Title	Effect of temporal fine structure on speech intelligibility modeling
Author(s)	Chen, F; Guan, T; Wong, LLN
Citation	The 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS), Osaka, Japan, 3-7 July 2013. In IEEE Engineering in Medicine and Biology Society. Conference Proceedings, 2013, p. 4199-4202
Issued Date	2013
URL	http://hdl.handle.net/10722/185164
Rights	IEEE Engineering in Medicine and Biology Society. Conference Proceedings. Copyright © Institute of Electrical and Electronics Engineers.

35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Effect of Temporal Fine Structure on Speech Intelligibility Modeling

Fei Chen, Tian Guan, and Lena L. N. Wong

Abstract – **Temporal fine structure (TFS) carries important information for the speech perception of hearing-impaired listeners and for the design of novel prosthetic hearing devices. This study assessed the performance of present intelligibility indices for predicting the intelligibility of speech containing different amount of TFS information. Speech intelligibility data was collected from vocoded and wideband Mandarin sentences containing little/partial and intact TFS information, respectively, and was then subjected to the correlation analysis with existing intelligibility indices. It was found that, though performing well in predicting the intelligibility of vocoded or wideband speech separately, present intelligibility indices were not highly correlated with the intelligibility scores when a general function was used to map all intelligibility measures to intelligibility scores. Analysis further showed that the intelligibility prediction power could be significantly improved when multiple condition-dependent functions were used for mapping intelligibility measures to intelligibility scores.**

I. INTRODUCTION

Temporal envelope and temporal fine structure (TFS) have been recognized as two important acoustic cues for speech intelligibility [1]. The definition of temporal fine-structure varies across the literature [1-2]. Rosen defined fine structure as variations in the waveform within single periods of periodic sounds, with fluctuation rates ranging from 600 Hz to 10 kHz [2]. The most straightforward (mathematical) definition of TFS stems from the decomposition of a band-passed signal into its envelope and TFS components using the Hilbert transform [1]. The envelope captures the slowly varying modulations of amplitude in time, while the TFS component captures the rapid oscillations occurring at a rate close to the center frequency of the band. A number of recent studies have demonstrated that TFS cue contributes significantly to speech perception in noise, pitch perception, sound localization and tonal-language recognition [1, 3-5].

Recently studies showed that speech perception problems of the hearing-impaired listeners reflect their inabilities to use TFS information [6]. Vocoded speech has been used widely to simulate the listening performance of cochlear implants (CIs), the only medical treatment to restore partial hearing to a severely-to-profoundly deafened person [7]. In vocoder simulation, speech is processed in a manner similar to the CI speech processor, i.e., delivering the temporal envelope information and eliminating the TFS information, and presented to normal-hearing (NH) listeners for identification. A recent development of the CI technique is the combined electric-acoustic stimulation (EAS), which makes use of the residual acoustic hearing that many patients still have at low frequency. The EAS delivers the partial TFS cue at low frequency to the listeners, and its benefit in terms of better speech recognition in noise has been well documented in studies involving EAS patients [8].

Assessing the effect of TFS information for speech perception is important to guide the development of novel speech coding algorithms in prosthetic hearing devices, e.g., cochlear implants and hearing aids. However, considering the large algorithmic parametric space, and the large number of signal-to-noise ratio (SNR) levels needed to construct psychometric functions in noisy conditions, a large number of listening tests with (e.g., vocoded) speech are often needed to reach reliable conclusions. Alternatively, objective intelligibility prediction can be used to predict the intelligibility of speech containing different amount of TFS information.

While it is widely believed that the TFS cue has a notable impact on the performance of speech understanding, especially in noisy environments, it is still unknown how it would affect the performance of existing intelligibility indices in predicting the intelligibility of speech containing different amount of TFS information. The purpose of this study is to assess the performance of intelligibility indices in predicting the intelligibility of speech varying in the amount of TFS cue, i.e., little, partial and intact TFS cues in the tone-vocoded, EAS-vocoded and wideband speech, respectively. More specifically, the present study examines the effect of TFS on speech intelligibility modeling by using different functions to map the intelligibility measures to intelligibility scores. The intelligibility data of vocoded and wideband Mandarin speech was first collected from listening experiment, and was subsequently correlated with existing intelligibility indices to examine their performance in intelligibility prediction.

