CONTRIBUTIONS OF THE HIGH-RMS-LEVEL SEGMENTS TO THE INTELLIGIBILITY OF MANDARIN SENTENCES

Fei Chen and Lena L. N. Wong
Division of Speech and Hearing Sciences, The University of Hong Kong

ABSTRACT

Recent evidence suggests that segments carrying more spectral changes [e.g., consonant-vowel boundaries in the middle root-mean-square (RMS) level segments] are important to predict the intelligibility of English sentences. Nevertheless, considering the difference between Mandarin and English languages, it is hypothesized that the high-RMS-level segments might provide more perceptual information to the intelligibility of Mandarin speech. Two studies were conducted in this paper to assess the relative contributions of the high-RMS-level segments to the intelligibility of Mandarin sentences, i.e., speech perception and intelligibility prediction. Results show that 1) Mandarin sentences containing the high-RMS-level (i.e., above the overall RMS level of the whole utterance) segments are more intelligible (i.e., recognition rate up to 91%) than those with the middle-RMS-level segments; and 2) the high-RMS-level segments, which carry more vowel and tonal information, contribute more in predicting the intelligibility of Mandarin sentences in noise.

Index Terms – Speech perception, intelligibility prediction.

1. INTRODUCTION

Identifying speech segments carrying more intelligibility information is of great significance for us to understand the factors accounting for reliable speech perception, particularly in noisy environments, and subsequently design novel speech coding algorithms, e.g., for hearing assistive devices. Recently, studies based on a noise-replacement paradigm suggested a remarkable advantage of vowels versus consonants for (English) sentence intelligibility [1-4]. Cole *et al*. replaced vowel or consonant segments with speech-shaped noise, harmonic complexes or silence in sentences taken from the TIMIT corpus [1-2]. Their results showed that the vowel-only sentences (consonants replaced) led to a 2:1 intelligibility advantage over the consonant-only sentences (vowels replaced), i.e., word recognition rate 87.4% vs. 47.9%, regardless of the type of segmental replacement. This 2:1 advantage of vowels was later replicated by Kewley-Port *et al.* [3]. Stilp and Kluender recently suggested that cochlea-scaled entropy, not vowels, consonants or segment duration, best predicted (English) speech intelligibility [4]. They measured cochlea-scaled entropy in TIMIT sentences, and replaced portions of the sentences having high, medium, or low entropy with equal-level noise. Replacing low-entropy segments yielded relatively small impact on intelligibility while replacing high-entropy segments significantly reduced sentence intelligibility. A remarkably robust correlation was found with cochlea-scaled entropy predicting listeners' intelligibility scores.

However, sentence segmentation based on explicit vowel or consonant boundaries is extremely challenging to implement in practice even with using the most sophisticated phoneme detection algorithms. Chen and Loizou recently investigated the intelligibility prediction performance of intelligibility indices implemented using a relative root-mean-square (RMS) level sentence segmentation method [5]. The relative-RMS-level-based segmentation is implemented by dividing speech into short-term segments and classifying each segment into one of three regions according to its relative RMS intensity [6]. The high-RMS-level (H-level) region consists of segments at or above the overall RMS level of the whole utterance. The middle-RMS-level (M-level) region consists of segments ranging from the overall RMS level to 10 dB below (i.e., RMS–10 dB), and the low-level (L-level) region consists of segments ranging from RMS–10 dB to RMS–30 dB. Figure 1 shows an example sentence segmented into H-, M- and L-levels based on the above RMS threshold levels. For the most part, the H-level segments can be considered to be primarily vowels, the M-level segments contain vowel-consonant transitions, and the L-level segments are mostly weak consonants and pauses [5-6]. Chen and Loizou found that higher correlation of intelligibility prediction was obtained when including M-level speech segments containing a large number of consonant-vowel boundaries [5].

