Factors affecting the intelligibility of high-intensity-level-based speech

Danying Xu, Fei Chen, Fan Pan, and Dingchang Zheng

Citation: The Journal of the Acoustical Society of America **146**, EL151 (2019); doi: 10.1121/1.5122190
View online: https://doi.org/10.1121/1.5122190
View Table of Contents: https://asa.scitation.org/toc/jas/146/2
Published by the Acoustical Society of America

Factors affecting the intelligibility of high-intensity-level-based speech

Danying Xu and Fei Chen[a)]
Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China
11749122@mail.sustech.edu.cn, fchen@sustech.edu.cn

Fan Pan
College of Electronics and Information Engineering, Sichuan University, Chengdu, China
panfan@scu.edu.cn

Dingchang Zheng
Faculty of Health, Education, Medicine and Social Care, Anglia Ruskin University, Chelmsford, CM1 1SQ, United Kingdom
dingchang.zheng@anglia.ac.uk

Abstract: The present work examined factors affecting the intelligibility of high-intensity-level-based speech. Mandarin sentences were processed to contain only high-intensity segments confined by a 5-dB selected intensity range (SIR), with other segments replaced by noise. The processed stimuli were presented to normal-hearing listeners to recognize. The greatest intensity density occurred in the SIR with an upper boundary 3 dB below the peak intensity level, and this SIR yielded the highest intelligibility score in quiet. The SIR with the upper boundary at the peak intensity level yielded better intelligibility performance under noisy conditions, due largely to the relatively high effective signal-to-noise ratio.

[DDO]
Date Received: April 9, 2019 **Date Accepted:** July 29, 2019

1. Introduction

Studies of the perceptual impacts of various speech segments are currently ongoing (e.g., Cole *et al.*, 1996; Kewley-Port *et al.*, 2007; Fogerty and Kewley-Port, 2009; Stilp and Kluender 2010; Chen *et al.*, 2013; Chen *et al.*, 2015; Guan *et al.*, 2016; Shu *et al.*, 2016; Oxenham *et al.*, 2017; Aubanel *et al.*, 2018). The outcomes of these studies can provide us with important guidelines for the design of speech processing algorithms. The noise-replacement paradigm has long been used to study the relative perceptual impacts of different speech segments. Intact speech is split into regions using various segmentation methods, most commonly vowel-consonant (e.g., Cole *et al.*, 1996; Kewley-Port *et al.*, 2007), entropy-based (e.g., Stilp and Kluender, 2010), and level-dependent segmentation (e.g., Kates and Arehart, 2005).

Vowel-consonant segmentation has been employed in many studies to compare the relative perceptual importance of vowels and consonants. The results have consistently shown that vowels carry more intelligibility information than consonants do across English (e.g., Cole *et al.*, 1996; Kewley-Port *et al.*, 2007; Fogerty and Kewley-Port, 2009) and Chinese sentences and isolated words (Chen *et al.*, 2013; Chen *et al.*, 2015). In addition, the intelligibility ratio of vowel sentences (with consonants replaced by noise) to consonant sentences (with vowels replaced by noise) has been found to differ between English (2:1) (e.g., Cole *et al.*, 1996; Kewley-Port *et al.*, 2007) and Mandarin (3:1) (e.g., Chen *et al.*, 2013). However, the accurate identification of boundaries between vowel and consonant segments is difficult, even with sophisticated computation. Stilp *et al.* (2010) used cochlea-scaled entropy (CSE) to study speech-segment perceptual importance, arguing that CSE, rather than vowels, consonants, or time, best predicted speech intelligibility. Like vowel-consonant segmentation, CSE-based segmentation requires sophisticated computation to obtain speech segments of interest.

