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

Modeling Noise Influence to Speech Intelligibility Non-intrusively by Reduced Speech Dynamic Range

Fei Chen

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

fchen@sustc.edu.cn

Abstract

The noise influence to speech signal waveform can be characterized by reduced speech dynamic range (rDR). This motivated the present work to propose an rDR-based intelligibility measure (denoted as rDRm) that could be used to non-intrusively (i.e., do not require clean reference speech signal) predict speech intelligibility in noise and is computed only using the dynamic range extracted from the noise-corrupted speech. The rDRm indices were evaluated with intelligibility scores obtained from normal-hearing listeners presented with sentences corrupted by four types of maskers in a total of 22 conditions. High correlation (*r*=0.93) was obtained between rDRm values and listeners' sentence recognition scores, and this correlation was comparable to those computed with existing intrusive and non-intrusive intelligibility measures. This suggests that the dynamic range of speech signal may work as a simple but efficient predictor of speech intelligibility in noise, whose computation does not need access to the clean reference speech signal.

Index Terms: Speech intelligibility, intelligibility prediction, speech dynamic range.

1. Introduction

Human listening experiment plays an important role in speech perception studies, e.g., designing novel noise-suppression algorithms, developing intelligibility prediction model, etc. However, listening experiments are a time-consuming process requiring a large amount of manpower and experimental cost. To address these limitations, objectively predicting speech intelligibility has long attracted attention in this field. The intelligibility model may improve our understanding of the interaction between objective acoustic cues or environmental influences (e.g., noise and reverberation) and subjective speech recognition performance, and could potentially guide our design of novel speech processing strategies, for example, see [1-6]. Hence, a number of research efforts have been directed to the design of reliable intelligibility indices [2-5]. Unfortunately, due to the complexity of listening conditions, we still have many challenges, and one challenge is the development of non-intrusive intelligibility index.

While a large number of speech intelligibility indices have been developed, most of them are primarily intrusive in their computations [2-5]. In other words, a clean probe or reference signal is required when computing the intrusive intelligibility index, e.g., the well-known speech-transition index (STI). The fundamental principle of these intrusive indices is to compute an apparent signal-to-noise ratio (SNR). This apparent SNR may take various definitions in different intelligibility indices. For instance, the normalized covariance was used in the normalized covariance measure (NCM), which is a speech- and STI-based intelligibility index [3], and the apparent SNR was defined with the normalized spectrum coherence between clean reference signal and noisy speech signal [2]. Though these intrusive intelligibility indices successfully model subjective speech recognition performance in many listening environments (e.g., in noise and in reverberation), their computations require clean reference signal which may not exist in many application scenarios. Hence, lacking clean reference signal may limit the applicability of these intrusive intelligibility indices.

There are a few known measures that can predict non-intrusively or blindly (i.e., with no access to the reference signal) speech intelligibility, e.g., SRMR (speech-to-reverberation modulation energy ratio), ModA (modulation-spectrum area), and ABECm (across-band envelope correlation metric) [7-10]. Falk et al. proposed the non-intrusive measure SRMR for predicting the subjective quality of reverberant and dereverberated speech [7-8]. The SRMR measure was implemented in the modulation spectrum domain and evaluated on tasks of predicting the effects of coloration, reverberation tail and overall quality. In addition to quality assessment, the SRMR measure was also assessed indirectly for intelligibility prediction by correlating the output of other STI-based measures with the output of SRMR measures. Chen et al. developed a non-intrusive intelligibility index (i.e., ModA) for predicting the intelligibility of reverberant speech [9]. The ModA index was rooted in basic principles of STI theory. In general, STI predicts that intelligibility drops with reduction in speech envelope modulations, irrespective of the nature of that reduction, i.e., whether it is caused by additive (steady) noise or reverberation. The non-intrusive ModA measure was found to successfully predict the intelligibility of reverberant speech. Recent psychoacoustic studies have found that across-band envelope correlation (ABEC) carries important information for speech intelligibility. This motivated the development of an ABEC-based intelligibility measure that could be used to non-intrusively predict speech intelligibility in noise using only temporal envelope waveforms extracted from the noise-corrupted speech [10]. The ABECm value was computed by averaging the correlation coefficients of mean-removed envelope waveforms from adjacent frequency bands of the noise-corrupted speech, and was found to well predict the intelligibility of speech in noise.