Though the relative RMS-level sentence segmentation method sheds light on the relative importance of the level-dependent segments to speech intelligibility, few studies have been conducted to examine the relative perceptual contributions of the H- and M-level segments to speech recognition in listening experiment. It is also worth noting that while study in [5] showed the importance of vowel-consonant transitions captured by the M-level segments in English, the contributions of vowel-consonant transitions in intelligibility prediction might shift in other languages. In tonal languages (e.g., Mandarin Chinese), for instance, the H-level segments might carry perceptually more important information, and are better predictors of intelligibility than the M-level segments. This is attributed to the increased importance of F0 information needed for reliable tone recognition in Mandarin Chinese, as the F0 cues are present in the vowel-dominated H-level segments [7].

Hence, the aim of the present paper is to examine the relative importance of the high-RMS-level segments to the intelligibility of Mandarin speech. Two studies were conducted to assess the relative contributions of the high-RMS-level segments to 1) speech perception in listening experiment, and 2) intelligibility prediction of Mandarin speech in noise.

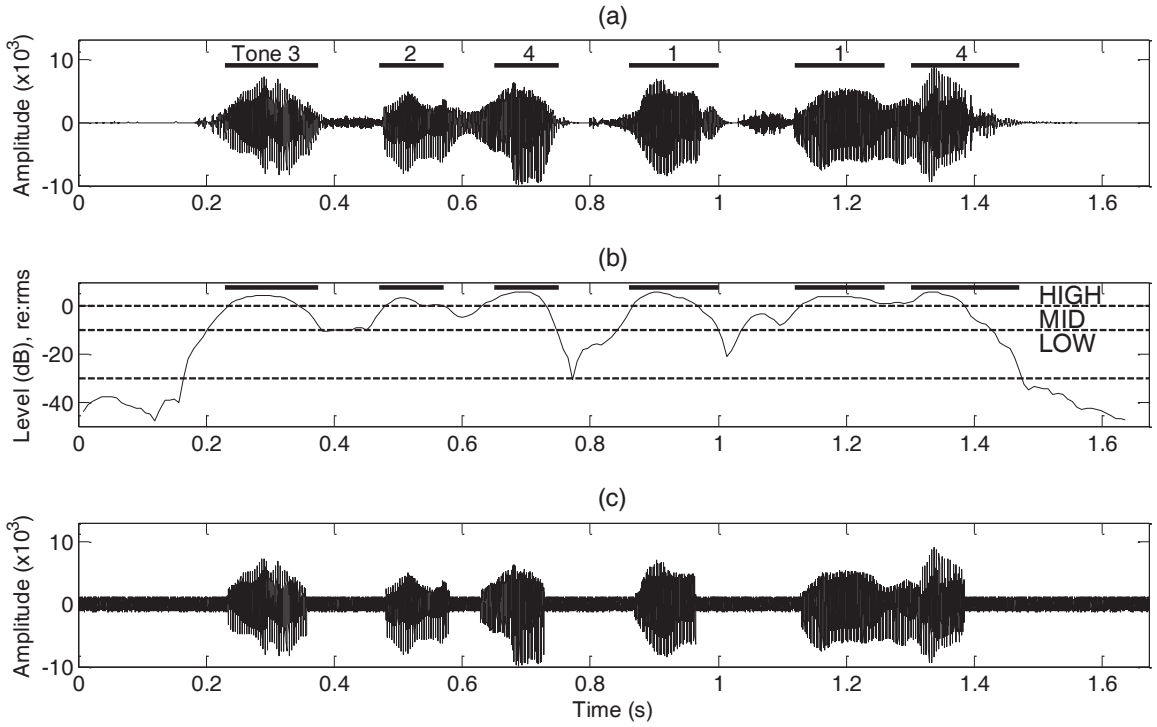

Fig. 1. Example waveforms of (a) a Mandarin sentence: "wo (in tone 3) xue (2) hui (4) kai (1) che (1) le (4)", (b) its relative RMS energy expressed in dB relative to the overall RMS level of the whole utterance, and (c) its high-RMS-level sentence. The solid lines in (a) and (b) delineate the voiced segments of the 6 Chinese words wherein the tones reside, while the numbers above the solid lines denote the tones of those words. Dashed lines in (b) show the boundaries of the high-, middle-, and low-RMS-level regions. The relative RMS threshold value is 0 dB for the high-level segmentation, and [0, –10] dB for the mid-level segmentation.