II. SPEECH INTELLIGIBILITY DATA

Nine (five male) NH listeners participated in the listening experiment. All subjects (aged from 23 to 42 yrs) were native speakers of Mandarin Chinese. The speech material consisted of Mandarin sentences taken from the Sound Express database [9]. All the sentences were produced by a female speaker, and recorded at a 22,050 Hz

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 General Research Fund (GRF), administered by the Hong Kong Research Grants council. This work was also supported Grant 31271056 from National Natural Science Foundation of China.

Fei Chen is with Division of Speech and Hearing Sciences, The University of Hong Kong, Hong Kong (e-mail: feichen1@hku.hk).

Tian Guan is with Graduate School at Shenzhen, Tsinghua University, Shenzhen, China.

Lena L. N. Wong is with Division of Speech and Hearing Sciences, The University of Hong Kong, Hong Kong.

sampling rate. Two types of maskers were used to corrupt the sentences, i.e. the continuous steady-state noise and two equal-level interfering female talkers. The fundamental frequencies of the target and two interfering talkers were 230, 232 and 235 Hz, respectively. The test sentences were processed by three signal processing (or TFS) conditions, preserving little, partial and intact TFS information, respectively.

The first processing condition (i.e., tone-vocoded) simulated the 8-channel electrical stimulation by using an 8-channel sinewave-excited vocoder. Signals were first processed through a pre-emphasis filter (2 kHz cutoff) with a 3 dB/octave roll-off and then band-passed into 8 frequency bands between 80 and 6 kHz (i.e. 80, 221, 426, 724, 1158, 1790, 2710, 4050, and 6 kHz) using sixth-order Butterworth filters. The envelope of the signal was extracted by full-wave rectification and low-pass (LP) filtering using a second-order 400 Hz Butterworth filter. Sinusoids were generated with amplitudes equal to the root-mean-square (RMS) energy of the envelopes (computed every 4 ms) and frequencies equal to the center frequencies of the bandpass filters. The sinusoids of each band were finally summed up and the level of the synthesized speech segment was adjusted to have the same RMS value as the original speech segment [9].

The second processing condition (i.e., EAS-vocoded) simulated the EAS stimulation. Signal was first LP filtered to 600 Hz using a sixth-order Butterworth filter. To simulate the effects of EAS for patients with residual hearing below 600 Hz, the LP stimulus was combined with the upper 5 channels of the 8-channel vocoder.

In the third processing condition (i.e., wideband), the corrupted Mandarin sentences were first processed through a pre-emphasis filter (2 kHz cutoff) with a 3 dB/octave roll-off, and then limited to the frequency range between 80 and 6 kHz using sixth-order Butterworth filters.

TABLE I lists the amount of TFS information and SNR levels for the above three signal processing conditions. The SNR levels were selected chosen to avoid ceiling/floor effects. The listening experiment was performed in a sound-proof room using a PC connected to a Tucker-Davis system 3. Stimuli were played to listeners monaurally through a Sennheiser HD 250 Linear II circumaural head-phone at a comfortable listening level. Prior to the test, each subject participated in a 10-minute training session to listen to a set of stimuli and familiarize them with the test procedure. During the testing session, the subjects were asked to write down all the words they heard. Each subject participated in a total of 36 (=10 tone-vocoded + 10 EAS-vocoded + 16 wideband) conditions. Twenty Mandarin sentences were used per condition, and none of the sentences were repeated across the conditions. The order of the test conditions was randomized across subjects. Subjects were given a 5-min break every 30 mins during the testing session. More detail on signal processing and test procedure is included in [9].

TABLE I. The amount of TFS information and SNR levels for the three TFS conditions.

