, respectively, indicating that the average duration of H-level segments was only slightly longer than that of M-level segments. However, paired *t*-tests showed that the duration difference was significant between paired segments. The analysis also showed that the mean durations of H- and M-level segments were 41.4 and 35.0 ms, respectively.

4. Procedure

The experiment was performed in a sound-proof room, and stimuli were played to listeners binaurally through an HD 650 circumaural head-phone (Sennheiser, Germany) set at a comfortable listening level (i.e., ∼65 dB sound pressure level). Before the actual testing session, each subject participated in a 10-min training session and was given six lists of 10 MHINT sentences. The training session familiarized the subjects with the testing procedure and conditions. During the training session, the subjects were allowed to read feedback (i.e., transcription of the training sentences) while they were listening to the sentences. Only six testing conditions [=2 types of maskers (i.e., SSN and 6-talker babble) at 5 dB input SNR level × 3 segmental conditions (i.e., H-only, M-only-adjusted, and M-only)] were used during training and the sentences used during testing were not the same as any of the training sentences. In the testing session, the order of the conditions was randomized across subjects, and the subjects were asked to repeat orally all of the words they heard. In addition, the lists were randomized across listeners. Each subject participated in a total of 12 conditions [=2 types of maskers (i.e., SSN and 6-talker babble) × 2 input SNR levels (i.e., 5 and 0 dB) × 3 segmental conditions (i.e., H-only, M-only-adjusted, and M-only)]. One list of ten Mandarin sentences was used per tested condition, and none of the sentences was repeated across the conditions. Subjects were allowed to listen to each stimulus a maximum of three times (consistent with a previous study, e.g., Chen *et al.*, 2013), and were required to repeat as many words as they could recognize. A simple custom-designed software interface was designed for the listening experiments; each participant used the software interface to control the auditory delivery of the processed stimuli. During the testing session, a tester accompanied the participant and scored his/her response online. A 5-min break was given to the subjects every 30 min to avoid listening fatigue. The intelligibility score for each condition was computed as the ratio between the number of correctly recognized words and the total number of words contained in each list of ten MHINT sentences.

B. Results

Mean recognition scores for all conditions in experiment 1 are shown in Fig. 4. Statistical significance was determined

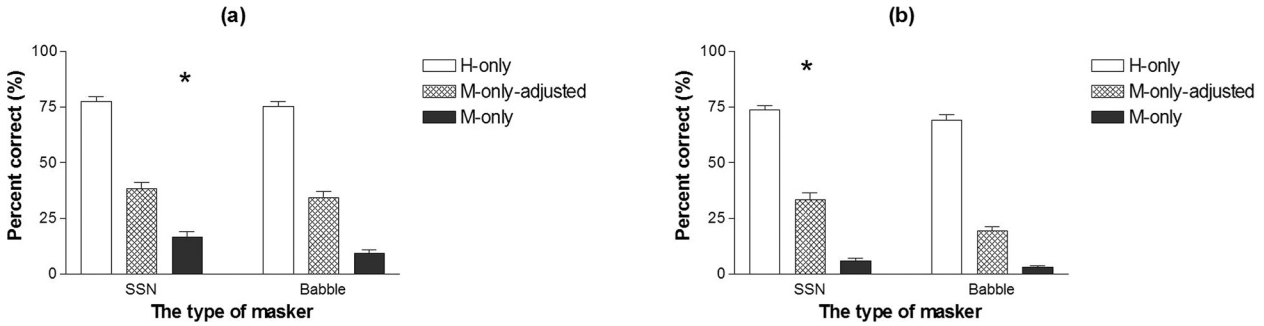

FIG. 4. Sentence recognition scores for all conditions at (a) SSN masker and (b) 6-talker babble masker. The error bars denote ±1 standard error of the mean. Asterisk denotes that the intelligibility score at SSN masker is significantly larger than that at babble masker.

by using the recognition score as the dependent variable, and the type of masker (SSN and babble) and segmental condition (i.e., H-only, H-only-adjusted, and M-only) as the two within-subject factors. The recognition score scores were first converted to rational arcsine units (RAUs) using the rationalized arcsine transform (Studebaker, 1985).

