A TIME-WEIGHTED METHOD FOR PREDICTING THE INTELLIGIBILITY OF SPEECH IN THE PRESENCE OF INTERFERING SOUNDS

Mingjie Song[1], *Fei Chen*[2], *Xihong Wu*[1], *Jing Chen*[1]

1.Department of Machine Intelligence, Speech and Hearing Research Center, and Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, China,100871
2.Department of Electrical and Electronic Engineering, Southern University of Science and Technology, No. 1088 Xueyuan Blvd., Shenzhen, China, 518055

ABSTRACT

The speech intelligibility index (SII) has been widely used as an objective method of predicting speech intelligibility, but its traditional form is most effective predicting speech intelligibility scores under stationary noise but not more challenging conditions (e.g., competing noise interference). To address this limitation, the present work extended the SII model to predict the intelligibility of speech in both steady speech-spectral noise (SSN) and dual-talker speech (DTS), by using a time-weighted function that accounted for the relative perceptual importance of vowels and consonants in speech intelligibility. The performance of the new time-weighted SII (TW-SII) was compared to the other two well-known methods, i.e., the time-averaged SII (TA-SII) and coherence SII (CSII). Experimental results showed the intelligibility prediction accuracy of the three methods was similar for speech in SSN, but the prediction by TW-SII was more accurate than those by TA-SII and CSII for speech in DTS. The possible applications and limitations of the present intelligibility model were analyzed and discussed.

Index Terms— speech intelligibility index, time-weighted, vowel, consonant

1. INTRODUCTION

Speech intelligibility is a measure of the effectiveness of speech comprehension, and it is an important index for the evaluation of listening environments, communication devices, or hearing treatment. The desire to evaluate speech intelligibility objectively has led to the development of several computational models. Among many, two representative models have been developed as international standards: speech intelligibility index (SII)[1] and speech transmission index (STI)[2]. The former model mainly addresses the effect of additive noise and bandwidth reduction, and the latter model mainly addresses the reverberation effect.This work focused on the effect of additive noise, hence it was based on SII.

The basic idea of SII is that the intelligibility of the speech signal is the sum of the contributions of several frequency bands [3]. The equation is given as

$$SII = \sum_{i=1}^{n} W_i \times A_i \tag{1}$$

where n is the number of frequency bands, and W_i and A_i are the values of the importance function and the audibility function at the ith band, respectively[4]. In the calculation procedure, the speech and noise signals firstly pass a set of bandpass filters, respectively. The signal-to-noise ratio (SNR) is calculated in each frequency band. Then, the SNR value is transferred to an audibility index (A_i) between 0 and 1 by normalizing SNRs to the dynamic range of the speech level, and finally the SII value is determined by a weighted summation of the audibility indexes across frequency bands with the band importance function. In the standard form, the SNR value is evaluated according to the long-term spectrum of the signals, which is reasonable when the additive noise is steady in amplitude.

When the interfering sounds fluctuate, listeners are able to catch glimpses of the speech during the short silent periods of the masking noise, leading to improved speech intelligibility. However, this effect is not taken into account for the traditional SII model, since the model is independent of the amount of fluctuation in the noise signal. To address the issue above, Rhebergen et al. extended the SII by splitting the signal in each frequency band into short-term frames [5][6]. The frame length varied across frequency bands. The calculation of A_i for each frame was the same as before, and denoted as $A_i(t)$ in equation (2). The A_i value of equation (1) was determined by averaging $A_i(t)$ values across all frames, using the following equation

$$SII = \sum_{i=1}^{n} [W_i \times \frac{1}{T} \sum_{t=1}^{T} A_i(t)] \tag{2}$$

The work was supported by the National Natural Science Foundation of China (No.61473008, No.61771023 and No.11590773), and a Newton alumni funding by the Royal Society, UK.

where T is the number of all frames, and $A_i(t)$ represents the A_i value for frame t of the ith frequency band. Experimental results showed that the prediction by this extended model was better than that of the traditional SII model for sinusoidally intensity modulated noise and real speech or speech-like maskers, but it was still not accurate [7].

