Published in final edited form as:
Biomed Signal Process Control. 2013 May ; 8(3): 311–314. doi:10.1016/j.bspc.2012.11.007.

Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure

Fei Chen[a], **Oldooz Hazrati**[b], and **Philipos C. Loizou**[b]

[a]Division of Speech & Hearing Sciences, The University of Hong Kong, Hong Kong, China

[b]Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX, USA

Abstract

Reverberation is known to reduce the temporal envelope modulations present in the signal and affect the shape of the modulation spectrum. A non-intrusive intelligibility measure for reverberant speech is proposed motivated by the fact that the area of the modulation spectrum decreases with increasing reverberation. The proposed measure is based on the average modulation area computed across four acoustic frequency bands spanning the signal bandwidth. High correlations ($r = 0.98$) were observed with sentence intelligibility scores obtained by cochlear implant listeners. Proposed measure outperformed other measures including an intrusive speech-transmission index based measure.

Keywords

Intelligibility prediction; non-intrusive measure; envelope modulation reduction

1. Introduction

Reverberation is known to change speech quality and can also impact speech intelligibility, as it blurs temporal and spectral cues and flattens formant transitions. The speech-transmission index (STI) has been shown to predict successfully the effects of reverberation, room acoustics, and additive noise (e.g., [1]–[2]). In its original form, the STI measure uses artificial signals (e.g., sinewave-modulated signals) as probe signals to assess the reduction in signal modulation in a number of frequency bands and for a range of modulation frequencies (0.6–12.5 Hz) known to be important for speech intelligibility. A number of extensions to the STI measure have been proposed requiring access to the clean reference signal [3]–[4]. However, in most real-world applications the clean reference signal is often not available. In such scenarios, a non-intrusive intelligibility measure would be more desirable. To our knowledge, no non-intrusive measure exists for predicting the intelligibility of reverberant speech.

Falk *et al*. [5] recently proposed a non-intrusive measure (i.e., SRMR, or Speech to Reverberation Modulation Energy Ratio) for predicting the subjective quality of reverberant

Address correspondence to: Fei Chen, Ph. D., Division of Speech and Hearing Sciences, The University of Hong Kong, Prince Philip Dental Hospital, 34 Hospital Road, Hong Kong, feichen1@hku.hk, Phone : (+852) 2859-0586, Fax : (+852) 2559-0060.

and de-reverberated speech. Their proposed measure was implemented in the modulation spectrum domain and compared against three standard intrusive/non-intrusive measures (i.e., ITU-T PESQ, ITU-T P.563, and ANSI ANIQUE+) on tasks of predicting coloration effects, reverberation tail effects and overall quality. In addition to quality assessment, their proposed measure was also evaluated indirectly for intelligibility prediction by correlating the output of other STI-based measures with the output of their measure. Aside from the indirect evaluation, no intelligibility listening tests were conducted to validate their measure [5].

Human listening tests are very important and necessary for the evaluation of intelligibility measures. The challenge faced, however, with evaluating the intelligibility of reverberant speech with normal-hearing listeners is that they generally perform extremely well (near 100%) even in highly reverberant environments [6]. That makes it difficult to evaluate any intelligibility measure for reverberant speech since it is necessary to include easy and difficult listening conditions spanning the whole 0–100% range. Cochlear implant (CI) listeners, on the other hand, perform poorly, even at moderate T_{60} values, when presented with reverberant speech [6]–[7]. For that reason, in the present study we are using cochlear implant listeners to evaluate the proposed non-intrusive intelligibility measure.

2. Non-intrusive intelligibility measure

The modulation transfer function, as used in the STI, is defined as the reduction in the modulation index of the intensity envelope of a reverberant (or noisy) signal relative to that of the original signal for modulation frequencies ranging from 0.5 to 16 Hz. The modulation index decreases, particularly in the higher modulation frequencies, as reverberation (T_{60}) increases owing to smearing of the envelope by the late reflections. Consequently, the area under the modulation spectrum decreases as the reverberation (T_{60}) increases. This is illustrated in Fig. 1 showing speech modulation spectra for different reverberations (T_{60}). From this, we propose the hypothesis that the modulation area can be used as a predictor for intelligibility of reverberant speech. Next, we describe the proposed non-intrusive measure which is based on the area of the modulation spectrum. Note that this measure only requires computation of the modulation spectrum of reverberant speech and does not need access to the input (anechoic) signal.