http://dx.doi.org/10.21437/Interspeech.2016-9

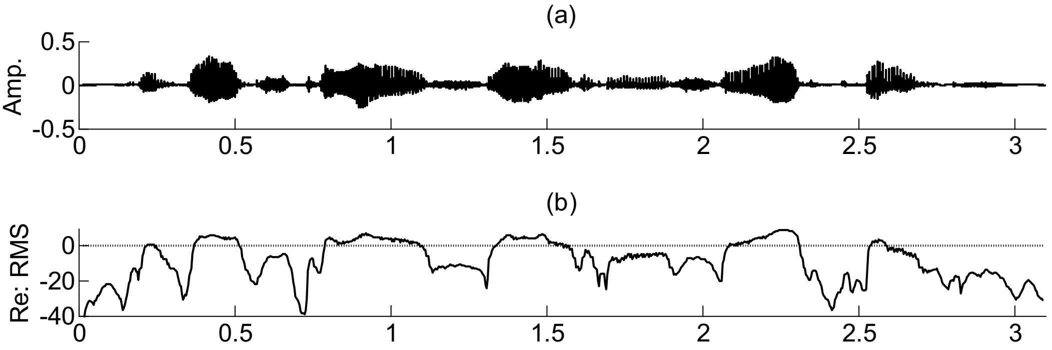

Figure 1. *(a) Example waveform of a clean sentence, and (b) its relative RMS energy expressed in dB relative to the overall RMS level of the whole utterance.*

The purpose of this study is to design a new non-intrusive intelligibility index for predicting the intelligibility of speech in noise. The proposed measure was motivated by the effect of noise interference in reducing speech dynamic range [11]. Acoustic analysis has showed that adding noise to speech signal may reduce its dynamic range (DR, see DR definition and example in Section 2), and this reduced dynamic range of speech in noise occurs in accordance with reduced speech intelligibility. Therefore, the present work developed a reduced dynamic range based intelligibility index for noisy speech, and assessed its performance for non-intrusively modelling the intelligibility of noise-corrupted speech.

2. Reduced dynamic range based intelligibility index

The proposed reduced speech dynamic range based intelligibility metric is computed as below. The noise-corrupted speech signal is divided into short (e.g., 16-ms) non-overlapping segments, and the root-mean-square (RMS) level of each segment is computed and further divided by the overall RMS level of the whole sentence. This will generate the relative-RMS-level waveform [2], as demonstrated in Fig. 1 (b). The speech dynamic range is defined as the difference between the maximum value and minimum value in the relative-RMS-level waveform in Fig. 1 (b). Figure 2 shows an example of noise influence to speech dynamic range. It is observed that the dynamic range of noisy speech signal varies in accordance with the SNR level. Note that Fig. 2 shows the dynamic range in response to a speech-spectrum shaped noise (SSN). Similar trends can also be observed for other types of masker signals.

3. Methods

3.1. Subjects and materials

A total of 8 normal-hearing (NH) subjects (age ranged from 18 to 27 yrs) participated in the listening tests to collect subjective intelligibility scores. All subjects were native speakers of American English, and were paid for their participation.

IEEE sentences were used as test material [12]. All sentences were produced by a male talker, and were taken from the CD in [1]. The masker signals included SSN, and three real-world recordings from different places: babble, car, and street. The sentences and four maskers were sampled at 25 kHz. Segments randomly selected from the maskers were added to the IEEE sentences at 10, –5, –3, 0 and 5 dB for babble masker, –15, –10 and –5 dB for car masker, –15, –10, –7, –5, –3, 0 and 5 dB for SSN masker, and - 15, –10, –7, –5, –3, 0 and 5 for street masker. The SNR levels were selected based on pilot data collected from one subject to avoid ceiling and floor effects.

