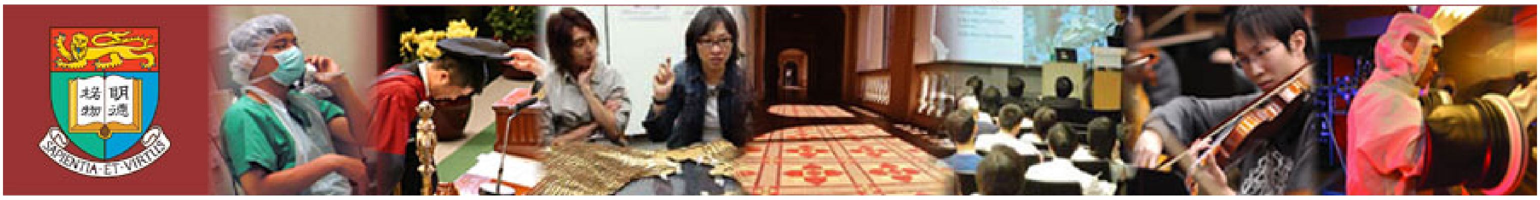

Non-intrusive Intelligibility Prediction for Mandarin Speech in Noise

Fei Chen
Division of Speech and Hearing Sciences
The University of Hong Kong, Hong Kong

Tian Guan
Graduate School at Shenzhen, Tsinghua University,
Shenzhen, China

***Abstract*–**Most existing intelligibility indices require access to the input (clean) reference signal to predict speech intelligibility in noise. In some real-world applications, however, only the noise-masked speech is available, rendering existing indices of little use. The present study assessed the performance of an intelligibility measure that could be used to predict non-intrusively (i.e., with no access to the clean input signal) speech intelligibility in noise using only information extracted from the noise-masked speech envelopes. The proposed intelligibility measure (denoted as ModA) was computed by integrating the area of the modulation spectrum (within 0.5 Hz to 10 Hz) of the noise-masked envelopes extracted in four acoustic bands. The ModA measure was evaluated with intelligibility scores obtained by normal-hearing listeners presented with Mandarin sentences corrupted by three types of maskers. High correlation (r=0.90) was obtained between ModA values and listener's intelligibility scores, suggesting that the modulation-spectrum area could be potentially used as a simple but efficient predictor of speech intelligibility in noisy conditions.

***Keywords*–**Non-intrusive intelligibility index, intelligibility prediction.

I. INTRODUCTION

Intelligibility indices, such as the speech intelligibility index (SII) [1], coherence-based speech intelligibility index (CSII) [2] or speech transmission index (STI) [3-4], require access to the clean reference signal to predict speech intelligibility. For the computation of the SII index, for instance, the clean reference signal is needed in order to compute the signal-to-noise ratio (SNR) in each frequency band. For the CSII measure, the reference signal is needed in order to compute the signal-to-distortion ratio and for the STI measure the reference signal is needed in order to compute the modulation transfer function [4-5]. It should be noted that the STI measure, in its original formulation [3], does not require access to the clean reference speech signal, as it makes use of artificial test signals (i.e., sinusoidally modulated noise) as input. Other speech intelligibility measure includes the spectro-temporal modulation index (STMI) [6]. These methods have been shown to work successfully in predicting speech intelligibility in a variety of acoustic conditions involving noise, reverberation, or both.

In clinical applications where the main interest is to predict hearing-impaired listener's ability to recognize speech in noise, there is access to the reference clean signal. In certain realistic telecommunication applications (e.g., quality of service monitoring), however, we do not have access to the clean reference signal, but only the noise-masked signal. This presents a challenge with existing intelligibility measures. While a number of blind techniques, also referred to as non-intrusive techniques, have been proposed to predict the quality of speech transmitted over VoIP networks, processed via low-rate codecs [7-8] or subjected to reverberation [9], there are limited known measures that can predict blindly (i.e., with no access to the reference signal) speech intelligibility in noise.