TFS conditions	Amount of TFS information	SNR level (dB)
tone-vocoded	little	–4, 0, 4, 8, 12
EAS-vocoded	partial (at low frequency)	–4, –2, 0, 2, 4
wideband	intact	–14, –12, –10, –8, –6, –4, –2, 0

III. SPEECH INTELLIGIBILITY MEASURES

Present intelligibility indices employ primarily either temporal-envelope or spectral-envelope information to compute the index. For the temporal-envelope based measure, we examined the intelligibility prediction performance of the normalized covariance measure (NCM) [10], while for the spectral-envelope based measure, we investigated the coherence-based speech intelligibility index (CSII) measure [11].

The NCM index is similar to the speech-transmission index [12] in that it computes a weighted sum of transmission index (TI) values determined from the envelopes of the probe and response signals in each frequency band [10]. Unlike the traditional STI measure, however, which quantifies the change in modulation depth between the probe and response envelopes using the modulation transfer function, the NCM index is based on the covariance between the probe and response envelope signals computed in each band. The NCM index makes use of the envelopes extracted for the whole utterance to compute the TI value of each band. The TI values are subsequently converted to an apparent SNR and mapped to the NCM index taking values between 0 and 1.

The speech intelligibility index (SII) [13] is based on the principle that the intelligibility of speech depends on the proportion of spectral information that is audible to the listener and is computed by dividing the spectrum into 20 bands (contributing equally to intelligibility) and estimating the weighted average of the SNRs in each band. The modified coherence-based SII index (i.e., CSII) [11] uses the base form of the SII procedure, but with the SNR term replaced by the signal-to-distortion ratio, which is computed using the coherence function between the probe and response signals.

These two measures have been shown to yield high correlations with the intelligibility of vocoded and wideband speech [9, 14], and noise-masked speech processed by noise reduction algorithms [15]. More details regarding the definition and implementation of the NCM and CSII measures can be found in [10, 15].

IV. RESULTS

The Pearson's correlation coefficient (r) was used to assess the intelligibility prediction performance of intelligibility indices. The average intelligibility scores obtained by NH listeners were subjected to the correlation analysis with the corresponding values obtained by the CSII and NCM indices.

TABLE II shows the resulting correlation coefficients between sentence recognition scores and the CSII and NCM measures. It is seen that the CSII and NCM measures well predict the intelligibility of wideband or vocoded speech.

TABLE II. Correlation coefficients (r) between sentence recognition scores and intelligibility indices. Asterisk indicates that the difference in correlation coefficients is statistically significant (p<0.05) between strategies of one general mapping and multiple condition-dependent mapping.

TFS conditions		CSII		NCM	
		r	(a, b)	r	(a, b)
Separated	tone-vocoded	0.91	(289.5, –1.6)	0.92	(269.5, –57.7)
	EAS-vocoded	0.89	(168.6, 16.7)	0.89	(154.6, –14.4)
	wideband	0.90	(272.6, 29.8)	0.83	(146.9, 9.5)
Combined (wideband + vocoded)	one general mapping	0.70	(175.8, 26.2)	0.73	(137.3, 1.4)
	multiple condition-dependent mapping	0.91 *	(272.6, 29.8) (289.5, –1.6) (168.6, 16.7)	0.87 *	(146.9, 9.5) (269.5, –57.7) (154.6, –14.4)

The correlation coefficients range from r=0.83 to 0.92. However, when predicting the intelligibility scores of speech from all TFS conditions (i.e., wideband, tone-vocoded and EAS-vocoded), the prediction performance dramatically drops to correlation r=0.70 and 0.73 for the CSII and NCM measures, respectively. Figure 1 show the scatter plot of intelligibility scores against the CSII measures. A linear function was used for mapping the CSII values to intelligibility scores in each TFS condition, as:

$$y_{pre} = a \cdot x + b, \quad (1)$$

where y_{pre} and x are the predicted intelligibility score and intelligibility measure (e.g., CSII), respectively, and a and b are the fitting parameters. The solid, dotted and dashed lines (see TABLE II for values of fitting parameters a and b) in Fig. 1 linearly map the intelligibility measures to intelligibility scores in the wideband, tone-vocoded, and EAS-vocoded conditions, respectively. It is seen that, though performing well in predicting the intelligibility of wideband or vocoded speech separately, the CSII measures are not highly correlated with the intelligibility scores when a general function is used to map all intelligibility measures to intelligibility scores. Figure 2 (a) shows the scatter plot of intelligibility scores against the predicted scores when the CSII measures are mapped with a general function determined from the all 36 intelligibility data in Fig. 1.

