INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Assessing Level-dependent Segmental Contribution to the Intelligibility of Speech Processed by Single-channel Noise-suppression Algorithms

Tian Guan [1], *Guangxing Chu* [1], *Fei Chen* [2], *Feng Yang* [3]

[1] Research Centre of Biomedical Engineering, Graduate School at Shenzhen, Tsinghua University, Shenzhen, China

[2] Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China

[3] Shenzhen Children's Hospital, Shenzhen, China

guantian@sz.tsinghua.edu.cn,fchen@hku.hk,hkufrank@163.com

Abstract

Most existing single-channel noise-suppression algorithms cannot improve speech intelligibility for normal-hearing listeners; however, the underlying reason for this performance deficit is still unclear. Given that various speech segments contain different perceptual contributions, the present work assesses whether the intelligibility of noisy speech can be improved when selectively suppressing its noise at high-level (vowel-dominated) or middle-level (containing vowel-consonant transitions) segments by existing single-channel noise-suppression algorithms. The speech signal was corrupted by speech-spectrum shaped noise and two-talker babble masker, and its noisy high- or middle-level segments were replaced by their noise-suppressed versions processed by four types of existing single-channel noise-suppression algorithms. Experimental results showed that performing segmental noise-suppression at high- or middle-level led to decreased intelligibility relative to noisy speech. This suggests that the lack of intelligibility improvement by existing noise-suppression algorithms is also present at segmental level, which may account for the deficit traditionally observed at full-sentence level.

Index Terms: Noise suppression, speech intelligibility, segmental contribution.

1. Introduction

The objective of speech enhancement (or noise suppression) is to improve one or more perceptual aspects of noisy speech, most notably, quality and intelligibility [1]. However, improving speech quality might not necessarily lead to improvement in speech intelligibility. In fact, in many cases improvement in quality might be accompanied by a decrease in intelligibility [e.g., 2-3]; however, the underlying reason for being unable to improve intelligibility is still unclear.

When performing noise suppression, the present speech enhancement algorithms commonly lead to non-linear distortion contained in noise-suppressed speech (e.g., musical noise) [4-6]. Measuring the effect of non-linear distortion is a challenging task considering that many types of distortions exist due to the varieties of noise-suppression algorithms. An earlier study by Kim and Loizou simplified the classification of non-linear distortions and classified the distortions into either amplification or attenuation distortion [4], where the amplification distortion and attenuation distortion referred to the scenarios that the envelope (or magnitude) spectrum of the noise-suppressed speech was larger and smaller, respectively, than that of clean speech. They found that amplification distortion was practically harmful to speech intelligibility, and in contrast, attenuation distortion did not impair speech intelligibility [4]. This implies that, while amplification and attenuation distortions co-exist in noise-suppressed speech, they need to be treated differently for speech intelligibility prediction.

The above-mentioned lack of intelligibility improvement by noise-suppression processing was demonstrated at full-sentence level. Many studies have found that various speech segments carry different amount of intelligibility information [e.g., 7-10]. For instance, vowel-only speech (with consonants replaced by noise) is more intelligible than consonant-only speech (with vowels replaced by noise) [7-8]. When speech signal is segmented by the relative root-mean-square (RMS)-level based segmentation, the high-level (H-level) and middle-level (M-level) regions consists of segments at or above the overall RMS level of the whole utterance and segments ranging from the overall RMS level to 10 dB below (i.e., RMS–10 dB), respectively (see example in Fig. 1). For the most part, H-level segments include vowels and semivowels, while M-level segments include consonants and vowel-consonant transitions [9]. Earlier studies have found that H- and M-level segments carry different perceptual contributions for speech intelligibility [10].