In principle, the importance of speech segments for intelligibility varies along time according to their phoneme categories [8][9][10][11]. Cole et al. studied the relative contribution of vowels and consonants to sentence intelligibility in English [8]. When the vowel or consonant segments were replaced with speech-shaped noise within a sentence, it was found that the vowel-only sentences had a higher intelligibility than the consonant-only sentences. Kewley-Port et al. also confirmed the intelligibility advantage of vowels for both young normal-hearing listeners and elderly hearing-impaired listeners [9]. Chen et al. investigated the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility, and reported similar results that vowels were more important than consonants [10]. These speech perception findings suggest that it is necessary to explore a time-weighted method for speech intelligibility prediction when the SII model was applied.

To take account of the relative perceptual importance of vowels and consonants in sentences, the traditional SII model was extended with a time-weighted function to predict speech intelligibility in steady speech-spectrum noise (SSN) and in the typical competing speech, dual-talker speech (DTS). Details of the new method will be described in Section 2. The performance of the new method was evaluated by analyzing the correlation between the predicted scores and human listeners scores, and the correlation was also compared with two existing methods. The experiments and results will be described in Sections 3 and 4, respectively. Finally, the main findings will be discussed and concluded in Section 5.

2. MODEL

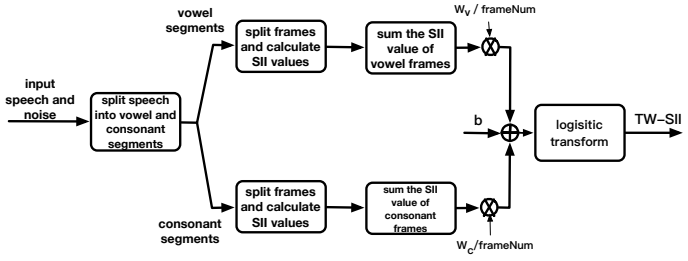

Fig. 1. Flow diagram for the calculation of the TW-SII model. A detailed description is given in the main text.

The new time-weighted speech intelligibility index (TW-SII) was calculated in this way: firstly signals were divided into vowel and consonant segments, and then they were split into frames with 8-ms frame length. The SII value of each frame was calculated using the traditional SII model, and they were summed across frames for vowels and for consonants, respectively, denoted as $SIIv$ and $SIIc$; finally the SII of the sentence was calculated by equations (3) and (4). Figure 1 shows the flow diagram of this procedure. The basic idea to conduct the time-weighted function was to allocate different weight to vowels and consonants when the SII values were integrated across time. A similar time weighted procedure was proposed for estimating the speech intelligibility by Kates et al., called coherence speech intelligibility index(CSII) [12], in which signals were divided into three categories (low-, medium-, and high-level) according to signal amplitude. We adopted and modified their equations for the time-weighted manipulation in this work.

A linear weighting of the vowel and consonant SII values was used as equation (3). Then the weighted sum was transformed to a SII value between 0 and 1 using a logistic function as equation (4). The weights were constrained to be positive. The constrained optimization method was used to obtain the weight value by fitting the model prediction scores and real listener intelligibility scores [12].

$$sum = (W_c * SII_c + W_v * SII_v)/frameNum + b \quad (3)$$

$$SII = \frac{1}{1 + exp^{-sum}} \quad (4)$$

where SII_v and SII_c represent the SII values of vowels and consonants, W_v and W_c represent the weight of vowels and consonants, $frameNum$ represents the total number of frames of a sentence, and b represents a bias caused by speech materials.

3. EXPERIMENT

Firstly, a human behavioral experiment was conducted to collect data for the fitting procedure and obtaining the weight values. Secondly, the performance of the new model was evaluated by comparing the scores predicted by the model and listener scores, and to the performance of the time-averaged SII model (TA-SII) and coherence SII model (CSII).