Instead of using one general function to map all intelligibility measures to intelligibility scores, we tried to use multiple functions for the above mapping purpose. This is done by using the three linear functions (i.e., three fitting lines in Fig. 1) for mapping the intelligibility measures of wideband, tone-vocoded or EAS-vocoded sentences towards their intelligibility scores, separately. The rational for this is that the mapping (or fitting) functions differ substantially when applied to intelligibility data collected from different TFS conditions, as shown in Fig. 1.

TABLE II lists the resulting correlation coefficients between all intelligibility scores and the predicted scores when multiple condition-dependent functions are used for intelligibility prediction. The correlations are improved to 0.91 and 0.87 for the CSII and NCM measures, respectively. Statistical analysis was performed as per Steiger [16]. When compared to the standard normal curve rejection points of ± 1.96, the correlation coefficient computed with multiple condition-dependent mapping functions is found to be significantly (p<0.05) higher than those obtained with one general mapping function. Figure 2 (b) shows the scatter plot of intelligibility scores against the predicted scores when the CSII measures are mapped with multiple (i.e., three) mapping functions.

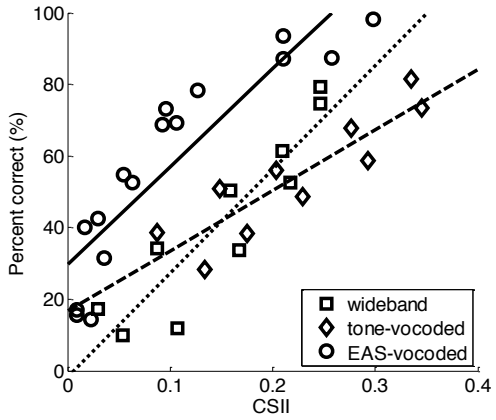

Figure 1. Scatter plot of intelligibility scores against the CSII measures. The solid, dotted and dashed lines linearly map the intelligibility measures to intelligibility scores for the three TFS conditions, respectively.

V. DISCUSSION AND CONCLUSION

Recent studies have found that, for predicting the intelligibility of cochlear-implant vocoded speech (both English and Mandarin) with little TFS cue, envelope information was sufficient to yield a high prediction [9, 14]. Such measures included the NCM measure, which primarily employed the temporal envelope cue for intelligibility prediction. Consistently, it is seen in TABLE II that the NCM measures well predicted the intelligibility of tone-vocoded (r=0.92) and EAS-vocoded (r=0.89) Mandarin. However, when considering the impact of TFS cue on speech intelligibility, it was found that the envelope