Given the difference of perceptual contributions at segmental level, a question is raised: whether the present speech enhancement algorithms can improve the intelligibility of segmentally noise-corrupted speech? In other words, if only the selected (H- or M-level) segments of noisy speech are processed by noise suppression processing, would this lead to improved intelligibility relative to noisy speech (containing all noisy segments)? The answer to this question can improve our insights on the lack of intelligibility improvement at both full-sentence and segmental levels. For instance, if there is intelligibility improvement when noisy speech is noise-suppressed at segmental level, the intelligibility improvement deficit at full-sentence level may be attributed to the integration of distortion contained in different speech segments. The purpose of this study is to assess the segmental contribution to the intelligibility of noise-suppressed speech. More specifically, we will examine the perceptual

http://dx.doi.org/10.21437/Interspeech.2016-43

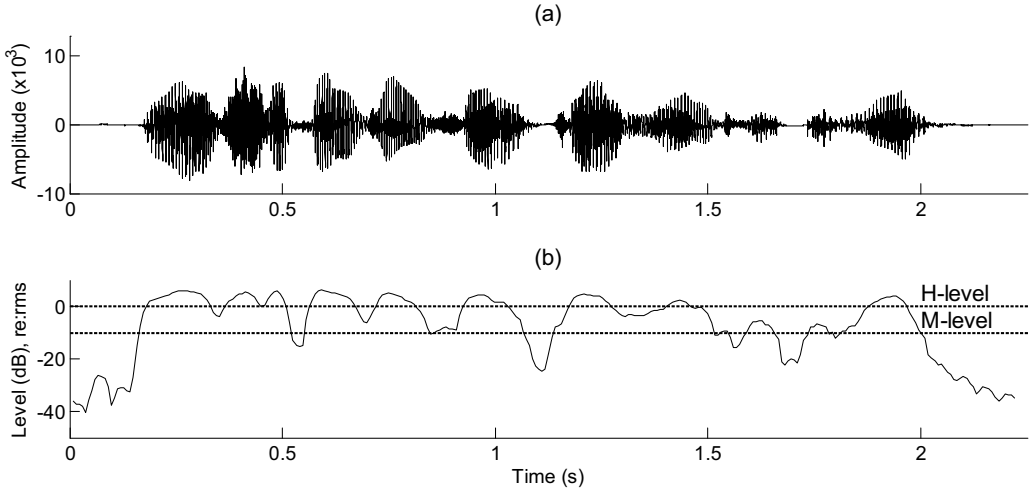

Figure 1. *Example waveforms of (a) a sentence and (b) its relative RMS energy expressed in dB relative to the overall RMS level of the whole utterance. Dashed lines in (b) show the boundaries of the high-, and middle-RMS-level regions. The relative RMS threshold level(s) is 0 dB for the H-level segmentation, and [0, –10] dB for the M-level segmentation.*

contributions of H-level (vowel-dominated) and M-level (containing vowel-consonant transitions) segments to the intelligibility of noise-suppressed speech.

2. Experiment

2.1. Subjects and materials

Ten (5 male and 5 female, aged 18 to 24 yrs) normal-hearing (NH) native Mandarin listeners participated in the experiment. The sentence material consisted of sentences taken from the Mandarin Hearing in Noise Test (MHINT) database [11]. There were totally 24 lists in the MHINT corpus. Each MHINT list had 10 sentences, and each sentence contained 10 keywords. All the sentences were produced by a male speaker, and their duration was around 2.8±0.3 second. A steady-state speech-spectrum shaped noise (SSN) and a two male-voice (2-talker) babble were used to corrupt the MHINT sentences at –10 and –5 dB signal-to-noise ratio (SNR) levels, respectively. The SNR levels were chosen to avoid the ceiling/floor effects. Note that the SSN and 2-talker maskers were used in this experiment because they represented two different types of masking, i.e., steady-state and competing.

2.2. Signal processing

The noise-corrupted sentences were processed by four different speech enhancement algorithms, which included the generalized KLT approach [12], the Log Minimum Mean Square Error (logMMSE) algorithm[13], the multiband spectral-subtractive algorithm [14], and the Wiener algorithm based on *a priori* SNR estimation (Wiener) [15]. These four algorithms were selected because they covered the four most-used types of single-channel speech enhancement methods (i.e., subspace approach, statistical-modeling approach, spectral-subtractive approach, and Wiener-filtering approach), and they represented the state-of-the-art noise-suppression techniques. The parameters used in the implementation of these algorithms were the same as those published, and the Matlab code for the above speech enhancement algorithms was taken from [1].