The time-domain waveform of the reverberant signal is first limited (i.e., normalized) within a fixed amplitude range (i.e., [−0.8, 0.8] in this study), and then decomposed into N bands spanning the signal bandwidth (300–7600 Hz in this study). The frequency decomposition is implemented with a series of fourth-order Butterworth filters, with center frequencies spaced along the cochlear frequency map in equal steps and computed according to the cochlear frequency-position function [8]. The temporal envelope of each band is computed using the Hilbert transform and then down-sampled to the rate of $2 \times f_{cut}$ Hz, thereby limiting the envelope modulation rate to f_{cut} Hz (e.g., f_{cut} = 10 Hz). The mean-removed envelope is band-pass filtered through 1/3-octave-band spaced filters with center frequencies ranging from 0.5 to 8 Hz. The mean-removed root-mean-square (RMS) output of each 1/3-octave band is subsequently computed to form the modulation spectrum of each acoustic frequency band (see example in Fig. 1). The 13 modulation indices (covering 0.5–10 Hz) are summed up to yield the area A_i under the modulation spectrum of each acoustic frequency band. Finally, the A_i values are averaged across all acoustic frequency bands to produce the average modulation-spectrum area (ModA) as:

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

where ModA denotes the average (across all acoustic frequency bands) modulation area, and $N = 4$ is the number of bands used in the present study. The lower cutoff frequencies of the four band-pass filters were {300, 775, 1735, and 3676 Hz}. A normalization step is also implemented to limit the values of ModA within the range [0, 1]. This is done by replacing the band A_i values in (1) with the normalized values computed as $A_i' = A_i / c_i$, where c_i are normalization constants computed by averaging the A_i values of 10 sentences in quiet (for this study, c_i = [2.83, 3.76, 2.65, 4.22]). Sentences in quiet were used since their modulation spectra have the largest areas.

Note that the normalization processing (i.e., by normalization constants c_i) is NOT essential for the ModA measure, and it only performs to restrict the values of ModA within the range of [0, 1]. We do this mainly because most present intelligibility indices (e.g., STI) are limited to the range of [0, 1]. If there is no need for a bounded ModA value (i.e., ≤ 1), this normalization procedure can be removed when computing the ModA measure. In addition, the normalization processing does NOT affect of the correlation analysis in the later section.

For comparative purposes, the SRMR measure [5] and the intrusive normalized covariance measure (NCM) [3] are also evaluated in the present study. The NCM index falls in the family of (intrusive) speech-based STI measures [3], and is implemented here with $N = 4$ bands and modulation rate f_{cut} = 10 Hz. The NCM has been shown previously [9] to predict reliably the intelligibility of noise-corrupted speech processed via speech enhancement algorithms (only the effects of additive noise were investigated in [9]).

3. Intelligibility listening tests

Sentences taken from the IEEE database were used as test material [10]. The sentences from one male talker were originally recorded at 25 kHz and down-sampled to 16 kHz for this study. Head related transfer functions (HRTFs) recorded in a 5.5 m × 4.5 m × 3.1 m (length × width × height) room with a total volume of 76.8 m^3 [11] were used to simulate most reverberant conditions (T_{60} = 1.0 s) with a 1 m distance between the single-source signal and the microphone. HRTFS recorded in a 10.06 m × 6.65 m × 3.4 m (length × width × height) room with a total volume of 227.5 m^3 [12] and a 5.5 m single-source to microphone distance were used to simulate other reverberant conditions. The initial average reverberation time of the latter experimental room (T_{60} = 0.8 s) had been reduced to T_{60} = 0.6, and 0.3 s by adding floor carpeting and absorptive panels on the walls and ceiling. The HRTFs had a direct-to-reverberant ratio (DRR) of −0.5, −3.0, −1.8, and 1.8 dB at T_{60} = 1.0, 0.8, 0.6, and 0.3 s, respectively. Speech-shaped noise with the same long-term spectrum as the sentences in the IEEE corpus was added to the reverberant signals to generate the reverberant + noisy conditions.