For results at 5 dB input SNR level in Fig. 4(a), two-way analysis of variance (ANOVA) with repeated measures indicated significant effects of the type of masker ($F[1, 11] = 7.28$, $p < 0.05$), segmental condition ($F[2, 22] = 504.67$, $p < 0.001$), and a non-significant interaction ($F[2, 26] = 1.38$, $p = 0.27$) between the type of masker and segmental condition. One-way ANOVA with repeated measures was conducted in each type of masker to further analyze the effect of segmental condition. Alpha level for statistical significance was Bonferroni corrected, and only those tests with a p-value lower than 0.017 (=0.05/3) were considered as significant. At both types of maskers, the results showed significant differences in the performance between segmental conditions H-only and M-only-adjusted, segmental conditions H- and M-only, and segmental conditions M-only-adjusted and M-only ($ps < 0.001$). A paired t-test was conducted in each segmental condition to further analyze the effect of the type of masker. The results showed there was a significant ($p < 0.05$) difference in the performance between the two types of maskers at the M-only condition, but not at the H-only and M-only-adjusted conditions.

For results at 0 dB input SNR level in Fig. 4(b), two-way ANOVA with repeated measures indicated significant effects of the type of masker ($F[1, 11] = 16.93$, $p < 0.05$), segmental condition ($F[2, 22] = 326.36$, $p < 0.001$), and a non-significant interaction ($F[2, 22] = 3.08$, $p = 0.07$) between the type of masker and segmental condition. One-way ANOVA with repeated measures was conducted in each type of masker to further analyze the effect of segmental condition. Alpha level for statistical significance was Bonferroni corrected, and only those tests with a p-value lower than 0.017 (=0.05/3) were considered as significant. At both types of maskers, the results showed significant differences in the performance between segmental conditions H-only and M-only-adjusted, segmental conditions H- and M-only, and segmental conditions M-only-adjusted and M-only ($ps < 0.001$). A paired t-test was conducted in each segmental condition to further analyze the effect of the type of noise. The results showed there was significant ($p < 0.05$) difference in the performance between the two types of maskers at the M-only-adjusted condition, but not at the H- and M-only conditions.

The above results showed that H-only sentences were much more intelligible than M-only sentences, though H-level segments occupied almost the same duration as M-level segments within the sentences, demonstrating a clear intelligibility advantage of H-level segments over M-level segments. In addition, when the effective SNR level of M-level segments was equalized to that of H-level segments, H-only sentences remained more intelligible than M-only sentences (see Fig. 4). This finding indicates that an effective SNR level is not the key factor accounting for the intelligibility advantage of H-level segments over M-level segments.

III. EXPERIMENT 2: CONTRIBUTIONS OF LEVEL-SEGREGATED SEGMENTS TO THE INTELLIGIBILTY OF NOISE-SUPPRESSED SPEECH

The purpose of experiment 2 was to compare the intelligibility of H- and M-only sentences in a noise-suppressed condition generated by processing with commonly-used single-channel noise-suppression algorithms.

A. Methods

1. Subjects

Eleven (seven males and four females) new NH native-Mandarin listeners (aged from 22 to 25 yrs) participated in this experiment. All participants were undergraduate students from Southern University of Science and Technology, and were paid for their participation in this study. All subjects had NH, as determined by having measured pure-tone thresholds (at 250–8000 Hz) lower than 25 dB hearing level.

2. Materials

The speech materials and maskers were the same as those used in experiment 1. A noise segment of the same length as the clean intact speech signal was randomly cut out of the noise recordings, appropriately scaled to reach the desired input SNR level, and finally added to the speech signals at the input SNR levels of −10 and −5 dB for SSN and babble masker, respectively. The input SNR levels were chosen based on the known performance with full-segment sentences.