Following the speech enhancement processing of the above four stimuli (i.e., KLT, logMMSE, MB and Wiener) and the reference noisy stimuli, we synthesized two types of stimuli containing 1) enhanced speech at H-level segments while the rest segments were noisy, and 2) enhanced speech at M-level segments while the rest segments were noisy. This study assessed six segmental processing conditions to the intelligibility of noise-corrupted speech. The condition of (full-sentence) noisy speech (noted as Noisy) gives the reference intelligibility score, and the condition of segmentally clean speech (i.e., replacing noisy M- or H-level segments with their clean versions, noted as CLN) shows the best intelligibility score for noise-suppressed conditions. Four additional conditions use noise-suppressed segments (i.e., by four noise-suppressed algorithms) to replace their corresponding noisy segments.

The relative RMS-level based segmentation is implemented by dividing speech into short-term (16-ms in this study) segments and classifying each segment into H- or M-level regions according to its relative RMS intensity [4]. The threshold levels of 0 and –10 dB split speech into H-level and M-level, and these two threshold levels were selected as they were originally proposed in [9]. Note that the RMS-level segmentation is implemented on the clean speech signal in this study. Figure 1 shows an example sentence segmented into H-level and M-level based on the above RMS threshold levels.

2.3. Procedure

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. Each subject participated in a total of 24 conditions [=2 maskers (i.e., SSN at –10 dB SNR and 2-talker masker at –5 dB SNR) × 2 segmental noise-suppression conditions (i.e., H-level and M-level) × 6 signal processing conditions (i.e., Noisy, KLT, logMMSE, MB, Wiener, and CLN)]. Different sentence lists were presented to each listener for different test conditions. The order of the test conditions was randomized across subjects. Subjects were allowed to listen to the sentences 3 times at most, and were instructed to orally repeat all the words that they could recognize. The

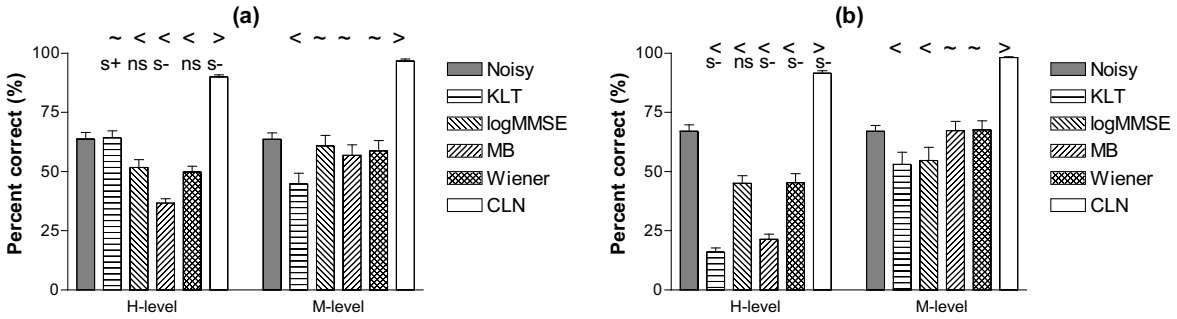

Figure 2. *Mean sentence intelligibility scores as a function of segmental condition at (a) SSN masker and –10 dB SNR, and (b) 2-talker masker and –5 dB SNR. The error bars denote ±1 standard error of the mean. '~', '<' and '>' denote that the intelligibility score of segmentally enhanced speech (by KLT, logMMSE, MB, Wiener or CLN) is non-significantly ($p>0.05$) smaller/larger, significantly ($p<0.05$) smaller, and significantly ($p<0.05$) larger, respectively, than that of noisy speech at the same group of tested condition. 'ns', 's–', and 's+' mean that the intelligibility score of segmentally enhanced speech at H-level is non-significantly ($p>0.05$) smaller, significantly ($p<0.05$) smaller, and significantly ($p<0.05$) larger, respectively, than its paired test condition at M-level.*

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 each list of 10 MHINT sentences.

3. Results

Figure 2 (a) shows the mean sentence recognition scores of all tested conditions when the interfering masker is SSN at –10 dB SNR. Statistical significance was determined by using the percent recognition score as the dependent variable, and segmental condition (H-level and M-level) and noise-suppression processing (Noisy, KLT, logMMSE, MB, Wiener and CLN) as the two within-subject factors. Two-way analysis of variance (ANOVA) with repeated measures indicated a non-significant effect (F[1, 9]=3.496, p=0.094) of segmental condition, significant effect (F[5, 45]=62.282, p<0.001) of noise-suppression processing, and a significant interaction (F[5, 45]=12.483, p<0.001) between segmental condition and noise-suppression processing.