Eleven cochlear-implant users (fitted with the Nucleus Freedom device) were recruited for the listening tests. The stimuli were presented to the CI users unilaterally through the auxiliary input jack of the SPEAR3 processor in a double-wall sound proof booth (Acoustic Systems, Inc.). Prior to testing, the listeners adjusted the volume level to a comfortable level, and the volume level was fixed throughout the tests. The CI listeners participated in a total of 21 conditions [13], involving: (a) anechoic (quiet) condition, (b) four reverberant (T_{60} = 0.3, 0.6, 0.8, and 1.0 s) conditions, (c) four reverberant + noisy (combinations of T_{60} = 0.6 and 0.8 s with SNR = 5 and 10 dB) conditions, and (d) 12 conditions involving reverberant sentences processed via an ideal reverberant mask algorithm [7] (in T_{60} = 0.6 and 0.8 s and SNR levels of 5 and 10 dB using three different binary mask threshold values of −8, −10 and −12 dB). A total of 420 IEEE sentences were used in the listening tests (20 sentences/condition). None of the sentences was repeated across conditions. The order of the test conditions was randomized across the subjects to minimize order effects. The responses of

each individual were collected and scored offline based on the number of words identified correctly.

4. Results and discussion

The average (across all subjects) intelligibility scores obtained in 21 conditions (see Sec. 3) were correlated with the average ModA values computed from 20 sentences used in each condition. Table 1 shows the correlations of the ModA, NCM, and SRMR measures with sentence intelligibility scores. Highest correlation ($r = 0.98$), with smallest prediction error (5.5%), was obtained with the proposed ModA measure. Note that this correlation further improves to $r = 0.99$ if only the 9 non-processed conditions are used in the analysis, namely the conditions wherein the reverberant speech was not processed by the ideal reverberant mask algorithm [7]. Fig. 2(a) shows the scatter plot of sentence recognition scores against the ModA values. A logistic function was used to map the ModA values to sentence intelligibility scores as follows:

$$y = \frac{1}{1+e^{-(b_1 \times \mathrm{ModA} - b_2)}} \times 100, \quad (2)$$

where y is the predicted intelligibility score (in percent), and b_1, b_2 are the fitting parameters given in Table 2. Figs. 2(b) and 2(c) show the corresponding scatter plots for the NCM and SRMR measures. Statistical tests [14] revealed that the correlation coefficient of the ModA measure is significantly ($p < 0.05$) higher than that obtained with the SRMR measure (i.e., $r = 0.98$ vs. 0.92), but not significantly higher than that with the NCM measure (i.e., $r = 0.98$ vs. 0.96). Note that in general the shape of the logistic mapping function might differ depending on the speech material (e.g., sentences, non-sense syllables, etc.) used and whether normal-hearing listeners or cochlear-implant listeners are used as subjects. The shallow segment of the mapping function (Fig. 2(a)) spanning values of ModA < 0.5 is indicative of the difficulty CI listeners experience in reverberant conditions [7]. For normal-hearing listeners, for instance, an STI value of 0.5 would predict 70% correct [2, Fig. 9]. In contrast, a value of ModA = 0.5 predicts a 20% score for the CI listeners (Fig. 2(a)) tested in the present study.

Further analysis was done to assess the influence of modulation rate (highest modulation frequency), number of bands and speaker gender on the prediction power of the ModA measure.