3. Signal processing

In this experiment, we selected four representative noise-suppression algorithms, including the generalized Karhunen-Loeve transform (KLT) approach (Hu and Loizou, 2003), the Log Minimum Mean Square Error (logMMSE) algorithm (Ephraim and Malah, 1985), the multi-band spectral subtraction (MB) algorithm (Kamath and Loizou, 2002), and the Wiener algorithm based on a prior SNR estimation (Scalart and Filho, 1996), to process noise-masked sentences. These four algorithms encompass the four most commonly-used types of signal-channel noise-suppression methods, namely, the subspace approach, statistical-modeling approach, spectral-subtractive approach, and Wiener-filtering approach (see review in Loizou, 2007). Briefly, the spectral-subtractive algorithm is implemented with an estimate of the clean signal spectrum generated by subtracting an estimate of the noise spectrum from the noise-masked speech spectrum (Kamath and Loizou 2002). The Wiener filter uses *a priori* SNR statistics to design a gain function that suppresses low-SNR segments and preserves high-SNR segments. Owing to its simple model structure, the Wiener filter provides only moderate noise reduction, but at a relatively low computational cost (Scalart and Filho, 1996). For the KLT method, the noise-masked speech signal is first projected into orthogonal subspaces. Those KLT parts representing the signal subspace are modified by a gain function determined by the estimator, while the remaining KLT parts representing the noise subspace are nulled. Finally, the enhanced signal is obtained from the inverse KLT of the modified parts (Hu and Loizou, 2003). The statistical-modeling approach employs statistical models with optimization criteria (e.g., minimum mean square error) to estimate the magnitude spectrum of the speech signal of interest (Ephraim and Malah, 1985). For each algorithm, the implementation parameters reported in the above study referenced for each were used. A statistical-model-based voice activity detector (VAD) was used in all algorithms, except with the subspace method, to update the noise spectrum during speech-absent segments (Sohn *et al.*, 1999). For the subspace method, we used the VAD method in Mittal and Phamdo (2000) with the threshold value set to 1.2. Detailed descriptions of the algorithms tested can be found in Hu and Loizou (2007) and Loizou (2007). The MATLAB code used to implement the above four noise-suppression algorithms was obtained from the connected discourse in Loizou (2007).

Noise-masked sentences were first processed by noise-suppression algorithms. Subsequently, either H- or M-level segments (classified from the clean speech signal) of noise-suppressed sentences were retained, and the remaining segments were replaced with silence to generate H- or M-only noise-suppressed sentences. Waveforms of H- and M-only noise-suppressed (masked by SSN at −10 dB SNR, and processed by Wiener filtering) sentences are shown in Fig. 2. Spectrograms of H- and M-only noise-suppressed sentences are shown in Fig. 3. Before noise-masking, noise-suppression, and silence-replacement processing, the third condition, i.e., M-only-adjusted, was first processed by computing the effective SNR level of concatenated H-level segments, and scaling the level of concatenated M-level segments to yield the same effective SNR level as that of concatenated H-level segments.

4. Procedure

The experimental procedure used in experiment 2 was essentially the same as that used in experiment 1. Again, in the training session in which subjects were familiarized with the testing procedure and conditions, each subject was given six lists of ten sentences (different from those used in the testing session) and allowed to read transcriptions while listening to the sentences. However, in experiment 2, each subject was exposed to a total of 24 conditions [=2 types of maskers (i.e., SSN and 6-talker babble) × 3 segmental conditions (i.e., H-only, M-only-adjusted, and M-only) × 4 signal processing conditions (i.e., KLT, logMMSE, MB, and Wiener)], which were randomized across subjects. As in experiment 1, one list of ten sentences was presented per condition, and none of the sentences was repeated across the conditions.

B. Results

Mean recognition scores of Mandarin sentences for all conditions are shown in Fig. 5. Statistical significance was determined by using the recognition score as the dependent variable, and signal processing condition (i.e., KLT, logMMSE,