Level-dependent segmentation has also attracted considerable interest, due mainly to its relative ease of implementation. Kates and Arehart (2005) used this method to predict speech intelligibility by classifying speech segments into three levels

[a)]Author to whom correspondence should be addressed.

with the following cutoffs: the overall root-mean-square (RMS) level (in dB), 10 dB below the RMS level, and 30 dB below the RMS level. Figure 1(b) shows an example of the intensity contour of a Mandarin sentence, normalized to the overall intensity level. Traditionally, segments with intensities above 0 dB (relative to the overall intensity level) are referred to as high (H) level, those with intensities between the RMS and RMS −10 dB are defined as middle (M) level, and those between the RMS −10 dB and RMS −30 dB are considered to be low (L) level. These three regions have been found to have different phonetic constitutions: H-level segments are mainly vowels, L-level segments are largely consonants, and M-level segments contain more vowel-consonant transitions (e.g., Kates and Arehart, 2005; Chen and Loizou, 2012). Early work showed that M-level results had the best predictive performance in the modeling of speech intelligibility in noise with segmental information, where the speech intelligibility data were collected with full-segment sentences in noise (e.g., Kates and Arehart, 2005; Ma *et al.*, 2009; Chen and Loizou, 2012). Using the noise-replacement paradigm, Guan *et al.* (2016) recently showed that H-level speech segments carried more intelligibility information than did M-level segments, and that this advantage was not due to the high intensity levels. Shu *et al.* (2016) found no significant difference in the intelligibility performance of segmentally processed Mandarin Chinese stimuli containing the same amounts of H-level or high-CSE segments. Oxenham *et al.* (2017) confirmed that CSE correlated strongly with the speech level. They also noted that the importance of speech segments to overall intelligibility was best predicted by their relative intensity, not by CSE. Aubanel *et al.* (2018) recently examined the superiority of CSE over traditional linguistic and psychoacoustic characteristics, and found that CSE was related closely to intensity and captured similar speech regions.

Early studies showed the advantages of the use of level-dependent segmentation to study the relative importance of various speech segments, including its ease of implementation and performance comparable to that of other sophisticated segmentation methods. In addition, they showed that H-level segments have particular importance due to their strong correlation with speech intelligibility. The aim of this work was to further study the relative importance of these H-level segments, motivated by the following two points. First, traditional three-level segmentation does not enable examination of the perceptual impacts of segments in narrow selected intensity ranges (SIRs), i.e., within the 10-dB intensity range of H-level segments (Kates and Arehart, 2005) [Fig. 1(b)]. In other words, although segments of interest are easy to select with level-based segmentation, the number of segments within a narrow SIR is unclear. If the distribution of H-level segments is not uniform across the 10-dB H-level intensity range, identification of the narrow intensity range containing more segments (characterized by intensity density in this work) and assessment of their perceptual impacts are important. Second, peak intensity regions (within the 10-dB H-level intensity range) have attracted special interest in speech perception studies, due partially to their ease