Post hoc pairwise comparisons at H-level condition showed that the recognition score of CLN sentences was significantly larger than that of Noisy sentences; however, the scores of noise-suppressed sentences (by noise-suppression algorithms) were either non-significant (p>0.05) different from or significantly (p<0.05) smaller than that of Noisy sentences. This finding was also seen from the six conditions at M-level in Fig. 2 (a).

Post hoc pairwise comparisons at the same noise-suppression processing showed that the paired scores at Noisy, logMMSE and Wiener conditions were non-significantly (p>0.05) different between H- and M-level conditions, the scores at MB or CLN and H-level condition was significantly (p<0.05) smaller its paired score at M-level condition, and the score at KLT and H-level condition was significantly (p<0.05) larger that its paired score at M-level condition.

Figure 2 (b) shows the mean sentence recognition scores of all conditions when the interfering masker is 2-talker babble at –5 dB SNR. Statistical significance was determined by using the percent recognition score as the dependent variable, and segmental condition and noise-suppression processing as the two within-subject factors. Two-way ANOVA with repeated measures indicated a significant effect (F[1, 9]=66.233, p<0.001) of segmental condition, significant effect (F[5, 45]=110.429, p<0.001) of noise-suppression processing, and a significant interaction (F[5, 45]=15.143, p<0.001) between segmental condition and noise-suppression processing.

Post hoc pairwise comparisons at H-level condition showed that the recognition score of CLN sentences was significantly larger than that of Noisy sentences; however, the scores of noise-suppressed sentences (by noise-suppression algorithms) were all significantly (p<0.05) smaller than that of Noisy sentences. This finding was also seen from the six conditions at M-level in Fig. 2 (b). That is, the recognition score of CLN sentences was significantly larger than that of Noisy sentences; however, the scores of noise-suppressed sentences (by noise-suppression algorithms) were all significantly (p<0.05) smaller than that of Noisy sentences.

Post hoc pairwise comparisons at the same noise-suppression processing showed that the paired score at logMMSE condition was non-significantly (p>0.05) different between H- and M-level conditions, the scores at all other five conditions were significantly (p<0.05) smaller their paired scores at M-level condition.

4. Discussion and conclusions

While many earlier studies found that existing single-channel speech enhancement algorithms could not improve or even deteriorate the intelligibility of noise-corrupted speech at full-sentence level by NH listeners, the present work further found that they could not lead to intelligibility improvement at segmental level. When only H- or M-level segments were processed by noise suppression algorithms, the intelligibility score of stimuli containing noise-suppressed H- or M-level segments (while the other segments remained noisy) was not larger than that of noisy speech (i.e., with all noisy segments). However, when the H- or M-level segments were replaced by their clean versions, Fig. 2 shows that the intelligibility score could be significantly improved relative to that of noisy speech. This indicates that at the presence of noise corruption,

speech intelligibility could be improved even when only selectively performing noise-suppression at selected H- or M-level segments. However, as mentioned earlier, noise-suppression processing is marked with extra distortion which is detrimental for speech intelligibility [e.g., 4]. The present work showed that the distortion was equally distributed at H-level and M-level. That is, the distortion at either H- or M-level is harmful to the intelligibility of segmentally noise-suppressed speech, yielding a decreased intelligibility score relative to noisy speech.