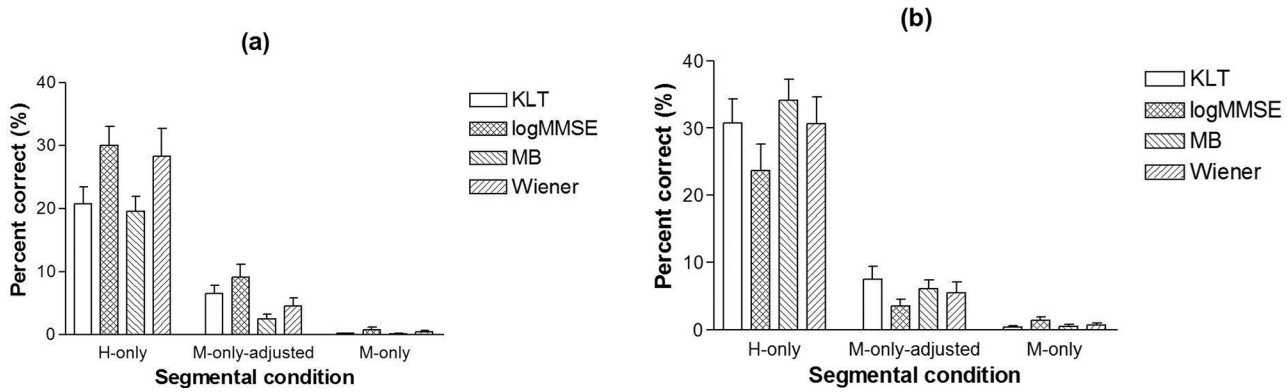

FIG. 5. Sentence recognition scores for all conditions at (a) SSN masker at −10 dB SNR and (b) 6-talker babble masker at −5 dB SNR. The error bars denote ±1 standard error of the mean.

MB, and Wiener) and segmental condition (i.e., H-only, M-only-adjusted, and M-only) as the two within-subject factors. The recognition scores were first converted to RAUs using the rationalized arcsine transform (Studebaker, 1985).

For SSN masker (at 10 dB SNR) results in Fig. 5(a), Mauchly's test indicated that the assumption of sphericity was violated for the signal processing condition [χ^2 (5) $= 11.96$, $p < 0.05$], segmental condition [χ^2 (2) $= 6.30$, $p < 0.05$] and interaction between the signal processing condition and segmental condition [χ^2 (20) $= 33.72$, $p < 0.05$]; therefore the degree of freedom was corrected using Greenhouse-Geisser estimates of sphericity (signal processing condition: $\varepsilon = 0.6$; segmental condition: $\varepsilon = 0.7$; interaction between signal processing condition and segmental condition: $\varepsilon = 0.5$). Two-way ANOVA with repeated measures indicated a significant effect of signal processing condition ($F[1.7, 17.4] = 9.60$, $p < 0.001$), segmental condition ($F[1.3, 13.3] = 153.28$, $p < 0.001$), and a non-significant interaction ($F[2.9, 29] = 1.85$, $p > 0.05$) between the signal processing condition and segmental condition. One-way ANOVA with repeated measures was conducted in each signal processing condition to further analyze the effect of segmental condition. Alpha level for statistical significance was Bonferroni corrected, and only those tests with a p-value lower than 0.008 (=0.05/6) were considered as significant. In all four signal processing conditions, the results showed significant differences in the performance between segmental conditions H-only and M-only-adjusted, segmental conditions H- and M-only, and segmental conditions M-only-adjusted and M-only ($ps < 0.001$).

For babble masker (at −5 dB SNR) results in Fig. 5(b), Mauchly's test indicated that the assumption of sphericity was violated for the segmental condition, χ^2 (2) $= 12.44$, $p < 0.05$, therefore the degree of freedom was corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.6$). Two-way ANOVA with repeated measures indicated a significant effect of segmental condition, $F[1.1, 11.4] = 170.74$, $p < 0.001$), a non-significant effect of signal processing condition ($F[3, 30] = 1.52$, $p > 0.05$), and a non-significant interaction between the signal processing condition and segmental condition ($F[6, 60] = 2.24$, $p > 0.05$). One-way ANOVA with repeated measures was conducted in each signal processing condition to further analyze the effect of segmental condition. Alpha level for statistical significance was Bonferroni corrected, and only those tests with a p-value lower than 0.008 (=0.05/6) were considered as significant. In all four signal processing conditions, the results showed significant differences in the performance between segmental conditions H-only and M-only-adjusted, segmental conditions H- and M-only, and segmental conditions M-only-adjusted and M-only ($ps < 0.001$).

IV. DISCUSSION AND CONCLUSIONS

A. Contributions of level-dependent segments to sentence intelligibility

Under noise-masked conditions, H-only sentences remained highly intelligible, with intelligibility scores above 70.0% (see Fig. 4). Many perceptual studies have suggested that the human auditory system has a remarkable capacity for understanding speech even in challenging conditions (e.g., Miller and Licklider, 1950; Warren, 1970; Powers and Speaks, 1973; Remez *et al.*, 1981). The brain may construct the meaning of interrupted sentences (e.g., H-only) based on language experience, expectations, contextual cues, and linguistic rules in a top-down processing (e.g., Chen *et al.*, 2014). This form of cognitive construction may account, at least in part, for the high intelligibility of H-only sentences observed in this study. It is noteworthy that, in this study, H-only sentences were consistently more intelligible than M-only sentences. Our finding that the intelligibility advantage of H-level sentences over M-level sentences was retained even when the inherent SNR advantage of M-level segments was obviated with the use of M-only-adjusted segments suggests that the intelligibility advantage of H-level segments over M-level segments cannot be fully attributed to a SNR effect. However, as seen in Fig. 4, the M-only-adjusted condition showed improved intelligibility relative to the M-only condition, indicating that to some extent the level of the segment mattered to intelligibility.