Chen et al. recently proposed a non-intrusive intelligibility index (denoted as ModA) for predicting the intelligibility of reverberant speech [10]. The proposed measure is rooted in the basic principle 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. Adding noise to speech, for instance, causes a reduction in modulations across all modulation frequencies, while the effect of reverberation is modulation-frequency dependent and acts like a low-pass filter on the modulation spectrum. The relative reduction in modulations is evident when plotting the modulation spectra of corrupted (by masking noise) speech envelopes. Figure 1 shows the modulation spectra of the noise-masked speech envelopes computed for an acoustic frequency band and at different SNR levels. It is seen that, as more noise is added to the signal, the modulation spectrum of the noise-masked envelopes becomes flat and shifts down (across all modulation frequencies) relative to the modulation spectrum of the clean envelope [11]. Consequently, the area under the modulation spectrum is reduced as the SNR level decreases. This suggests that the modulation area can serve as an index of speech intelligibility, since it potentially reflects the reduction in speech modulation due to the additive noise.

The non-intrusive intelligibility index ModA was found to successfully predict the intelligibility of reverberant speech [10]. In present study, we will further test the hypothesis that the modulation-spectrum area can be used as a predictor of speech intelligibility in noisy conditions. The intelligibility prediction power of the ModA measure will be evaluated using a total of 24 noisy conditions at several SNR levels in three types of real-world environments [speech-shaped noise (SSN), babble, and street].

This research was supported by Faculty Research Fund, Faculty of Education, The University of Hong Kong, by Seed Funding for Basic Research. This work was also supported Grant 31271056 from National Natural Science Foundation of China.

978-1-4799-2827-9/13/$31.00 ©2013 IEEE

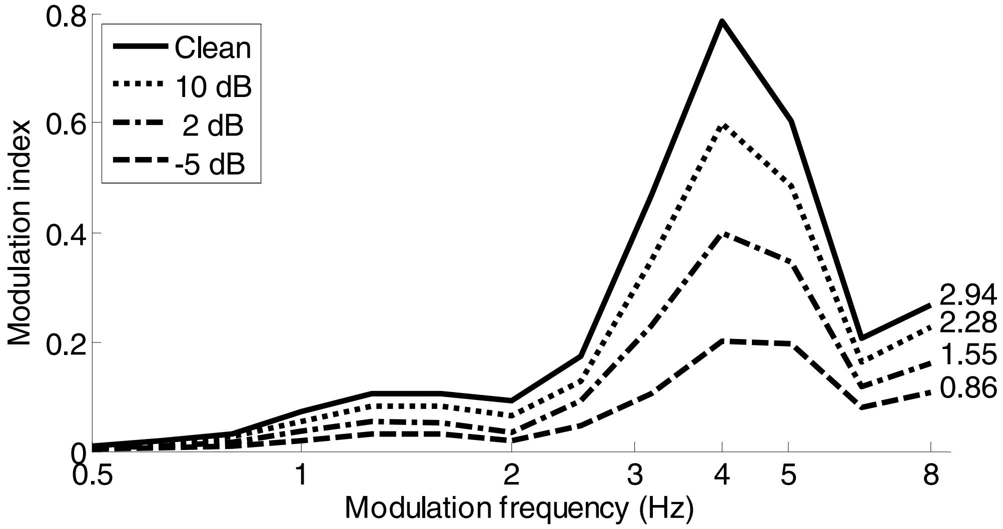

Fig. 1. Speech modulation spectra computed in four SNR levels for an acoustic frequency band spanning 775–1735 Hz. Numbers at right of each curve indicate the area of the corresponding modulation spectrum.

II. SPEECH INTELLIGIBILITY DATA

A. Subjects

Nine (five male and four female) normal-hearing (NH) subjects participated in the listening tests. All subjects were native speakers of Mandarin Chinese, and were paid for their participation. The subjects' age ranged from 18 to 34 yrs.