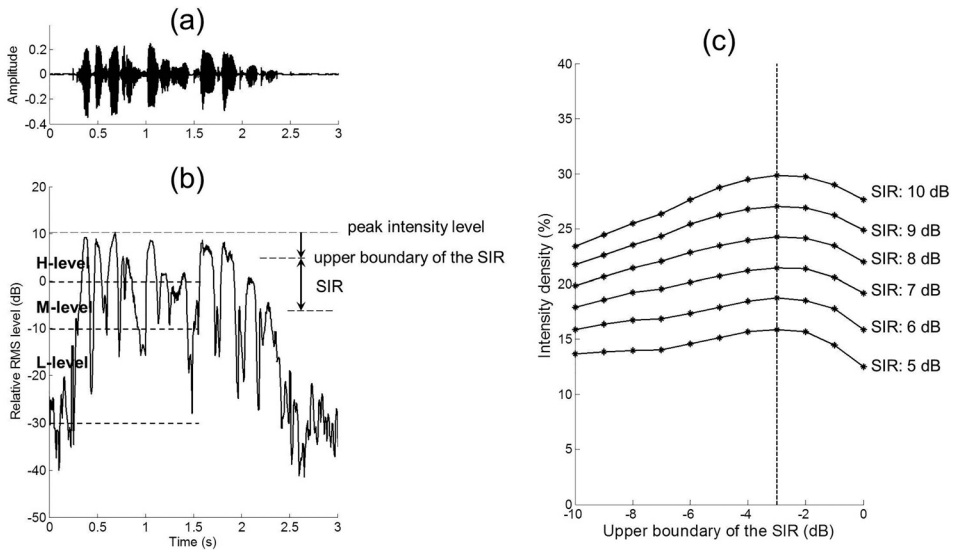

Fig. 1. An example of a speech waveform (a) and its relative root-mean-square (RMS)-level curve (b). (c) The intensity density distribution computed with a fixed selected intensity range (SIR) and the upper boundary of the SIR relative to the peak intensity level. The dashed line indicates the upper boundary of the SIR yielding the greatest intensity density.

of selection and favorable signal-to-noise ratios (SNRs) under noisy conditions. Thus, investigation of how noise interference affects the perceptual contributions of peak intensity regions is warranted.

2. Methods

2.1 Subjects and materials

Ten subjects (seven males and three females) with normal hearing (i.e., pure-tone thresholds not exceeding the 20-dB hearing level at octave frequencies of 125–8000 Hz in both ears) participated in this experiment. All subjects were native speakers of Mandarin Chinese and were paid for their participation. The speech material consisted of sentences extracted from the Mandarin Hearing in Noise Test (MHINT) database (Wong *et al.*, 2007). The MHINT corpus comprises 24 lists of 10 sentences each, with 10 keywords per sentence. The original recordings of MHINT sentences (Wong *et al.*, 2007) were used in this study. All sentences were spoken by a male native Mandarin Chinese speaker at a fundamental frequency of 75–180 Hz, and were recorded at a sampling rate of 16 kHz.

The effect of noise masking on speech perception is commonly represented by two mechanisms: energetic masking by steady-state noise, and informational masking by other speech characteristics (Carhart *et al.*, 1967; Watson, 2005). Hence, two types of masker signal were used in the present work to corrupt the sentences: steady-state speech-spectrum-shaped noise (SSN) and two-talker babble. To generate the SSN masker, a finite impulse response filter was designed based on the average spectrum of the MHINT sentences, and white noise was filtered and scaled to the sentences' long-term average spectrum and level. The two-talker babble masker contained speech from two equal-level interfering male native Mandarin talkers, and the resulting noise was intelligible to listeners. Specially, the two interfering talkers clearly read two different short stories (about 2 min) at a normal speaking rate, but were not synchronized in onsets of their readings. The two recorded speech signals were adjusted to have equal intensity level and mixed to generate the two-talker babble masker. Note that the original noise signal used in the development of the MHINT was not applied in this study. 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 0-dB input SNR levels for each masker condition. The input SNR level was chosen based on known performance from a pilot study with full-segment sentences.

2.2 Signal processing

Relative RMS level-based segmentation was performed by dividing speech into short-term (16-ms in this study) segments and classifying each segment into one of three regions according to its relative RMS intensity (Kates and Arehart, 2005). We adopted the threshold levels proposed by Kates and Arehart (2005): 0, –10, and –30 dB. Figure 1(b) shows an example of segmentation of a speech waveform into H-, M-, and L-levels based on these RMS thresholds, wherein the RMS-level waveform was normalized relative to the overall RMS level.

This work studied the perceptual contributions of high-intensity segments in the SIR shown in Fig. 1(b). The upper boundary of the SIR was set relative to the peak intensity level. For instance, a –3-dB boundary is 3 dB below the peak intensity level. For this study, intensity density was defined as the number of segments within an SIR and was calculated using all 240 MHINT sentences. SIR values were chosen at 1-dB intervals from 10 to 5 dB, and the upper boundaries were chosen at 1-dB intervals from 0 dB (i.e., the peak intensity level) to –10 dB. Figure 1(c) shows the distribution of intensity density, which is not uniform within the high-intensity region. Specifically, with a fixed SIR value (e.g., 5 dB), the SIR with an upper boundary of –3 dB, not 0 dB, has the most segments. This pattern is consistent across all SIR values examined.