Recent studies have demonstrated a marked intelligibility advantage of vowel-only utterances over consonant-only utterances (e.g., Kewley-Port *et al.*, 2007; Fogerty and Kewley-Port, 2009), which is consistent with our present finding of H-only sentences, which are particularly vowel-prominent, being more intelligible than M-only sentences. Vowel segments carry many prominent acoustic cues for speech perception, such as formants and harmonic structures. Additionally relevant for a Chinese study such as this is the fact that the Mandarin Chinese language does not permit consonant clusters and has lost syllable-final plosives. In tonal languages, including Mandarin Chinese, fundamental frequency ($F0$) cues, which carry important information for lexical tone recognition, exist mainly in the vowel-dominated H-level segments. This property may have enabled the listeners in our study to get sufficient tone information for interpretation from H-only sentences, but not M-only sentences.

The presently observed H-only advantage should be considered in the context of published results on the intelligibility advantage of vowel-only sentences. Fogerty and Humes (2012) investigated three acoustic properties that are present in consonants and vowels and are informative for understanding monosyllabic words and sentences. They observed better performance for vowel-only sentences over consonant-only sentences in all processing conditions. More recently, Fogerty (2014) investigated the importance of overall segment amplitude and intrinsic segment amplitude modulation of consonants and vowels for sentence intelligibility. His results underscored the importance of vowel-envelope modulations for intelligibility of interrupted sentences. Note that the finding of the H-level intelligibility advantage over M-level does not retract the perceptual significance of M-level information, as the M-level segments carry important information (e.g., vowel-consonant transitions) to supplement the H-level segments in speech perception.

High entropy normally indicates a high potential for information. Hence, cochlea-scaled spectral entropy is used

as a measure of the relative (un)predictability (potential information) of acoustic signals. Stilp and Kluender (2010) reported a highly robust correlation between their CSE measures and listeners' intelligibility scores. While the CSE method makes no distinction between spectral changes occurring at vowels, consonants, or vowel-consonant transitions, vowel-consonant transitions are particularly well represented in M-level segments, which include prominent spectral changes that may be even more robust in the presence of noise (e.g., Chen and Loizou, 2012). Previous studies have shown that M-level segments carry important acoustic cues (e.g., vowel-consonant transition) for modeling speech intelligibility (e.g., Kates and Arehart, 2005; Ma *et al.*, 2009; Chen and Loizou, 2012). However, the present work suggests that H-level segments have a larger perceptual impact upon sentence understanding than M-level segments under both noise-masked and noise-suppressed conditions. Hence, further research is warranted to examine the functional roles of level-dependent (e.g., H- and M-level) segments in speech recognition and intelligibility prediction.

B. Contributions of noise-suppressed level-dependent segments to sentence intelligibility

Many studies have attempted to address the challenge of why traditional noise-suppression algorithms fail to improve speech intelligibility in NH listeners (e.g., Loizou and Kim, 2011). In the present work, H-level noise-suppressed segments were found to be much more intelligible than M-level noise-suppressed segments across all of the noise-suppression algorithms examined. We speculate that H-level segments, due to their high intensity, might have less distortion than M-level segments. To test this hypothesis, we applied an objective metric to assess the amount of signal distortion (relative to clean speech signal) contained in H- and M-level segments. For this purpose, we chose the normalized covariance metric (NCM) measure, which is a speech-based speech transmission index (TI) (Goldsworthy and Greenberg, 2004). The NCM is computed based on the apparent SNR between processed (i.e., noise-suppressed in this study) and probe (i.e., clean) speech signals. Its values are within the range of 0 to 1, such that the closer the value is to 1, the less distorted the noise-suppressed signal is relative to the reference clean speech signal (see the Appendix for a detailed explanation of NCM computation). The NCM values computed for the eight noise-suppressed conditions (i.e., H- and M-level-only) in this study are shown in Fig. 6 (NCM values shown are averages of ten values computed for ten sentences). The NCM values computed for H-level segments are higher than those computed for M-level segments, supporting the notion that noise-suppressed H-level segments contain less distortion than noise-suppressed M-level segments. This distortion difference may account, perhaps in part, for the intelligibility difference between H- and M-only noise-suppressed sentences.