2. INTELLIGIBILITY OF THE HIGH-RMS-LEVEL MANDARIN SENTENCES

2.1. Speech intelligibility data

Eight (four male and four female) normal-hearing (NH) native-Mandarin-Chinese listeners participated in the experiment. The subjects' age ranged from 23 to 34 yrs, with the majority being graduate students at The University of Hong Kong.

The sentence material consisted of sentences taken from the Mandarin Hearing in Noise Test (MHINT) database [8]. There were totally 12 lists in the MHINT corpus. Each MHINT list had 20 sentences, and each sentence contained 10 keywords. All the sentences were produced by a male speaker, with fundamental frequency ranging from 75 to 180 Hz.

To create the level-dependent (e.g., H- and M-level) sentences, the relative-RMS-level-based sentence segmentation was first implemented to divided 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 [5-6]. The H-level segments were then kept intact, while the rest (i.e., M- and L-level) segments were replaced with white noise having the same RMS value with those replaced segments [1]. The relative high-level RMS threshold (i.e., $hRMS_{thr}$) was originally proposed as 0 dB relative to the overall RMS level of the whole utterance [6]. The plots of relative RMS energy of a Mandarin sentence (i.e., expressed in dB relative to the RMS level of the whole utterance) and its high-RMS-level sentence are exemplified in Figs. 1 (b) and (c). To examine the extent to what the relative high-level RMS threshold (i.e., $hRMS_{thr}$) value would influence the intelligibility of the high-RMS-level sentences, this study varied the $hRMS_{thr}$ value from 6 to –4 dB. In other words, the H-level region consists of segments at or above $hRMS_{thr}$ dB relative to the overall RMS level of the whole utterance.

The experiment was performed in a sound-proof room, and stimuli were played to listeners monaurally through a Sennheiser HD 250 Linear II circumaural head-phone at a comfortable listening level. The subjects were asked to write down the words they heard. Each subject participated in a total of 6 conditions (i.e., 6 relative high-level RMS threshold values of 6, 4, 2, 0, –2, and –4 dB). None of the sentences were repeated across the conditions, and the order of the test conditions was randomized across subjects. The intelligibility score for each condition was computed as the ratio between the number of the correctly recognized words and the total number of words contained in 20 sentences.

2.2. Results

Table 1 shows the average intelligibility scores obtained by NH listeners in the various conditions. As expected, higher

Table 1. Mean sentence recognition scores obtained by NH listeners in various conditions, and the average duration for the relative RMS-level based segmentation for Mandarin sentences. Asterisk denotes that the intelligibility score is significantly (p<0.05) higher than that of the M-level condition.

	High-RMS-level						Mid-RMS-level
	$hRMS_{thr}$=6 (dB)	4	2	0	–2	–4	
Intelligibility score (%)	34.3	61.1	79.6 *	**90.8 ***	95.1 *	98.6 *	**72.4**
Average duration (%)	9.1	15.8	22.2	**27.8**	33.3	38.7	**23.2**

Table 2. Correlation coefficients obtained with the relative RMS-level segmentation based NCM measures. Asterisk denotes that the correlation coefficient is significantly (p<0.05) higher than that computed with the M-level segments or all segments.