B. Materials

The speech material consisted of Mandarin sentences taken from the Sound Express database [12]. All the sentences were produced by a female speaker, with fundamental frequency of 230 ± 65 Hz. The masker signals included speech-shaped noise, and two real-world recordings from different places: babble (cafeteria) and street. The sentences and three maskers were sampled at 16 kHz. Segments randomly selected from the masker signals were added to the Mandarin sentences at different SNR levels, i.e., (–12, –10, –8, –6, –4, –2, and 0 dB) for the babble masker, (–14, –12, –10, –8, –6, –4, –2, and 0 dB) for the SSN masker, and (–14, –12, –11, –10, –8, –6, –4, –2, and 0 dB) for the street masker. SNR levels were chosen with each type of masker, based on pilot data collected from one subject, to avoid ceiling effects.

C. Procedure

The noise-masked sentences were presented binaurally to the listeners in a double-walled sound-proof booth via a circumaural headphone at a comfortable listening level. Twenty sentences were used for each condition, and none of the sentence lists were repeated. NH listeners participated in a total of 24 conditions (i.e., 7, 8, and 9 SNR levels for the babble, SSN and street maskers, respectively). The order of the test conditions was randomized across subjects. Subjects were given a 5-min break every 30 minutes of testing. The intelligibility score was calculated in percentage by dividing the number of words correctly recognized by the total number of words in a particular testing condition.

III. SPEECH INTELLIGIBILITY MEASURE

The implementation of the non-intrusive measure is as follows [10]. The waveform of the noise-masked signal is first limited within a fixed amplitude range (i.e., [–0.8, 0.8]), and decomposed into N frequency bands within the signal bandwidth (300–7600 Hz in this study) by using a series of fourth-order Butterworth filters. The center frequencies of Butterworth filters were spaced along the cochlear frequency map in equal steps according to the cochlear frequency-position function [13]. The temporal envelope of each band is computed using the Hilbert transform, and down-sampled to the rate of $2\times f_{\mathrm{cut}}$ Hz to limit the envelope modulation rate to f_{cut} Hz. The mean-removed envelope is then band-pass filtered through 1/3-octave-band spaced six-order Butterworth filters with center frequencies ranging from 0.5 to 10 Hz. The mean-removed root-mean-square output of each 1/3-octave band is subsequently computed to form the modulation spectrum of each acoustic frequency band, as exemplified in Fig. 1. The 13 modulation indices covering 0.5–10 Hz are summed up to yield the area $\mathrm{A_i}$ under the modulation spectrum of each acoustic frequency band. Finally, averaging the $\mathrm{A_i}$ values across all frequency bands leads to the average modulation-spectrum area (ModA) as:

$$\mathrm{ModA} = \frac{1}{N}\sum_{\mathrm{i=1}}^{N} \mathrm{A_i}, \qquad (1)$$

where ModA denotes the average modulation area, and $N = 4$ is the number of frequency bands used in the present study. The lower cutoff frequencies of the four band-pass filters were (300, 775, 1735, and 3676 Hz).