As the aim of the present study was to investigate which narrow SIR within the 10-dB intensity range of H-level segments (defined in Kates and Arehart, 2005) carried more perceptual impact, this work chose a small SIR (i.e., 5 dB) and three values of the upper boundary of the 5-dB SIR (i.e., 0, –3, and –6 dB), yielding three narrow SIRs (i.e., 0 to –5 dB, –3 to –8 dB, and –6 to –11 dB, relative to the peak intensity level) within the 10-dB intensity range of H-level segments. Therefore, the present work was carried out in the context of the H-level intensity region. Note that the third narrow SIR (i.e., –6 to –11 dB, relative to the peak intensity level) contained some segments (i.e., –10 to –11 dB, relative to the peak intensity level) within the M-level intensity region.

For the clean sentences, noise-replaced stimuli were created based on a selected set of segments confined by a 5-dB SIR with an upper boundary of 0, –3, or –6 dB. The segments within the SIR were preserved, and the remaining segments were replaced by SSN scaled to 16 dB below the level of the intact speech waveform (Fogerty and Kewley-Port, 2009). Because the purpose of this experiment was to assess which segments (confined by an SIR and its upper boundary) of the original speech signal contained more intelligibility information when the speech signal was mixed with interfering noise, noise was added to a clean sentence (at a 0-dB input SNR) prior to the replacement of out-level segments with noise. The noise-masked sentences were also edited such that the selected segments (those selected for the clean sentences) were retained and the remaining segments were replaced with noise. The sound examples of the processed sentences for different conditions will be provided upon request.

2.3 Procedure

The experiment was conducted in a soundproof booth. Test stimuli were played to the participants binaurally through Sennheiser (HD 25 II) circumaural headphones (Sennheiser, Germany) at a comfortable listening level of 65 dB sound pressure level. Each sentence could be repeated twice. Before the experiment, participants completed a practice session consisting of 30 sample sentences (10 sentences for each condition of quiet, SSN at 0-dB input SNR, and two-talker at 0-dB input SNR, and all three conditions were with a 0-dB upper boundary of a 5-dB SIR) that differed from those used in the experimental session. During the experimental session, participants orally repeated all keywords that they could recognize. Each participant was presented with a total of nine conditions [three noise conditions (quiet, SSN at 0-dB input SNR, and two-talker at 0-dB input SNR) × three upper boundaries of a 5-dB SIR (0, –3, and –6 dB)]. One list of 10 sentences was used per test condition, and no sentence was repeated across conditions. The test condition order was randomized across subjects, and subjects were given a 5-min break after every 30 min of testing. 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 that condition.

3. Results

Statistical significance was determined using the percent recognition score as the dependent variable, and the noise condition and upper boundary of the 5-dB SIR as within-subject factors. Recognition scores were converted to rational arcsine units using the transform of Studebaker (1985). Figure 2(a) shows the mean sentence recognition results for all test conditions. Two-way repeated-measures analysis of variance revealed significant effects of the noise condition ($F_{2,18} = 88.91$, $p < 0.001$) and