4.1. Influence of modulation rate

To further assess whether including higher (>10 Hz) modulation frequencies would improve the correlation of the ModA measure, we examined the correlations obtained with modulation frequencies up to 120 Hz. The correlations obtained with different modulation rates are tabulated in Table 3. As can be seen, there was no improvement in correlation when the modulation rate increased. Though the correlation was decreased to 0.97 when using higher modulation rate (>40 Hz), statistical tests indicated that this correlation difference was not significant ($p < 0.05$).

The SRMR measure [5] computes the ratio between the sum of the average per-modulation band energy obtained from (1–4) and (5–K) modulation frequencies [5], where K is the highest modulation frequency which is adapted to the speech signal under test. The highest modulation frequency allowed in the implementation of the SRMR measure was 128 Hz. The present study demonstrates that the envelope information contained in low-frequency modulation rates (<10 Hz) and obtained using 1/3-octave resolution is sufficient for predicting non-intrusively the intelligibility of reverberant speech.

4.2. Influence of number of bands

Table 3 shows the correlations obtained when the number of bands N increased to 20. As observed in Table 3, there was no improvement in correlation when the number of bands increased. Though the correlation coefficient was decreased to 0.96 when using a larger number of bands (i.e., $N = 20$), statistical tests indicated that this correlation difference was not significant ($p < 0.05$).

The SRMR measure was implemented with a 23-channel gammatone filterbank, with center frequencies ranging from 125 Hz to nearly half the sampling rate [5]. The analysis in our present study showed that the number of filters did not significantly affect the prediction power of the ModA measure. To some extent, this outcome is also consistent with that obtained from the intrusive NCM measure [9]. Ma *et al.* found that the number of bands did not influence the performance of the NCM measure in intelligibility prediction for noise-corrupted speech processed via speech enhancement algorithms [9].

4.3. Influence of speaker gender

Figure 3 shows the averaged values of the ModA measure computed with TIMIT sentences [15] produced by 10 female speakers and 10 male speakers in four reverberant conditions. As can be seen, the ModA values are nearly the same for both genders, suggesting that the ModA measure is not influenced by the gender of the speaker. This outcome is expected, since the ModA measure is implemented in the modulation domain, and does not capture any F0-specific information from speech.

5. Conclusions

A non-intrusive intelligibility measure for reverberant (and/or reverberant + noisy) speech was proposed. The proposed measure was based on the computation of the modulation spectrum area within a narrow range of modulation frequencies (0.5–10 Hz) known to be important for speech intelligibility [9]. High correlations ($r = 0.98$) were observed with sentence intelligibility scores obtained by cochlear implant listeners. The proposed measure outperformed other measures, including an intrusive speech-based STI measure [3], [9]. When considering the implementation of the SRMR measure [5], the outcome of the present study highlights the importance of capturing the information contained in the low-frequency envelope modulations (<10 Hz) with sufficient resolution (1/3-octave).

Acknowledgments

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 Grant No. R01 DC010494 from the National Institute of Deafness and other Communication Disorders, NIH. The authors would like to thank Dr. Tiago H. Falk for providing the MATLAB implementation of the SRMR measure.

References

1. Houtgast T, Steeneken HJM. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. J Acoust Soc Amer. 1985; 77(3):1069–1077.
2. Steeneken HJM, Houtgast T. A physical method for measuring speech transmission quality. J Acoust Soc Amer. 1980; 67(1):318–326. [PubMed: 7354199]
3. Goldsworthy R, Greenberg J. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. J Acoust Soc Amer. 2004; 116(6):3679–3689. [PubMed: 15658718]
4. Payton K, Braida L. A method to determine the speech transmission index from speech waveforms. J Acoust Soc Amer. 1999; 106(6):3637–3648. [PubMed: 10615702]