Most (if not all) noise-suppression algorithms involve a gain reduction stage, in which the mixture envelope or spectrum is multiplied by a non-linear gain function (i.e., taking values from 0 to 1) with the intent of suppressing background noise [1]. The shape and choice of the gain function varies across algorithms, but independent of its shape, when the gain function is applied to the mixture envelopes (or spectra), it introduces distortion to the envelopes (or spectra). The gain function is normally influenced by the SNR level, and a low SNR level may severely affect the accuracy of gain function estimation. This study assessed the effect of two level-dependent speech segments, i.e., H-level and M-level, for improving the intelligibility of noise-suppressed speech. The H-level region consists of segments at or above the overall RMS level of the whole utterance, while the M-level region consists of segments ranging from the overall RMS level to 10 dB below (i.e., RMS–10 dB). Hence, the intensity of speech segment at H-level is larger than that at M-level, or the local SNR level at H-level may be larger than that at M-level. This would lead to the assumption that H-level segments may contain less distortion (due to its relatively accurate SNR estimation) than M-level segments. However, as seen in Fig. 2, the present work showed that H-level noise-suppressed speech did not always lead to better intelligibility performance than its M-level counterpart. This suggests that the effect of distortion (i.e., caused by noise-suppression) to speech intelligibility is not affected the segmental level. Though H-level and M-level contain speech segments with different levels, there is no H-level advantage to reduce the detrimental effect of distortion to intelligibility. Instead, it is interesting to see that, at 2-talker babble condition in Fig. 2 (b), the distortion from noise-suppression at M-level yielded a much higher intelligibility score than that at H-level.

Earlier work showed that M-level segments, which contained more vowel-consonant transitions, carried important perceptual information for speech intelligibility in noise [e.g., 10]. The present study also demonstrated this M-level advantage for speech recognition. Figure 2 (b) shows that the intelligibility score of M-level enhanced speech is significantly (p<0.05) larger than its paired score of H-level enhanced speech, except at logMMSE condition where the difference of two scores at M-level and H-level is non-significant [noted at 'ns' in Fig. 2 (b)]. The same finding is seen in Fig. 2 (a) where speech signal was corrupted by SSN masker at –10 dB SNR. The mean scores of most M-level enhanced conditions are significantly (p<0.05) or non-significant larger than their paired scores of H-level enhanced conditions, with exception of KLT condition. Hence, these results suggest that M-level might carry more perceptional importance for speech enhancement processing.

In conclusion, the present work assessed the segmental contribution of noise suppression for improving speech intelligibility. When only H- or M-level segments of noise-corrupted speech were processed by existing single-channel speech enhancement algorithms and the other segments were corrupted by noise, the intelligibility of segmentally noise-suppressed speech was not improved relative to the noisy speech containing all noisy segments. Taking findings from earlier work, this study showed that single-channel speech enhancement algorithms may not improve the intelligibility of noise-corrupted sentences at both full-sentence and segmental levels. Future work can design algorithms aiming to diminish the distortion at segmental level according to the acoustic difference and perceptual importance of various speech segments.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61571213 and 31271056). This work was also supported by Shenzhen Medical Engineering Laboratory for Human Auditory-Equilibrium Function, and Shenzhen Municipal Science and Technology Innovation Committee (Grant No. CYJ20140416141331555).

7. References

[1] P.C. Loizou, Speech *Enhancement: Theory and Practice,* Taylor & Francis Group, Boca Raton, 2007.

[2] Y. Hu and P.C. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Am.*, 122, 1777–1786, 2007.

[3] J.F. Li, L. Yang, J. Zhang, Y.H. Yan, Y. Hu, M. Akagi, and P.C. Loizou, "Comparative intelligibility investigation of single-channel noise-reduction algorithms for Chinese, Japanese, and English," *J. Acoust. Soc. Am.*, 129, 3291–3301, 2011.

[4] P.C. Loizou and G.B. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. Audio Speech Lang. Process.*, 19, 47–56, 2011.

[5] F. Chen and P.C. Loizou, "Impact of SNR and gain-function over- and under-estimation on speech intelligibility," *Speech Commun.*, 54, 272–281, 2012.

[6] G.B. Kim and P.C. Loizou, "Gain-induced speech distortions and the absence of intelligibility benefit with existing noise-reduction algorithms," *J. Acoust. Soc. Am.*, 130, 1581–1596, 2011.

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

[8] F. Chen, L. L. N. Wong, and Y. W. Wong, "Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility," *J. Acoust. Soc. Am.*, 134, EL178–EL184, 2013.

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

[10] F. Chen and Loizou, "Contribution of cochlea-scaled entropy versus consonant-vowel boundaries to prediction of speech intelligibility in noise," *J. Acoust. Soc. Am.*, 131, 4104–4113, 2012.

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

[12] Y. Hu and P.C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process.*, 11, 334–341, 2003.

[13] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, 33, 443–445, 1985.

[14] S. Kamath and P.C. Loizou, "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*, 2002.

[15] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Processing*, pp. 629–632, 1996.