3.1. Subjective experiment

3.1.1. Subjects and materials

Twelve native-Mandarin-Chinese listeners with normal hearing (age range 18 to 25 years) participated in the experiment. There were 4 men and 8 women participants. They were all students of Peking University and paid for their participation. The speech material consisted of sentences taken from the Mandarin Speech Perception (MSP) corpus, which includes 10 lists of 10 sentences, with 7 words per sentence [10]. All of these sentences were spoken by a female speaker. In addition, the start and end sampling points of vowels, consonants

and silence were labeled for each of the sentences in this corpus, which is useful for the calculation of TW-SII. They are labeled manually by experienced phoneticians based on the acoustic landmarks observed in the spectrograms. There were two types of maskers SSN and DTS. A finite impulse response filter was designed based on the average long-term spectrum of the MSP sentences, and white noise was filtered to produce the SSN signal. The DTS masker contained two equal-level interfering female talkers.

3.1.2. Procedure

The experiment was performed in an anechoic chamber and the sound was played to listeners through a loudspeaker (Dynaudio Acoustics, BM6A). The sound level of the target speech was calibrated at 55 dB SPL. The masker signal displayed one second earlier than the target, and they ended simultaneously. Before the formal experiment, each subject participated in a 10-min training session and was given two lists of ten MSP sentences. There were four SNRs for each type of masker: -10, -8, -6, and -4 dB for SSN; -8, -6, -4 and 0 dB for DTS. Totally, eight conditions (2 masker types$\times$4 SNRs) were tested for each subject. During the test, subjects listened to each trial and verbally repeated the words they heard. Their responses were recorded and the correct words were counted by the experimenter.

3.2. Objective intelligibility evaluation

Signals were filtered into 1/3-octave bands for the following three objective models when the SII values were calculated by equation (1). As the speech used in this study was Mandarin Chinese, the frequency important function based on Mandarin was used for all models[13].

3.2.1. TW-SII

The method described in Section 2 was used for the TW-SII model. The data collected from the subjective experiments were divided two groups: the data of 8 subjects were used as training set to fit the weight values, and the rest of the data from 4 subjects were used as testing set to compare the scores predicted by the model with listener scores.

3.2.2. TA-SII

Since the temporal resolution of the auditory system is frequency dependent [14], time constants for the lower frequency bands are larger than those for the higher bands. Frame lengths ranging from 64ms at the lowest band (160Hz) to 2ms at the highest band (8000Hz) was used [15]. SII was calculated by equation (2).

3.2.3. CSII

The experiment based on CSII used the same training set and test set as TW-SII. The calculation procedure of CSII was also similar as TW-SII, except that 1) the frame length was 16ms; 2) three categories were defined as high- ($L_{frame} \geq L_{sentence}$), middle-($L_{frame} - L_{sentence} \geq -10dB$), and low-level ($L_{frame} - L_{sentence} < -10dB$), where L_{frame} and $L_{sentence}$ represent the RMS level of the frame and the whole sentence, respectively; 3) SII calculation was replaced by the coherence calculation. More details of this method can be found in [12].

4. RESULTS

4.1. Subjective experiment

Figure 2 shows the mean scores (correct proportion) averaged across 12 subjects as a function of SNRs for both masking conditions. It was obvious that speech intelligibility varied in a reasonable range. Speech intelligibility was higher in SSN than in DTS at a certain SNR, since prominent information masking was included in the latter masker [16][17][18]. Speech intelligibility in SSN was more sensitive to the variation of SNR than in DTS, since the former masker was mainly dominant by energetic masking. These results were consistent with previous studies [19].

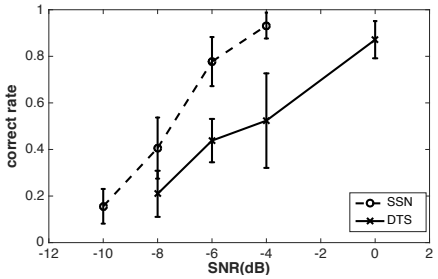

Fig. 2. Average scores as a function of SNR for SSN (dotted curve) and DTS masking (solid curve), respectively. Error bars represent standard deviation.