	NCM_{high}						NCM_{mid}	NCM_{all}
	$hRMS_{thr}$=6 (dB)	4	2	0	–2	–4		
Correlation coefficient (r)	0.72	0.94	0.98	**0.98 ***	0.98 *	0.98 *	**0.91**	**0.95**

intelligibility score is obtained when the relative H-level RMS threshold value decreases. The intelligibility score is 34% when the $hRMS_{thr}$ value is 6 dB (i.e., selecting 9% of the highest-RMS-level segments), and is 99% when the $hRMS_{thr}$ value is –4 dB (i.e., selecting 39% of the highest-RMS-level segments). When the $hRMS_{thr}$ value is set to 0 dB, as originally proposed in [6], the intelligibility score is about 91%. Table 1 also gives the intelligibility score of the mid-level sentences (i.e., energy level ranges from overall RMS level to 10 dB below). It is seen that, though the H- (i.e., above the overall RMS level of the whole utterance) and M-level (i.e., from RMS to RMS–10 dB) segments have almost the same averaged duration (i.e., 28% vs. 23%), the high-RMS-level sentences yield a significantly higher recognition rate than the mid-RMS-level sentences (i.e., 91% vs. 72%) [9].

3. INTELLIGIBILITY PREDITION USING THE HIGH-RMS-LEVEL SEGMENTS

3.1. Speech intelligibility data

The speech intelligibility data was taken from the intelligibility evaluation of noisy Mandarin sentences listened by a total of 9 new NH listeners [7]. Briefly, sentences taken from the Sound Express database [10] were corrupted by the SSN and 2-talker maskers at 8 SNR levels (i.e., –14, –12, –10, –8, –6, –4, –2, and 0 dB). All the sentences were produced by a female speaker. Two types of maskers were used to corrupt the sentences. The first masker was continuous steady-state noise, and the second was two equal-level interfering female talkers. The SNR levels were chosen to avoid ceiling/floor effects. The corrupted Mandarin sentences were first processed through a pre-emphasis filter (2000 Hz cutoff) with a 3 dB/octave roll-off, and then band-limited to the frequency range between 80 and 6000 Hz. The noisy speech sentence files were presented to the listeners in a double-walled sound-proof booth via Sennheiser's HD 250 Linear II circumaural headphones at comfortable listening levels. Twenty sentences were used for each condition, and none of the sentence lists were repeated. The intelligibility scores were obtained from NH listeners in a total of 16 conditions (=2 maskers × 8 SNR levels).

3.2. Speech intelligibility measures

In this study, we examined the intelligibility prediction performance of the normalized covariance measure (NCM), which is classified as a speech-transmission index (STI) based measure [11-12]. The NCM index computes a weighted sum of transmission index (TI) values determined from the envelopes of the probe (input) and response (output) signals in each frequency band [12]. Unlike the traditional STI measure, however, which quantifies the change in modulation depth between the probe and response envelopes using the modulation transfer function, the NCM index is based on the covariance between the probe and response envelope signals computed in each band. In its original implementation, the NCM index makes use of the envelopes extracted for the whole utterance to compute the TI value of each band (TI values are subsequently converted to an apparent SNR and mapped to the NCM index taking values between 0 and 1). In the present study, we modified the NCM index as follows to account for the relative RMS-level segmentation methods, i.e., to account for a selected set of (e.g., high-level) segments entering its computation. Using the probe signal, the time instances of the segments of interest (e.g., the H- or M-level segments) were first determined according to the relative-RMS-level segmentation method. The envelopes falling within each of the selected segments were then concatenated into one composite envelope of each frequency band. This was done for both the probe and response stimuli. Finally, the corresponding composite (concatenated) probe and response envelopes were used to compute the TI values for each band and subsequently the NCM index. More details regarding the definition and implementation of the NCM measure can be found in [12-13].