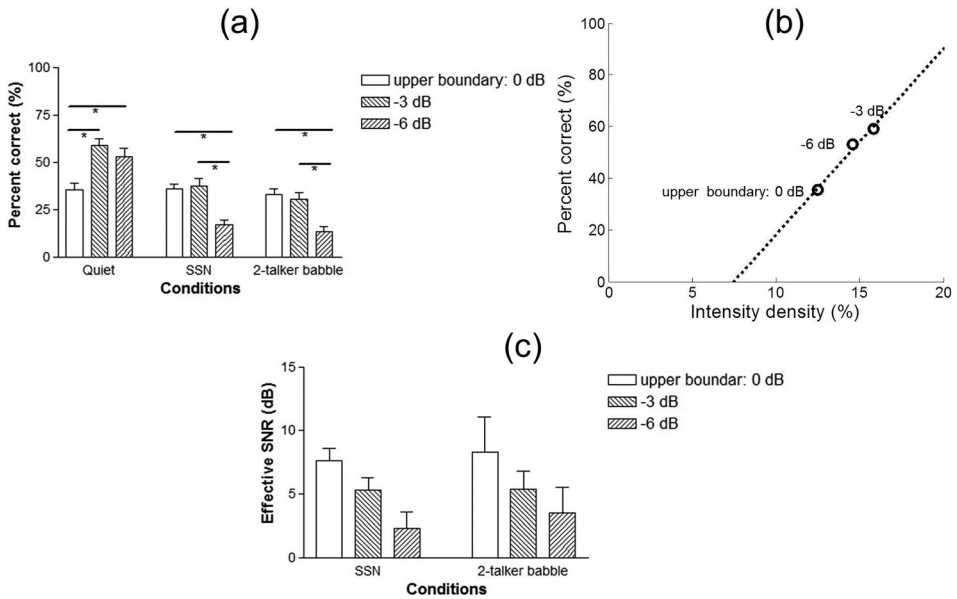

Fig. 2. (a) Mean recognition scores under all test conditions. The error bars denote standard errors of the mean, and the asterisks denote significant differences between paired recognition scores ($p < 0.017$). (b) Scatter plot of intensity densities and recognition scores for the three test conditions in quiet. (c) Effective SNR levels under all noisy conditions. The error bars denote standard deviation.

upper boundary of the SIR ($F_{2,18} = 25.40$, $p < 0.001$), and significant interaction between them ($F_{4,36} = 12.66$, $p < 0.05$).

Post hoc pairwise comparison of conditions with the same noise background was performed to analyze the effect of the upper boundary of the 5-dB SIR. The Bonferroni-corrected statistical significance level was set at $p < 0.017$ ($\alpha = 0.05$). Under the quiet condition, scores were significantly higher for the –3-dB and –6-dB boundaries than for the 0-dB boundary ($p < 0.017$). Under the SSN and two-talker babble conditions, scores were significantly higher for the 0-dB and –3-dB boundaries than for the –6-dB boundary ($p < 0.017$).

Post hoc pairwise comparison of conditions with the same upper boundary of the 5-dB SIR was performed to analyze the effect of noise background. The Bonferroni-corrected statistical significance level was set at $p < 0.017$ ($\alpha = 0.05$). Under the condition of 0 dB upper boundary, there was no significant ($p > 0.017$) difference among the scores of the three noise conditions. Under the condition of –3-dB upper boundary, there was significant ($p < 0.017$) difference among the scores of the three noise conditions. Similarly, under the condition of –6-dB upper boundary, there was significant ($p < 0.017$) difference among the scores of the three noise conditions.

Figure 2(b) shows a scatter plot of intensity densities and recognition scores for the three test conditions (i.e., three upper boundary values of 0, –3, and –6 dB of a 5-dB SIR) in quiet. A high coefficient of correlation between them (0.99) was obtained. Note that under the two noise conditions, the correlation coefficients between intensity densities and recognition scores were low, i.e., 0.07 under the SSN condition and 0.25 under the two-talker babble condition, respectively.

4. Discussion and conclusions

Many early studies showed the perceptual importance of H-level speech segments under quiet, noisy, and noise-suppressed conditions (e.g., Cole *et al.*, 1996; Guan *et al.*, 2016). The present work extended this research by examining the perceptual contributions of speech segments within a narrow SIR (i.e., 5 dB) in the H-level region (Kates and Arehart, 2005). First, the intensity density distribution within a narrow SIR range was analyzed. Intensity density was greatest in the SIR with an upper boundary 3 dB below the peak intensity level, not in the peak intensity region [Fig. 1(c)]. This finding suggests that the number of greatest-intensity segments is smaller than that of segments with intensity at 3 dB below the peak intensity level within a fixed SIR (e.g., 5 dB). Hence, when applying level-based segmentation, researchers should take into account the potential for substantial variability in the number of segments with the same intensity ranges. In addition, as the intensity density is defined as the number of speech segments within a SIR, the SIR with a higher intensity density has more speech segments or a longer segment duration. This variance in intensity density or duration was correlated strongly with recognition scores for H-level sentences, as shown in this study [Fig. 2(b)]. Hence, to some extent, these findings suggest that under a special constraint, e.g., the same intensity range, intensity density is a good predictor of speech recognition performance under quiet condition. For instance, the SIR ranging from –3 to –8 dB had a longer duration than that ranging from 0 to –5 dB, and yielded better recognition performance under quiet condition [59.0% vs 35.4%; Fig. 2(b)] in this study. However, such correlation between intensity densities and recognition scores was low under noisy conditions, suggesting that intensity density is not a good predictor of speech recognition performance under noisy conditions.