Most noise-suppression algorithms modulate a gain function upon the input speech signal. Because consonants have a lower amplitude (or lower effective SNR level) and are acoustically more noise-like than vowels, noise-suppression processing may apply a small gain to suppress weak consonants, producing large-magnitude distortions in M-level segments. To examine the effect of effective SNR level on the intelligibility of H- and M-level sentences, a new condition (i.e., M-only-adjusted) was designed in experiment 2 to make M-level segments have the same effective SNR as H-level segments. Our experiment 2 results showed that although the intelligibility scores for the M-only-adjusted condition were better than those for the M-only condition, they were still significantly lower than those of the H-only condition regardless of which of the four noise-suppression algorithms had been employed. Hence, similar to our findings in experiment 1, we found in experiment 2 that the intelligibility advantage of H-only noise-suppressed segments over M-only noise-suppressed segments could not be fully attributed to a differential SNR level.

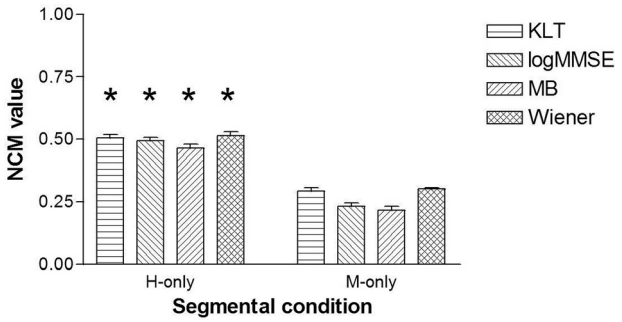

FIG. 6. NCM values for noise-suppressed conditions with SSN masker. The error bars denote ±1 standard error of the mean, and the asterisk indicates that the NCM value at H-only condition is significantly larger than that at M-only condition.

In this study, we observed a consistent perceptual advantage for H-level segments under both noise-masked and noise-suppression conditions as well as with SNR-adjusted stimuli. Hence, our findings indicate that the intelligibility advantage of H-level segments may be dependent upon other factors beyond signal processing method and SNR difference. For example, linguistic composition might account for the perceptual advantage of H-level segments in noise-masked and noise-suppressed conditions. These findings may guide the design of future single-channel noise-suppression algorithms. In particular, research efforts should be devoted to discerning the particular perceptual contributions of H-level segments and diminishing the potentially detrimental influence of heavily distorted M-level segments on speech intelligibility.

C. Limitations of the present work

Language type affects noise-masking and noise-suppression performance. The present work assessed level-dependent perceptual contributions in a tonal language, i.e., Mandarin. The property of a vowel system may shed light on the perceptual contributions of level-dependent segments in a specific language. English has an extensive vowel system, and its H-level segments may provide a voicing cue for some consonants and nasality for nasals as well as transitions. For Mandarin, its monophthongs are not as extensive as those in English, but its diphthongs and nasalized vowels are quite prominent. Together with lexical tones, H- and M-level segments are again important for intelligibility performance in

Mandarin. However, it is unclear about the perceptual importance of level-dependent segments in languages with a very restricted vowel system, such as Japanese and Spanish, which warrants further investigation.

This study examined the perceptional contributions of H- and M-level segments defined by a relative RMS-level based segmentation method. The relationship among the three segmentation methods (i.e., vowel-consonant, entropy based, and RMS-level based) has not been studied extensively, mainly because they were proposed in different fields (linguistics, information theory, and acoustics, respectively). Although it is unclear which segmentation method is best, we chose to use RMS-level based segmentation mainly because of its relative ease to implement (compared with the vowel-consonant based segmentation) and almost equal durations of H- and M-level segments. Further study is warranted to compare the perceptual contributions of speech regions segmented with different segmentation methods.