3.3. Results

The average intelligibility scores obtained by NH listeners were subjected to correlation analysis with the corresponding values obtained by the NCM measure implemented using the relative RMS-level sentence-segmentation method. More specifically, correlation analysis was performed between the mean (across all subjects) intelligibility scores obtained in each of the 16 testing conditions and the corresponding mean (computed

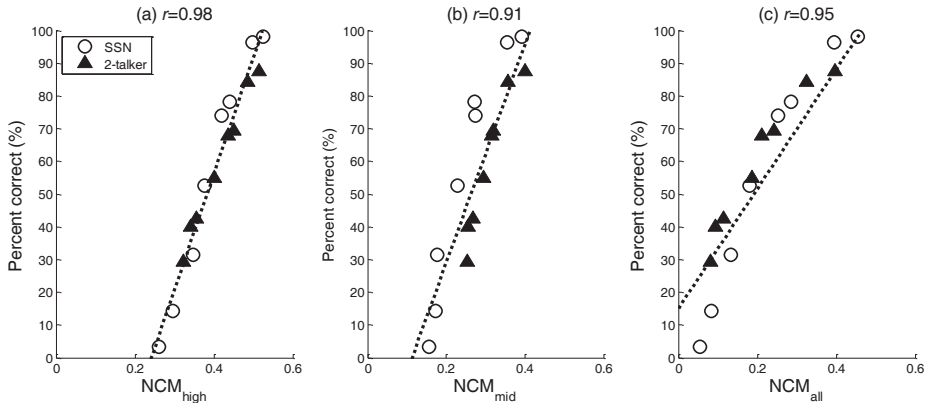

Fig. 2. Scatter plots of Mandarin sentence recognition scores against the (a) NCM_{high}, (b) NCM_{mid}, and (c) NCM values. The relative RMS threshold vale is 0 dB for the high-level segmentation, and [0, –10] dB for the mid-level segmentation.

across the 20 sentences used in each condition) intelligibility index values obtained in each condition. The Pearson's correlation coefficient (r) was used to assess the performance of the intelligibility measures to predict intelligibility scores.

Table 2 shows the correlation coefficients r between speech recognition scores and the NCM measures implemented with the relative-RMS-level segmentation method. Figure 2 shows the scatter plots of Mandarin sentence recognition scores against the predicted NCM_{high}, NCM_{mid} and NCM values. It is seen in Table 2 and Fig. 2 that the H-level (i.e., the $hRMS_{thr}$ value of 0 dB) based NCM measure (i.e., NCM_{high}) predicts the intelligibility statistically better ($p<0.05$) than that computed with the M-level segments or all segments (i.e., r=0.98 vs. 0.91 or 0.95) [9]. It is also noted that, though the correlation coefficient improves when varying the $hRMS_{thr}$ value from 6 dB to 0 dB (i.e., r=0.72 to 0.98), no further improvement in correlation is observed when including more H-level segments (or using smaller $hRMS_{thr}$ value, e.g., –2 dB) into computing the NCM_{high} measure.

4. DISCUSSION AND CONCLUSION

The human auditory system has a remarkable capacity for understanding speech in adverse conditions. The brain may retrieve the meaning of the distorted sentences by using *a priori* knowledge, language experience, expectations, contextual cues and linguistic rules involved in a top-down processing. The top-down processing of the central auditory system for high-level speech perception has been reported in many listening conditions, e.g., interrupted speech [14]. Hence, it is not surprising that, though the high-RMS-level (with $hRMS_{thr}$ value of 0 dB) segments occupy only about 28% of the total sentence duration, it is still quiet intelligible, i.e., with recognition score up to 91%.

Besides the contributions of contextual cues, linguistic cues may also account for the benefits of the H-level sentences. It has been reported that, for the most part, the H-level segments include vowels (and semivowels) while the M-level segments include vowel-consonant transitions [5-6]. In other words, acoustic cues on lexical tone identification mainly exist in the vowel-dominated H-level segments with strong energy, as shown in Fig. 1 (c), and listeners could get sufficient tonal information from the high-RMS-level Mandarin sentences. Studies have suggested that vowels contained more intelligibility information [1, 3] for English speech perception. Though it has not been verified with Mandarin speech, it is reasonable to expect that the H-level segments may carry more intelligibility information (e.g., vowels and lexical tones) to yield a high sentence recognition performance.