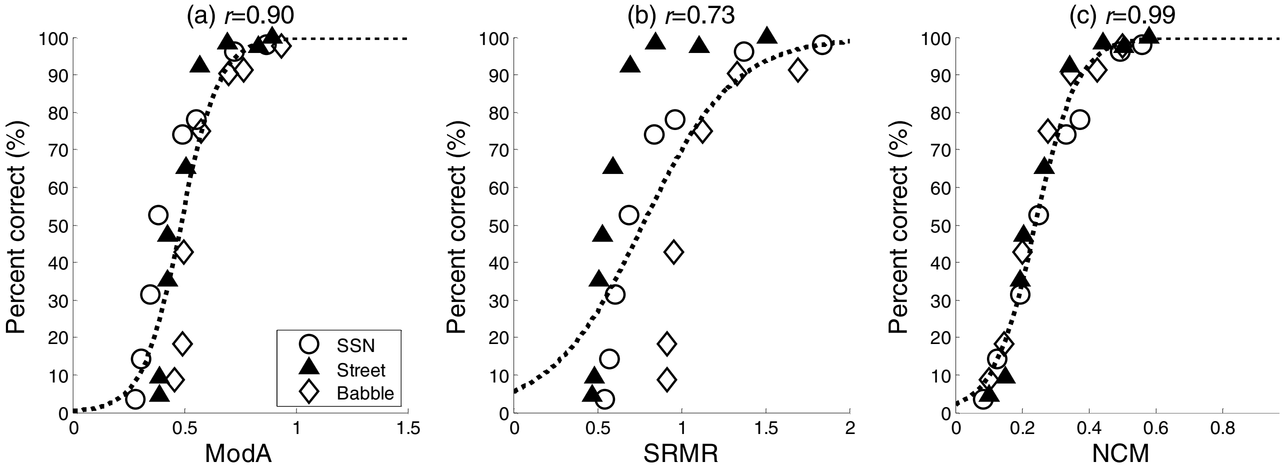

Fig. 2. Scatter plots of sentence recognition scores against the (a) ModA, (b) SRMR, and (c) NCM values.

IV. RESULTS

The Pearson's correlation coefficient (*r*) and the standard deviation of the prediction error (σ_e) were used to assess the performance of the speech intelligibility measures. The average intelligibility scores obtained by NH listeners in each condition were subjected to correlation analysis with the corresponding average values obtained in each condition by the intelligibility measures.

For comparative purposes, the SRMR (or speech to reverberation modulation energy ratio) measure [9] and the normalized covariance measure (NCM) [14] were also evaluated in the present study. The SRMR measure is a non-intrusive measure for predicting the quality and intelligibility of reverberant and dereverberated speech [9]. The NCM index is an intrusive speech-based STI measure, and is implemented here with $N = 20$ bands and modulation rate f_{cut} =100 Hz [15]. Table 1 shows the resultant correlations coefficients of the ModA, SRMR, and NCM measures with sentence intelligibility scores. The ModA measure well predicts the intelligibility scores, i.e., r=0.90. The highest correlation (r=0.99), with the smallest prediction error (5.9%), is obtained with the NCM index. Figure 2 (a), (b) and (c) show the scatter plots of sentence recognition scores against the ModA, SRMR and NCM values, respectively. A logistic function was used to map the intelligibility values to sentence intelligibility scores in Fig. 2. Statistical test [16] reveals that the correlation coefficient of the ModA measure is significantly (p<0.05) higher than that obtained with the SRMR measure (i.e., r=0.90 vs. 0.73), but significantly lower than that obtained with the NCM measure (i.e., r=0.90 vs. 0.99).

Further analysis was performed to assess the influence of modulation rate (f_{cut}) and number of bands (N) on the prediction power of the ModA measure. To assess whether including higher modulation frequencies (e.g., f_{cut}>10 Hz) would improve the correlation performance of the ModA measure, we examined the correlations obtained with modulation frequencies up to 50 Hz. The correlations obtained with different modulation rates are shown in Table 2. As can be seen, there is no improvement in correlation when the modulation rate increases. Table 2 also shows the correlations obtained when the number of bands N increases to 20. As observed in Table 2, there is slight improvement in correlation coefficient when the number of bands increases. Though the correlation coefficient is increased to 0.92 when using a larger number of bands (i.e., N=20), statistical test indicates that this correlation difference is not significant (p>0.05).

Table 1. Correlation coefficients (*r*) and standard deviations of the prediction error (σ_e) between sentence recognition scores and the ModA, SRMR, and NCM values. Asterisk denotes that the correlation coefficient is significantly (p<0.05) different from that of the ModA measure.

	ModA	SRMR	NCM
r	0.90	0.73 *	0.99 *
σ_e (%)	15.4	24.5	5.9

Table 2. Correlation coefficients (*r*) between sentence recognition scores and the ModA values.