5. Falk T, Zheng C, Chan WY. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. IEEE Trans Audio, Speech Signal Process. 2010; 18(7):1766–1774.
6. Nabelek AK, Pickett JM. Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners. J Speech, ang Hear Res. 1974; 17(4): 724–739.
7. Kokkinakis K, Hazrati O, Loizou PC. A channel-selection criterion for suppressing reverberation in cochlear implants. J Acoust Soc Amer. 2011; 129(5):3221–3232. [PubMed: 21568424]
8. Greenwood D. A cochlear frequency-position function for several species – 29 years later. J Acoust Soc Amer. 1990; 87(6):2592–2605. [PubMed: 2373794]
9. Ma J, Hu Y, Loizou PC. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J Acoust Soc Amer. 2009; 125(5):3387–3405. [PubMed: 19425678]
10. Loizou, PC. Speech Enhancement: Theory and Practice. Boca Raton: Florida: CRC Press LLC; 2007.
11. Van den Bogaert T, Doclo S, Wouters J, Moonen M. Speech enhancement with multichannel Wiener filter techniques in multi-microphone binaural hearing aids. J Acoust Soc Amer. 2009; 125(1):360–371. [PubMed: 19173423]
12. Neuman AC, Wroblewski M, Hajicek J, Rubinstein A. Combined effects of noise and reverberation on speech recognition performance of normal-hearing children and adults. Ear Hear. 2010; 31(3):336–344. [PubMed: 20215967]
13. Hazrati O, Loizou PC. The combined effect of reverberation and noise on speech intelligibility by cochlear implant listeners. Int J Audiol. 2012; 51(6):437–443. [PubMed: 22356300]
14. Steiger JH. Tests for comparing elements of a correlation matrix. Psychol Bull. 1980; 87(2):245–251.
15. Garofolo, J.; Lamel, L.; Fisher, W., et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus. Philadelphia: Linguistic Data Consortium; 1993.

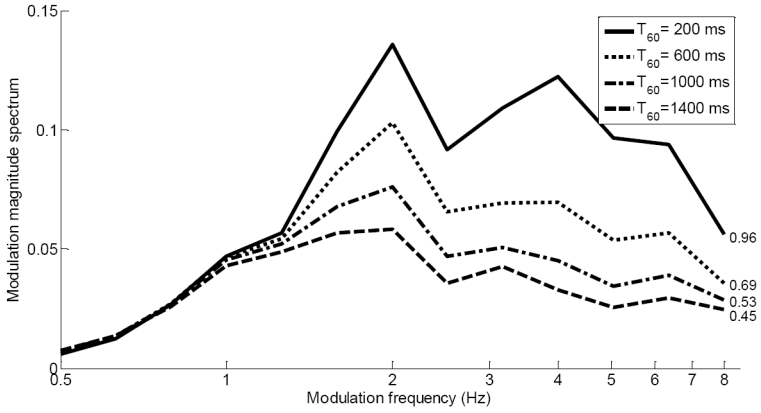

Fig. 1.

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

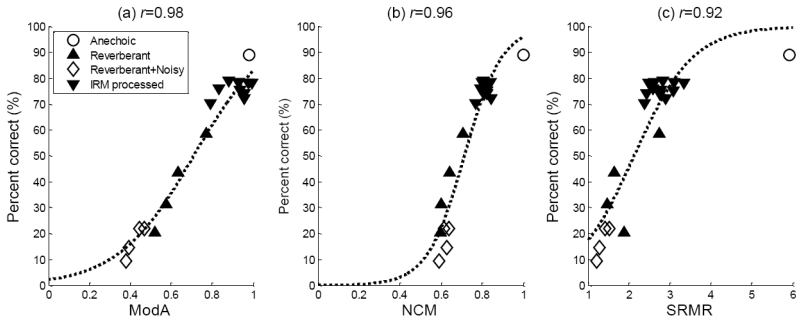

Fig. 2.

Scatter plots of sentence recognition scores against the: (a) ModA, (b) NCM, and (c) SRMR values. 'IRM' denotes the ideal reverberant mask algorithm in [7].