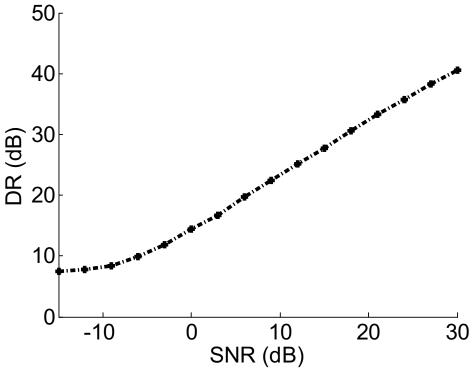

Figure 2. *Dynamic range in dB as a function of SNR level. DRs are computed from 10 sentences, and the masker signal is a speech-spectrum shaped noise.*

3.2. Procedure

The noise-corrupted sentences were presented binaurally to the listeners in a double-walled sound-proof booth via a circumaural headphone at a comfortable listening level (selected by the listeners). Each testing cost less than one hour and there was a 5-minute practice session performed before the actual test to familiarize the participants with the testing procedure. Twenty IEEE sentences were used for each condition, and none of the sentences were repeated. NH listeners participated in a total of 22 test conditions (=5, 3, 7 and 7 SNR levels for babble, car, SSN and street maskers, respectively). The test condition order was randomized across subjects, and subjects were given a 5-min break every 30 minutes of testing to avoid fatigue. 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 each condition.

Table 1. Correlation coefficients (r) between sentence recognition scores and the intelligibility measures examined in this study.

	Non-intrusive				Intrusive	
	rDRm	ABECm	ModA	SRMR	NCM	CSII$_{mid}$
r	0.93	0.96	0.93	0.91	0.91	0.95
N	–	16	4	23	16	16
f_{cut} (Hz)	–	20	10	2–20	20	–
weight across bands	–	1/(N–1)	1/N	–	1/N	ANSI [13]
b_1	0.8	20.7	8.7	15.4	19.9	3.7
b_2	7.7	9.3	3.9	3.2	4.0	4.1

4. Results

For the purpose of comparison, this study also predicted the intelligibility of the noise-corrupted speech by using two commonly-used intrusive measures, i.e., the NCM measure [3] and the middle-level coherence-based speech intelligibility index (CSII$_{mid}$) [2]. The underlying hypothesis of the intrusive intelligibility measures was that measures assessing temporal-envelope (i.e., NCM) or spectral (i.e., CSII$_{mid}$) distortion should correlate highly with the intelligibility of noise-corrupted speech. Earlier studies found that the NCM and CSII$_{mid}$ measures well predicted the intelligibility of noise-corrupted and/or noise-suppressed sentences [4]. The subscript 'mid' in CSII$_{mid}$ means that the CSII measure is computed by using only mid-level region consisting of segments ranging from the overall RMS level to 10 dB below (i.e., RMS–10 dB). More details regarding the definition and implementation of the NCM and CSII$_{mid}$ measures can be found in [4]. Note that these two measures need access to the clean reference signal to predict speech intelligibility.

The average intelligibility scores obtained by NH listeners were subjected to correlation analysis with the corresponding values computed by the previously mentioned intelligibility measures. More specifically, correlation analysis was performed between the mean (across all subjects) intelligibility scores obtained in each of the 22 test conditions and the corresponding mean (computed across the 20 sentences 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 in noise. Table 1 shows the correlation coefficients between subjective intelligibility scores and the rDRm, ABECm, ModA, SRMR, NCM and CSIImid measures. More details on computing the ABECm, ModA, SRMR, NCM and CSII$_{mid}$ measures can be found in [2, 3, 7, 9, 10]. It is seen that the rDRm indices well predict intelligibility scores, i.e., r=0.93. Note that N, f_{cut} and weight across bands are parameters used in computing intelligibility indices. For instance, N and f_{cut} denote the number of spectral bands and low-pass cut-off frequencies to extract multi-band temporal envelope waveforms, and each channel is weighted by a band-importance function.

Figure 3 shows the scatter plot of listeners' sentence recognition scores against the DR values. A logistic function was used to map the DR (and other index) values to sentence intelligibility scores, as:

$$y = \frac{1}{1+e^{-(b_1 \times index - b_2)}} \times 100, \quad (1)$$

where y is the mapped intelligibility score in percentage, and (b_1, b_2) are the fitting parameters. Table 1 also provides the values of the resulting fitting parameters [b_1 and b_2 in Eq. (1)] of the logistic transfer functions used to correlate intelligibility measures to sentence recognition scores.