The present work compared the relative perceptual contributions of H-level segments and M-level segments to sentence intelligibility in noise-masked and noise-suppressed conditions. The following conclusions can be drawn:

(1) H-level segments carry more intelligibility information than M-level segments under both noise-masked and noise-suppressed conditions. This intelligibility advantage may be partially attributed to the fact that H-level segments include predominantly vowels and semivowels, and is consistent with previous findings demonstrating an advantage of vowels over consonants for speech perception (e.g., Fogerty and Kewley-Port, 2009; Chen *et al.*, 2013).
(2) The intelligibility advantage of H-level segments over M-level segments cannot be fully explained by intensity or effective SNR difference. H-only sentences remain much more intelligible than M-only sentences even when the SNRs of H- and M-level segments are matched.
(3) However, the higher SNR of H-level segments relative to that of M-level segments enables H-level segments to be less influenced by distortion from noise-suppression processing. The intelligibility deficiency of M-only noise-suppressed sentences might be related to the distortion from noise-suppression processing. Hence, the future design of noise-suppression algorithms could be targeted to alleviate distortion within M-level segments.

ACKNOWLEDGMENTS

This work was supported by the National Nature Science Foundation of China (Grant Nos. 61571213 and 31271056), the Basic Research Foundation of Shenzhen (Grant Nos. JCYJ20160429191402782 and JCYJ20160324163759208), and Shenzhen Medical Engineering Laboratory for Human Auditory-equilibrium Function. The authors are grateful to the two reviewers who provided valuable feedback that significantly improved the presentation of the manuscript.

APPENDIX

This appendix gives the procedure to compute the NCM measure (Holube and Kollmeier, 1996; Goldsworthy and Greenberg, 2004). The stimuli are first bandpass filtered into N bands spanning the signal bandwidth (300–7600 Hz in this study). The envelope of each band is extracted by using Hilbert transform and then downsampled to $2f_{\text{cut}}$ Hz, which limits the envelope modulation rate to f_{cut} Hz ($f_{\text{cut}} = 25$ Hz in this study). Note that an anti-aliasing low-pass filter is used prior to downsampling to eliminate aliasing artifacts. Let $x_i(t)$ and $y_i(t)$ be the downsampled envelope in the ith band of the clean signal and the processed signal, respectively. The normalized covariance in the ith frequency band is computed as

$$\rho_i = \frac{\sum_t \left(x_i(t) - \bar{x}_i\right)\left(y_i(t) - \bar{y}_i\right)}{\sqrt{\sum_t \left(x_i(t) - \bar{x}_i\right)^2}\sqrt{\sum_t \left(y_i(t) - \bar{y}_i\right)^2}}, \tag{A1}$$

where $\bar{x}_i$ and $\bar{y}_i$ are the mean values of $x_i(t)$ and $y_i(t)$, respectively. The SNR in each band is computed as

$$\text{SNR}_i = 10\log_{10}\left(\frac{\rho_i^2}{1-\rho_i^2}\right), \tag{A2}$$

and subsequently limited to the range of $[-15,15]$ dB in this study. The TI in each band is computed by linearly mapping the SNR values between 0 and 1 following:

$$\text{TI}_i = (\text{SNR}_i + 15)/30. \tag{A3}$$

Finally, the all transmission indices are averaged to produce the NCM index

$$\text{NCM} = \frac{\sum_{i=1}^{N} \text{TI}_i \times w_i}{\sum_{i=1}^{N} w_i}, \tag{A4}$$

where $W = (w_1 \cdots w_i \cdots w_N)^T$ denotes the weight vector applied to the transmission-index TI_i of N bands. There are several methods for choosing the weight vector W in Eq. (A4), with the most common being the articulation index weights (ANSI, 1997).

ANSI (**1997**). ANSI-S3.5-1997, *Methods for Calculation of the Speech Intelligibility Index* (American National Standards Institute, New York).
Benard, M. R., and Başkent, D. (**2014**). "Perceptual learning of temporally interrupted spectrally degraded speech," J. Acoust. Soc. Am. **136**, 1344–1351.
Chen, F., and Loizou, P. (**2012**). "Contributions of cochlea-scaled entropy and consonant-vowel boundaries to prediction of speech intelligibility in noise," J. Acoust. Soc. Am. **131**, 4104–4113.
Chen, F., Wong, L. L. N., and Hu, Y. (**2014**). "Effects of lexical tone contour on Mandarin sentence intelligibility," J. Speech Lang. Hear. Res. **57**, 338–345.
Chen, F., Wong, L. L. N., and Wong, Y. W. (**2013**). "Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility," J. Acoust. Soc. Am. **134**, EL178–EL184.