In addition, early work reported that the intelligibility of MHINT sentences only preserving H-level speech segments (i.e., a 10-dB SIR with the upper boundary at the peak intensity level) was around 90.8% under quiet condition (Chen and Wong, 2013). Figure 2(a) shows that under all quiet conditions with a 5-dB SIR, and their intelligibility scores range from 35.4% to 59.0%. Especially, with the upper boundary at the peak intensity level, the intelligibility score difference between the 10-dB- and 5-dB-SIR conditions (i.e., 90.8% vs 35.4%) indicates the perceptual benefit of a larger SIR.

Although the SIR ranging from –3 to –8 dB provided the best intelligibility information under quiet condition in this study, whether test stimuli in other languages would yield the same results is not known. Mandarin Chinese, used in this study, differs from English in several ways, particularly in the perceptual importance of vowels (e.g., Chen *et al.*, 2013; Fogerty and Chen, 2014). Vowels in Mandarin Chinese have much longer durations than English vowels and provide important perceptual cues, such as fundamental frequency contours. Hence, intensity densities, and thus the

perceptual importance of H-level segments, might differ between English and Mandarin. Thus, further investigation is warranted on this language effect.

This work also showed that the pattern of perceptual contribution differed between quiet and noisy conditions. Speech recognition performance for the SIR ranging from 0 to –5 dB did not differ significantly between the quiet and noisy conditions [see Fig. 2(a)]. However, performance declined significantly under the noisy condition for the SIR ranging from –3 to –8 dB, i.e., from 59.0% under quiet condition to 37.4% under the SSN condition and 30.3% under the two-talker babble condition [Fig. 2(a)]. This result might be partially attributable to the effective SNR level to different intensity range. In this study, the effective SNR was defined as the ratio (in log scale) between the energy of the selected clean speech segments confined by a SIR and that of the selected noise segments confined by the same SIR. Figure 2(c) shows the effective SNR levels for all noisy conditions. The SIR ranging from 0 to –5 dB was less affected by noise interference, due to its high intensity and SNR levels. Although this range had the least intensity density [Fig. 2(b)] among the three test conditions, the importance of the effective SNR level increased in noisy conditions. Hence, recognition performance remained high in noise. On the other hand, although recognition scores were highest under quiet conditions for the SIR ranging from –3 to –8 dB, the effective SNR level was lower for this SIR than for that ranging from 0 to –5 dB under noisy conditions. The intelligibility advantage under quiet conditions was compromised by the low effective SNR level, and there was no significant difference between their recognition scores under noisy condition, e.g., 35.8% and 37.4% for the SIRs ranging from 0 to –5 dB and from –3 to –8 dB, respectively, under the SSN condition. Similarly, for the SIR ranging from –6 to –11 dB, which had lower speech intensity and SNR levels, recognition scores were lowest under noisy conditions despite the high intensity density (or long duration).

The results of this work have useful implications. First, they can provide insight for the design of adaptive dynamic range optimization (ADRO) for hearing-impaired listeners using hearing aids or cochlear implants. ADRO is an amplification strategy that uses digital signal processing techniques to improve the audibility, comfort, and intelligibility of sounds for people who use cochlear implants and/or hearing aids (Blamey, 2005). The essential processing of ADRO is to use statistical analysis to select the most information-rich section of the input dynamic range, and then applies fuzzy logic rules to control the gain so that the selected section of the dynamic range is presented at an audible and comfortable level. Knowing which dynamic range within the 10-dB intensity range of H-level segments is most information-rich can provide cues on the design of ADRO strategy. Second, this work showed that the intensity density peaked at 3 dB below the peak intensity level, and the SIR with the greatest intensity density yielded the best recognition performance under quiet conditions. However, under noisy conditions, the perceptual contribution of peak intensity regions increased, due mainly to the high effective SNR level (or smaller effect of noise interference). Hence, under noisy conditions, peak intensity regions contain more reliable information within the H-level of the 10-dB intensity range. Consistently, the idea of employing the most energetic region for speech perception was developed in many early studies, e.g., the "N-out-of-M" strategy in the speech processing for cochlear implants (see review in Loizou, 1999). In this strategy, the speech signal was split into M frequency bands, and the processor selected the N (<M) envelope outputs with the greatest energy to modulate the electrical pulse trains and stimulate the auditory nerves.