This study shows that the H-level segments are more beneficial for predicting the intelligibility of Mandarin sentences. The correlation coefficients of the NCM measure are 0.98 and 0.91 when including the H- and M-level segments, respectively, for intelligibility prediction. This advantage (i.e., the H- against M-level segments) differs with that observed with English sentences. Previous findings showed that the M-level segments of English sentences normally have a stronger intelligibility power than the high-level segments [5, 13, 15]. Therefore, together with results from previous studies, this work reveals a language effect on the contributions of the relative high-RMS-level segments in intelligibility prediction.

In conclusion, this paper assessed the relative contributions of the high-RMS-level segments to the intelligibility of Mandarin speech. The high-RMS-level (i.e., above the overall RMS level of the whole utterance) segments carry more intelligibility information (i.e., intelligibility score up to 91%) than those middle-RMS-level segments in Mandarin sentences. The relative importance of lexical tone for Mandarin perception might partially account for the intelligibility advantage of the H-level sentences. In addition, the high-RMS-level segments contribute more than the middle-RMS-level segments in predicting the intelligibility of Mandarin sentences in noise, which differs with results obtained from English speech.

5. ACKNOWLEDGEMENTS

This research was supported by Faculty Research Fund, Faculty of Education, The University of Hong Kong, by Seed Funding for Basic Research, The University of Hong Kong, and by General Research Fund (GRF), administered by the Hong Kong Research Grants council.

6. REFERENCES

[1] R. Cole, Y. Yan, B. Mak, M. Fanty, and T. Bailey, "The contribution of consonants versus vowels to word recognition in fluent speech," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 853–856, 1996.

[2] J. Garofolo, L. Lamel, W. Fisher, et al. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia: Linguistic Data Consortium, 1993.

[3] D. Fogerty and D. Kewley-Port, "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility," *J. Acoust. Soc. Am.*, vol. 126, no. 2, pp. 847–857, Aug. 2009.

[4] C.E. Stilp and K.R. Kluender, "Cochlear-scaled entropy, not consonants, vowels or time, best predicts speech intelligibility," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 107, pp. 12387–12392, 2010.

[5] F. Chen and P.C. Loizou, "Contributions of cochlea-scaled entropy and consonant-vowel boundaries to prediction of speech intelligibility in noise," *J. Acoust. Soc. Am.*, vol. 131, no. 5, pp. 4104–4113, May 2012.

[6] J. Kates and K. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, vol. 117, pp. 2224–2237, Apr. 2005.

[7] F. Chen and P.C. Loizou, "Predicting the intelligibility of vocoded and wideband Mandarin Chinese," *J. Acoust. Soc. Am.*, vol. 129, no. 5, pp. 3281–3290, May 2011.

[8] L.L. Wong, S.D. Soli, S. Liu, N. Han, and M.W. Huang, "Development of the Mandarin hearing in noise test (MHINT)," *Ear Hear.*, vol. 28, pp. 70S–74S, 2007.

[9] J.H. Steiger, "Tests for comparing elements of a correlation matrix," *Psychological Bulletin,* vol. 87, pp. 245–251, 1980.

[10] TigerSpeech Technology: http://www.tigerspeech.com/ (last visited: Nov. 29, 2012)

[11] H. J. M. Steeneken and T. Houtgast, "A physical method for measuring speech transmission quality," *J. Acoust. Soc. Amer.*, vol. 67, no. 1, pp. 318–326, Jan. 1980.

[12] R. Goldsworthy and J. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Amer.*, vol. 116, no. 6, pp. 3679–3689, Dec. 2004.

[13] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Amer.*, vol. 125, no. 5, pp. 3387–3405, May 2009.

[14] M. Chatterjee, F. Peredo, D. Nelson, and D. Başkent, "Recognition of interrupted sentences under conditions of spectral degradation," *J. Acoust. Soc. Am.*, vol. 127, no. 2, pp. EL37–41, Feb. 2010.

[15] F. Chen and P.C. Loizou, "Predicting the intelligibility of vocoded speech," *Ear Hear.*, vol. 32, pp. 331–338, 2011.