Cole, R., Yan, Y., Mak, B., Fanty, M., and Bailey, T. (**1996**). "The contribution of consonants versus vowels to word recognition in fluent speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 853–856.
Ephraim, Y., and Malah, D. (**1985**). "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Process **33**, 443–445.
Fogerty, D. (**2014**). "Importance of envelope modulations during consonants and vowels in segmentally interrupted sentences," J. Acoust. Soc. Am. **135**, 1568–1576.
Fogerty, D., and Humes, L. E. (**2012**). "The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences," J. Acoust. Soc. Am. **131**, 1490–1501.
Fogerty, D., and Kewley-Port, D. (**2009**). "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility," J. Acoust. Soc. Am. **126**, 847–857.
Goldsworthy, R., and Greenberg, J. (**2004**). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," J. Acoust. Soc. Am. **116**, 3679–3689.
Holube, I., and Kollmeier, K. (**1996**). "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," J. Acoust. Soc. Am. **100**, 1703–1715.
Hu, Y., and Loizou, P. (**2003**). "A generalized subspace approach for enhancing speech corrupted by colored noise," IEEE Trans. Speech Audio Process. **11**, 334–341.
Hu, Y., and Loizou, P. (**2007**). "A comparative intelligibility study of single-microphone noise reduction algorithms," J. Acoust. Soc. Am. **122**, 1777–1786.
Kamath, S., and Loizou, P. (**2002**). "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. IV–4164.
Kates, J., and Arehart, K. (**2005**). "Coherence and the speech intelligibility index," J. Acoust. Soc. Am. **117**, 2224–2237.
Kewley-Port, D., Burkle, T. Z., and Lee, J. H. (**2007**). "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," J. Acoust. Soc. Am. **122**, 2365–2375.
Kim, G., and Loizou, P. (**2011**). "Gain-induced speech distortions and the absence of intelligibility benefit with existing noise-reduction algorithms," J. Acoust. Soc. Am. **130**, 1581–1596.
Li, J., Yang, L., Zhang, J., Yan, T., Hu, Y., Akagi, M., and Loizou, P. (**2011**). "Comparative intelligibility investigation of single-channel noise-reduction algorithms for Chinese, Japanese, and English," J. Acoust. Soc. Am. **129**, 3291–3301.
Loizou, P. (**2007**). *Speech Enhancement: Theory and Practice* (CRC, Boca Raton, FL).
Loizou, P., and Kim, G. (**2011**). "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," IEEE Trans. Audio, Speech, Lang. Process. **19**, 47–56.
Ma, J., Hu, Y., and Loizou, P. (**2009**). "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," J. Acoust. Soc. Am. **125**, 3387–3405.
Miller, G. A., and Licklider, J. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**, 167–173.
Mittal, U., and Phamdo, N. (**2000**). "Signal/noise KLT based approach for enhancing speech degraded by colored noise," IEEE Trans. Speech Audio Proc. **8**, 159–167.
Powers, G. L., and Speaks, C. (**1973**). "Intelligibility of temporally interrupted speech," J. Acoust. Soc. Am. **54**, 661–667.
Powers, G. L., and Wilcox, J. C. (**1977**). "Intelligibility of temporally interrupted speech with and without intervening noise," J. Acoust. Soc. Am. **61**, 195–199.
Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (**1981**). "Speech perception without traditional speech cues," Science **212**, 947–949.
Scalart, P., and Filho, J. (**1996**). "Speech enhancement based on a priori signal to noise estimation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 629–632.
Sohn, J., Kim, N., and Sung, W. (**1999**). "A statistical model based voice activity detection," IEEE Signal Process. Lett. **6**, 1–3.
Stilp, C. E., and Kluender, K. R. (**2010**). "Cochlear-scaled entropy, not consonants, vowels or time, best predicts speech intelligibility," Proc. Natl. Acad. Sci. U.S.A. **107**, 12387–12392.
Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455–462.
Warren, R. M. (**1970**). "Perceptual restoration of missing speech sounds," Science **167**, 392–393.
Wong, L. L., Soli, S. D., Liu, S., Han, N., and Huang, M. W. (**2007**). "Development of the Mandarin Hearing in Noise Test (MHINT)," Ear Hear. **28**, 70S–74S.