4.2. TW-SII

The weight values fitted for each of the masker types are shown as follows

$$SSN \qquad sum = 28.17 * SII_c + 74.70 * SII_v - 10.39 \quad (5)$$

$$DTS \qquad sum = 32.65 * SII_c + 15.68 * SII_v - 6.43 \quad (6)$$

For SSN masking, the consonant weight was much smaller than the vowel weight, indicating vowels contributed more than consonants. However, the relative importance between vowels and consonants was reversed for DTS masking. This

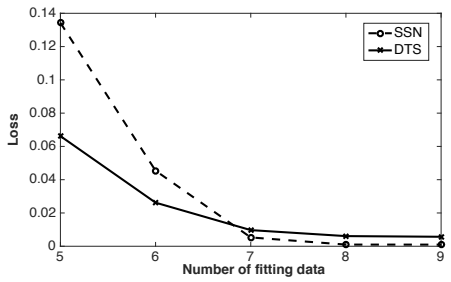

Fig. 3. The fitting function converges with the increase of the amount of fitting data. The number represents the amount of subjects.

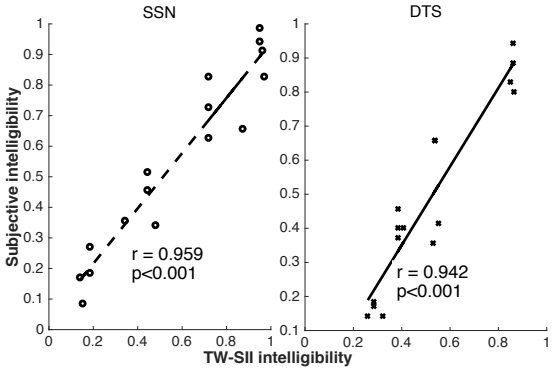

Fig. 4. The proportion of the MSP sentences identified correctly plotted versus the TW-SII model prediction.

pattern fit our assumption, because in SSN masking, consonants were heavily masked due to their low energy, but in DTS masking consonants could become important due to the increasing chances for glimpses of consonant in competing speech [20].

To test the suitability of the testing set size, the mean square error between the subjective intelligibility and the model predicted intelligibility was calculated as a function of the amount of data, as shown in Figure 3. This result indicated the weight values fitted from 8 subjects data were reliable.

Figure 4 shows the correct proportion identified by listeners versus the TW-SII model prediction for the test set. Pearsons correlation analysis indicated that the predicted scores were highly correlated with subjective scores for both SSN (r=0.959, p$<$0.001) and DTS masking (r=0.942, p$<$0.001).

4.3. Comparison with other objective intelligibility models

Similarly, the correlation analysis between the predicted scores and the subjective scores was also conducted for the result of the TA-SII model and CSII model. Table 1 shows the correlation values for each of the three models and each of the masker types. For SSN masking, the correlation was slightly higher for TW-SII than for TA-SII and CSII; for DTS masking, the correlation was markedly higher for TW-SII, indicating an advantage of the new TW-SII model.

Table 1. Correlations between the predicted scores and subjective scores for each model and each masker type. p value was less than 0.001 for all of them. Asterisk denotes that the correlation coefficient was significantly larger than that of TA-SII or CSII measure.

	SSN	DTS
TW-SII	0.959	0.942 *
TA-SII	0.911	0.873
CSII	0.957	0.851

5. DISCUSSION AND CONCLUSION

Earlier work has shown that speech segments (e.g., vowels and consonants) contained different amounts of intelligibility information [8][9][10][11]. Vowel-only sentences are more intelligible than consonant-only sentences. However, this important guideline on segmental contribution to speech intelligibility has not been reflected in existing intelligibility model. While many intelligibility models studied the effect of band-importance function to intelligibility [7][21], this study developed a new time weighted intelligibility model, which was based on the traditional SII model but incorporated the perceptional importance of vowels and consonants contained in speech signal.

The present TW-SII model showed benefit for predicting speech intelligibility in the presence of interfering sounds, especially in DTS masking, because the relative importance of consonants and vowels was introduced in this model. Although the effect of fluctuation existing in maskers was taken into account by calculating frame-based SII in the TA-SII model [5][6], the manipulation of averaging SII across frames smeared the relative importance of temporal segments, e.g., vowels and consonants in this work. In the CSII model, the temporal segments were divided into groups as high-, medium, and low-levels [12]. This partition was based on the purely physical characteristic rather than the perceptual characteristics used in this study. Hence, the prediction by the latter one could be more accurate in DTS masking.