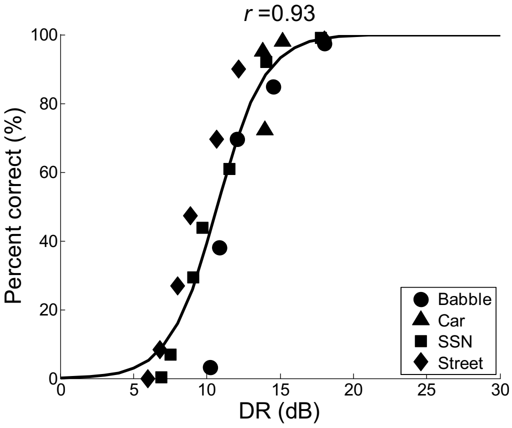

Figure 3. *Scatter plot of listeners' sentence recognition scores against the dynamic range values.*

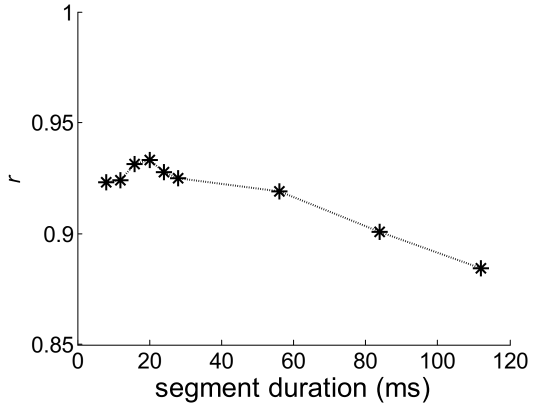

Figure 4. *The correlation coefficient of rDRm-based intelligibility prediction as a function of segment duration.*

It is observed in Table 1 that the prediction correlation of the rDRm index is comparable to those computed with the two intrusive intelligibility indices (i.e., 0.91 and 0.95 for the NCM and CSII$_{mid}$ measures, respectively); however, the computation of the rDRm index does not need access to clean reference signal. The prediction correlation of the rDRm index is also comparable to those computed with the three existing non-intrusive intelligibility measures, i.e., ABECm, ModA and SRMR.

Further analysis was done to assess the influence of segment duration [i.e., in computing the relative-RMS-level waveform, see Fig. 1 (b)] on the prediction power of the rDRm index. The rDRm-based intelligibility prediction correlation coefficients were computed with segment duration from 8 to 110 ms. It is seen in Fig. 4 that the largest correlation coefficient is obtained with segment duration 16 ms (i.e., the value used in Table 1 and Fig. 3). The correlation coefficients decrease when using long segment duration to compute dynamic range.

5. Discussion and conclusions

Speech perception studies have shown that many acoustic cues of speech signal are important predictor of speech intelligibility in noise. Hence, many studies designed non-intrusive intelligibility indices by using those important acoustic cues. For instance, the ModA index was developed based on the fact that adding noise to speech signal causes a reduction in modulations across all modulation frequencies [9]. Healy et al. found that when the correlations between the temporal envelopes of the speech signal were calculated, the rate at which the correlation fell closely matched the rate at which intelligibility fell for NH listeners [14]. This motivated the development of the ABECm index. The present work used the findings from speech perception studies that adding noise to speech signal leads to reduced speech dynamic rage and intelligibility score. Therefore, the development of most non-intrusive intelligibility indices were supported by findings from subjective speech perception studies, which not only improved our knowledge on the importance of various acoustic cues for speech intelligibility, but also guided our design of non-intrusive intelligibility index.