In conclusion, the present work assessed factors affecting the intelligibility of H-level segments. High-intensity segments were confined by a narrow SIR, and the SIR with an upper boundary 3 dB below the peak intensity level had the greatest intensity density. Under quiet conditions, the SIR starting 3 dB below the peak intensity level, rather than at this level, yielded the best speech recognition performance. In addition, under noisy conditions, the SIR starting at the peak intensity level was less affected by noise interference. Therefore, although the SIR starting at the peak intensity level had a much shorter duration, it yielded better speech recognition performance.

Acknowledgments

This work was supported by the Basic Research Foundation of Shenzhen (Grant No. JCYJ20170817110841907), the National Natural Science Foundation of China (Grant No. 61571213), and the Research Foundation of Department of Science and Technology of Guangdong Province (Grant No. 2018A050501001).

References and links

Aubanel, V., Cooke, M., Davis, C., and Kim, J. (**2018**). "Temporal factors in cochlea-scaled entropy and intensity-based intelligibility predictions," J. Acoust. Soc. Am. **143**, EL443–EL448.

Blamey, P. J. (**2005**). "Adaptive dynamic range optimization (ADRO): A digital amplification strategy for hearing aids and cochlear implants," Trends Amplif. **9**, 77–98.

Carhart, R., Tillman, T. W., and Johnson, K. R. (**1967**). "Release of masking for speech through interaural time delay," J. Acoust. Soc. Am. **42**, 124–138.

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., and Wong, L. L. (**2013**). "Contributions of the high-RMS-level segments to the intelligibility of mandarin sentences," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 7810–7815.

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.

Chen, F., Wong, M. L. Y., Zhu, S. F., and Wong, L. L. N. (**2015**). "Relative contributions of vowels and consonants in recognizing isolated Mandarin words," J. Phon. **52**, 26–34.

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.

Fogerty, D., and Chen, F. (**2014**). "Vowel spectral contributions to English and Mandarin sentence intelligibility," in *Proceedings of 15th Annual Conference of the International Speech Communication Association (InterSpeech)*, pp. 499–503.

Fogerty, D., and Kewley-Port, D. (**2009**). "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility," J. Acoust. Soc. Am. **126**, 847–857.

Guan, T., Chu, G. X., Tsao, Y., and Chen, F. (**2016**). "Assessing the perceptual contributions of level-dependent segments to sentence intelligibility," J. Acoust. Soc. Am. **140**, 3745–3754.

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.

Loizou, P. (**1999**). "Signal processing techniques for cochlear Implants," IEEE. Eng. Med. Biol. Mag. **18**, 34–46.

Ma, J. F., 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.

Oxenham, A. J., Boucher, J. E., and Kreft, H. A. (**2017**). "Speech intelligibility is best predicted by intensity, not cochlea-scaled entropy," J. Acoust. Soc. Am. **142**, EL264–EL269.

Shu, Y. L., Feng, X. X., and Chen, F. (**2016**). "Comparing the contributions of cochlear-scaled entropy and speech level to speech intelligibility," J. Acoust. Soc. Am. **140**, EL517–EL521.

Stilp, C. E., and Kluender, K. R. (**2010**). "Cochlea-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.

Watson, C. S. (**2005**). "Some comments on informational masking," Acta Acust. **91**, 502–512.

Wong, L. L. N., 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.