The conduction of TW-SII relied on the labeled information of vowels and consonants, which limit its application for a general speech corpus. However, it is possible to tag the segments automatically. We initially tried to obtain the labels automatically by combining cues of F0 contour, signal envelope and energy, and the TW-SII still showed promising performance. Further work is required to accomplish this totally automatic model.

6. REFERENCES

[1] ANSI ANSI, "S3. 5-1997, methods for the calculation of the speech intelligibility index," *New York: American National Standards Institute*, vol. 19, pp. 90–119, 1997.

[2] IEC STI Standard, "IEC 60268-16 objective rating of speech intelligibility by speech transmission index," 2005.

[3] H. Fletcher and J. C. Steinberg, "Articulation testing methods," *Bell Labs Technical Journal*, vol. 8, no. 4, pp. 806–854, 1929.

[4] K.D. Kryter, "Methods for the calculation and use of the articulation index," *Journal of the Acoustical Society of America*, vol. 34, no. 11, pp. 1689–1697, 1961.

[5] K.S. Rhebergen and N.J. Versfeld, "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2181–2192, 2005.

[6] K.S. Rhebergen, N.J. Versfeld, and W.A. Dreschler, "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," *Journal of the Acoustical Society of America*, vol. 120, no. 6, pp. 3988–97, 2006.

[7] J. Ma, Y. Hu, and P. C. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions.," *Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3387–405, 2009.

[8] R. A. Cole, Y.H. Yan, B. Mak, and M. Fanty, "The contribution of consonants versus vowels to word recognition in fluent speech," in *ICASSP*, 1996, pp. 853–856 vol. 2.

[9] D. Kewley-Port, T. Z. Burkle, and J. H. Lee, "Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners," *Journal of the Acoustical Society of America*, vol. 122, no. 4, pp. 2365–2375, 2007.

[10] F. Chen, L.LN. Wong, and E.YW. Wong, "Assessing the perceptual contributions of vowels and consonants to mandarin sentence intelligibility," *Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. EL178–EL184, 2013.

[11] M. J. Owren and G. C. Cardillo, "The relative roles of vowels and consonants in discriminating talker identity versus word meaning," *Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1727–1739, 2006.

[12] J.M. Kates and K.H. Arehart, "Coherence and the speech intelligibility index," *Journal of the Acoustical Society of America*, vol. 117, no. 4, pp. 2224–2237, 2005.

[13] J. Chen, Q. Huang, and X.H. Wu, "Frequency importance function of the speech intelligibility index for mandarin chinese," *Speech Communication*, vol. 83, no. C, pp. 94–103, 2016.

[14] B. Moore, *An Introduction to the Psychology of Hearing, 5th Edition*, Academic Press, 2012.

[15] B. R. Glasberg and B. C. J. Moore, "A model of loudness applicable to time-varying sounds," *Journal of the Audio Engineering Society*, vol. 50, no. 5, pp. 331–342, 2002.

[16] X.H. Wu, C. Wang, J. Chen, H.W. Qu, W.R. Li, Y.H. Wu, B.A. Schneider, and L. Li, "The effect of perceived spatial separation on informational masking of chinese speech," *Hearing Research*, vol. 199, no. 1, pp. 1–10, 2005.

[17] X.H. Wu, J. Chen, Z.G. Yang, Q. Huang, M.Y. Wang, and L. Li, "Effect of number of masking talkers on speech-on-speech masking in chinese," in *INTERSPEECH*, 2007, pp. 390–393.

[18] J. Chen, H.H. Li, L. Li, X.H. Wu, and B. Moore, "Informational masking of speech produced by speech-like sounds without linguistic content," *Journal of the Acoustical Society of America*, vol. 131, no. 4, pp. 2914–2926, 2012.

[19] Y. Tang, M. Cooke, and C. Valentini-Botinhao, "Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech," *Computer Speech & Language*, vol. 35, pp. 73–92, 2016.

[20] M. Cooke, "A glimpsing model of speech perception in noise," *Journal of the Acoustical Society of America*, vol. 119, no. 3, pp. 1562–73, 2007.

[21] 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.," *Journal of the Acoustical Society of America*, vol. 128, no. 6, pp. 3715–23, 2010.