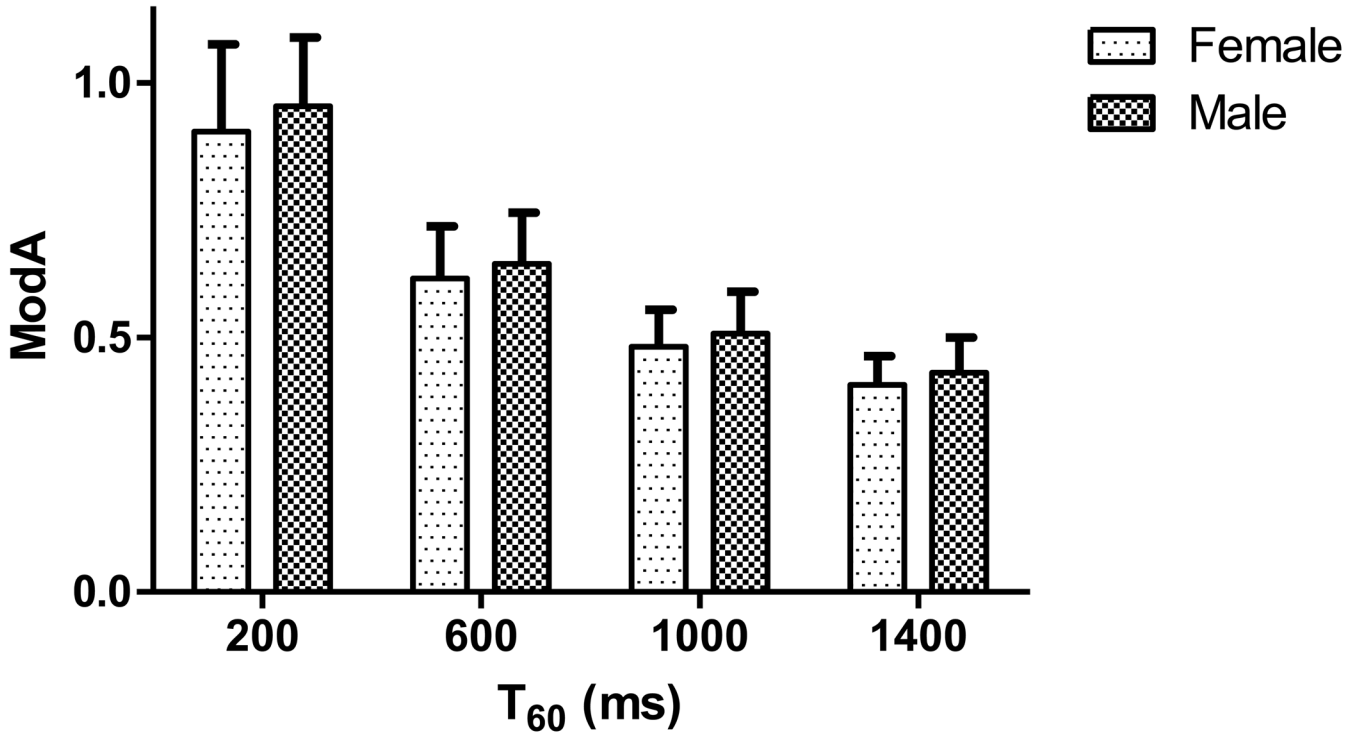

Fig. 3.
Average values of the ModA measure computed using sentences produced by 10 female speakers and 10 male speakers in four reverberant conditions.

Table 1

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

	ModA	NCM	SRMR
r	0.98	0.96	0.92*
σ_e	5.5%	7.1%	10.3%

Table 2

The fitting parameters used in the logistic function in Eq. (2) for mapping the objective ModA values to intelligibility scores (see Fig. 2).

	b_1	b_2
ModA	5.4	3.8
NCM	11.0	7.8
SRMR	1.4	3.0

Table 3

Correlation coefficients (r) between sentence recognition scores and the ModA values as a function of modulation rate (f_{cut}) and number of bands (N).

Modulation rate (N = 4)		Number of bands (f_{cut} = 10Hz)	
f_{cut} (Hz)	r	N	r
10	0.98	4	0.98
20	0.98	8	0.97
40	0.97	12	0.97
80	0.97	16	0.97
120	0.97	20	0.96