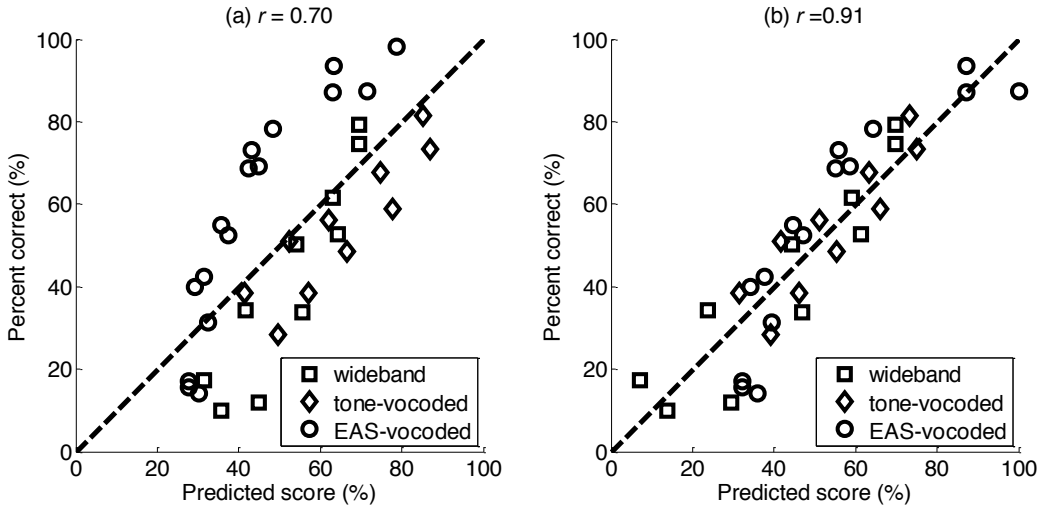

Figure 2. Scatter plots of intelligibility scores against the predicted scores when the CSII measures are mapped with (a) one general function, and (b) multiple condition-dependent functions.

information was not sufficient to well correlate with the intelligibility scores any more. Due to the contribution of TFS cue, the intelligibility of the wideband Mandarin speech was higher than that of the vocoded speech involving little or partial TFS cue. The envelope based (e.g., NCM) measure did not capture and quantify the TFS related distortion to account for its contribution for improved intelligibility score. Thus, it is not surprising that the resulting prediction performance of the NCM measure was degraded in the wideband plus vocoded conditions (i.e., r=0.73). The same effect of TFS also occurs to the CSII measure in predicting intelligibility of speech containing different amount of TFS information (i.e., r=0.70).

These findings suggest that it would be difficult to use one general mapping function to account for the effect of TFS on speech intelligibility modeling. In order words, in order to account for the effect of TFS on speech intelligibility modeling, multiple functions should be utilized to map intelligibility measures to intelligibility scores, with each characterizing the effect of specific TFS condition (e.g., tone-vocoded) on speech intelligibility.

In conclusion, the present study found that, though performing well in predicting the intelligibility of wideband or vocoded speech separately, present intelligibility indices (i.e., CSII and NCM) were not highly correlated with the intelligibility scores when a single function was used to map all intelligibility measures to intelligibility scores. The intelligibility prediction power could be significantly improved when multiple condition-dependent functions were used for mapping the intelligibility measures to intelligibility scores.

REFERENCES

[1] Z. Smith, B. Delgutte, and A. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, pp. 87–90, 2002.

[2] S. Rosen, "Temporal information in speech: acoustic, auditory and linguistic aspects," *Philos. Trans. R. Soc. London, Ser. B*, vol. 336, 367–373, 1992.

[3] F.G. Zeng, K.B. Nie, S. Liu, G. Stickney, E. Del Rio, Y.Y. Kong, and H. Chen, "On the dichotomy in auditory perception between temporal envelope and fine structure cues," *J. Acoust. Soc. Am.*, vol. 116, 1351–1354, 2006.

[4] Y.Y. Kong and F.G. Zeng, "Temporal and spectral cues in Mandarin tone recognition," *J. Acoust. Soc. Am.*, vol. 120, pp. 2830–2840, 2006.

[5] B.C. Moore, "The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people," *J. Assoc. Res. Otolaryngol.*, vol. 9, 399–406, 2008.

[6] C. Lorenzi, G. Gilbert, H. Carn, S. Garnier, and B.C. Moore, "Speech perception problems of the hearing impaired reflect inability to use temporal fine structure," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 103, 18866–18869, 2006.

[7] P.C. Loizou, "Introduction to cochlear implants," *IEEE Eng. Med. Biol. Mag.*, vol. 18, pp. 32–42, 1999.

[8] B. Gantz and C. Turner, "Combining acoustic and electric hearing," *Laryngoscope*, vol. 113, pp. 1726–1730, 2003.

[9] F. Chen and P.C. Loizou, "Predicting the intelligibility of vocoded and wideband Mandarin Chinese," *J. Acoust. Soc. Am.*, vol. 129, no. 5, pp. 3281–3290, 2011.

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

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

[12] 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.

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

[14] F. Chen and P.C. Loizou, "Predicting the intelligibility of vocoded speech," *Ear Hear.*, vol. 32, pp. 331–338, 2011.

[15] 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. Am.*, vol. 125, pp. 3387–3405, 2009.

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