Modulation rate (N=4)		Number of bands (f_{cut}=10 Hz)	
f_{cut} (Hz)	*r*	*N*	*r*
10	0.90	4	0.90
20	0.88	8	0.92
30	0.87	16	0.92
50	0.85	20	0.92

V. DISCUSSION AND CONCLUSION

In this study, the SRMR measure failed to well predict the intelligibility of the noise-masked Mandarin sentences (i.e., r=0.73). This might be attributed to the following possible reasons: 1) the SRMR index was proposed for predicting the intelligibility of reverberant and dereverberated speech, and probably needs to be optimized for predicting speech intelligibility in noise; 2) in computing the SRMR measure, a simple energy thresholding voice activity detection (VAD) algorithm was used to remove silence segments longer than 50ms, and should be replaced if noisy files with low SNR levels (e.g., from –14 to 0 dB in this study) are to be tested [9]; and 3) language effect might partially account for the deficiency of the SRMR measure in predicting the intelligibility of Mandarin sentences in noise. For instance, Chen recently evaluated the performance of several objective quality indices in predicting the intelligibility of vocoded speech, and found that language effect had an impact on the performance of predicting the intelligibility of vocoded Mandarin and English speech [17].

In conclusion, an intelligibility index (i.e., ModA) that requires no access to the clean (reference) signal was examined for predicting the intelligibility Mandarin speech in noise. Analysis of the data indicated that a relatively high correlation (r=0.90) was obtained between ModA values and NH listener's intelligibility scores in noise. Consistent with previous findings [10], the number of acoustic bands used in the computation of the ModA measure did not have a significant effect on the prediction power of the ModA measure (Table 2), and four acoustic bands were found to be sufficient. The findings in this study suggest that the modulation-spectrum area, capturing the reduction in envelope modulations caused by additive noise, can be used as a simple but efficient non-intrusive predictor of intelligibility in noisy conditions. However, the intelligibility prediction performance of the ModA measure is still relatively poorer (i.e., r=0.90 vs. 0.99) than that obtained with the traditional STI-based intelligibility index (i.e., NCM) that makes use of the input (clean) reference signal. This suggests that much work still needs to be done in order to further improve the predication power of the ModA measure.

REFERENCES

[1] ANSI, "Methods for calculation of the speech intelligibility index," S3.5–1997 (American National Standards Institute, New York).

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

[3] H.J. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Am.*, vol. 67, pp. 318–26, Jan. 1980.

[4] T. Houtgast and H.J. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, vol. 77, pp. 1069–1077, 1985.

[5] R. Drullman, J.M. Festen, and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Am.*, vol. 95, pp. 2670–2680, May 1994.

[6] M. Elhilali, T. Chi, and S. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Commun.*, vol. 41, pp. 331–348, Oct. 2003.

[7] ITU-T P.563, "Single-ended method for objective speech quality assessment in narrowband telephony applications," 2004 (International Telecommunication Union).

[8] D.S. Kim, "ANIQUE: An auditory model for single-ended speech quality estimation," *IEEE Trans. Speech Audio Process.*, vol. 13, pp. 821–831, Sep. 2005.

[9] 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, pp. 1766–1774, Sep. 2010.

[10] 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, pp. 311–314, May 2012.

[11] F. Dubbelboer and T. Houtgast, "The concept of signal-to-noise ratio in the modulation domain and speech intelligibility," *J. Acoust. Soc. Am.*, vol. 124, pp. 3937–3946, Dec. 2008.

[12] TigerSpeech Technology: http://www.tigerspeech.com/

[13] D. Greenwood, "A cochlear frequency-position function for several species – 29 years later," *J. Acoust. Soc. Amer.*, vol. 87, no. 6, pp. 2592–2605, Jun. 1990.

[14] R.L. Goldsworthy and J.E. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Am.*, vol. 116, pp. 3679–3689, Dec. 2004.

[15] F. Chen and P.C. Loizou, "Predicting the intelligibility of vocoded and wideband Mandarin Chinese," *J. Acoust. Soc. Am.*, vol. 129, pp. 3281–90, May 2011.

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

[17] F. Chen, "Predicting the intelligibility of cochlear-implant vocoded speech from objective quality measure," *J. Med. Biolog. Eng.*, vol. 32, pp. 189–193, 2012.