Note that many studies (including the present work) assessed the performance of non-intrusive intelligibility prediction for speech in noise, and obtained good prediction performance (see Table 1). Generally speaking, when noise signal is used to additively corrupt clean speech signal, this process does not contain much nonlinear distortion. Compared with many other speech corruption processes containing nonlinear distortion (e.g., by noise-suppression [1, 4]), this noise influence process is relatively simple. Hence, many non-intrusive intelligibility indices (e.g., ModA, ABECm and rDRm) well predict the intelligibility of speech in noise. When speech is subjected to nonlinear processing, most present intelligibility indices fail to successfully predict speech intelligibility, for example, see [4]. The present study also evaluated the rDRm index to predict the intelligibility scores of noise-corrupted speech processed through noise-suppression algorithms. The noise-suppressed sentences and their intelligibility scores were extracted from [4]. Briefly, masker signals including the real-world recordings from four different places: speech babble, car noise, street noise, and train noise were artificially added to the IEEE sentences at 0 and 5 dB SNR levels. The intelligibility scores obtained from the NH listeners in a total of 72 conditions (including 8 noise-suppression algorithms) were used to evaluate the predictive power of the rDRm index. It was found that the rDRm-based correlation coefficient was r=0.15, suggesting that the rDRm index was unable to well account for the intelligibility variance of speech containing nonlinear distortion.

In conclusion, an intelligibility index (i.e., rDRm) requiring no access to the clean reference signal was developed for predicting the intelligibility of speech in noise in this study. Analysis of the data indicated that a high correlation (r=0.93) was obtained between rDRm values and listeners' intelligibility scores in a total of 22 conditions, and the intelligibility prediction performance of the rDRm index was comparable to those of existing intrusive and non-intrusive intelligibility indices examined in this study. The findings in this study suggest that dynamic range, reflecting the influence of additive noise to clean speech signal, can be used as a simple but efficient non-intrusive predictor of speech intelligibility in noise.

6. Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61571213). This study was also supported by a grant from Neural and Cognitive Sciences Research Center, Southern University of Science and Technology, Shenzhen, China.

7. References

[1] P.C. Loizou, Speech Enhancement: Theory and Practice, 2ed, Taylor and Francis, Boca Raton, FL, 2013.

[2] J. Kates and K. Arehart, “Coherence and the speech intelligibility index,” J. Acoust. Soc. Am., vol. 117, no. 4 Pt 1, pp. 2224–2237, 2005.

[3] 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, 2004.

[4] 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, 2009.

[5] F. Chen and P.C. Loizou, “Analysis of a simplified normalized covariance measure based on binary weighting functions for predicting the intelligibility of noise-suppressed speech,” J. Acoust. Soc. Am., vol. 128, no. 6, pp. 3715–3723, 2010.

[6] F. Chen and P. C. Loizou, “Speech enhancement using a frequency-specific composite Wiener function,” in Proc. 35th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Dallas, TX, 2010, pp. 4726–4729.

[7] T.H. Falk, C.X. Zheng, and W.Y. Chan, “A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech,” IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 7, pp. 1766–1774, 2010.

[8] J.F. Santos, M. Senoussaoui, and T.H. Falk, “An improved non-intrusive intelligibility metric for noisy and reverberant speech,” Proc. 14th International Workshop on Acoustic Signal Enhancement, pp. 55–59, Juan les Pins, 2014.

[9] F. Chen, O. Hazrati, and P.C. Loizou, “Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure,” Biomed. Sig. Proc. Control, vol. 8, no. 3, pp. 311–314, 2013.

[10] F. Chen, “Predicting the intelligibility of noise-corrupted speech non-intrusively by across-band envelope correlation,” Biomed. Sig. Proc. Control, vol. 24, pp. 109–113, 2016.

[11] F. Chen, L. L. Wong, J. Qiu, Y. Liu, B. Azimi, and Y. Hu, “The contribution of matched envelope dynamic range to the binaural benefits in simulated bilateral electric hearing,” Journal of Speech, Language, and Hearing Research, vol. 56, no. 4, pp. 1166, 2013.

[12] IEEE, “IEEE recommended practice for speech quality measurements,” IEEE Trans. Audio Electroacoust., vol. 17, no. 3, pp. 225–246, 1969.

[13] ANSI, “Methods for calculation of the speech intelligibility index,” American National Standards Institute, New York, S3.5–1997.

[14] E.W. Healy, A. Kannabiran, and S.P. Bacon., “An across-frequency processing deficit in listeners with hearing impairment is supported by acoustic correlation,” J. Speech, Language, and Hearing Research, vol. 48, no. 5, pp. 1236–1242, 2005.