Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 000 (2016) 1–9
www.elsevier.com/locate/specom

Modeling speech intelligibility with recovered envelope from temporal fine structure stimulus

Fei Chen[a,*], Yu Tsao[b], Ying-Hui Lai[b,c]

[a] *Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Xueyuan Road 1088#, Xili, Nanshan District, Shenzhen, China*
[b] *Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan*
[c] *Department of Electrical Engineering, Yuan Ze University, Chung Li, Taiwan*

Received 17 June 2015; received in revised form 3 December 2015; accepted 30 January 2016
Available online xxx

Abstract

Temporal envelope and fine structure are two prominent acoustic cues for speech perception. Most existing speech-transmission-index-based metrics make use of the temporal envelope information and discard the temporal fine structure (TFS) cue to predict speech intelligibility. Recent studies have shown that the TFS stimulus synthesized with multiband TFS waveforms contains rich intelligibility information, which is reflected as the recovered envelope from the TFS stimulus. The present study first assessed the performance of using the recovered envelope from the synthesized TFS stimulus to predict the intelligibility of noise-distorted and noise-suppressed speech. The TFS stimulus was synthesized and fed as an input into the conventional normalized covariance measure (NCM) module. The results showed that the recovered envelope from the TFS stimulus predicted the intelligibility as well as the original envelope extracted from the wideband speech signal did. In addition, an additive intelligibility model was designed to combine the envelope from wideband speech and the recovered envelope from the TFS stimulus to predict speech intelligibility. The prediction power was significantly improved when these two envelope waveforms were integrated. The present study suggests that the recovered envelope from the TFS stimulus may be alternative acoustic information for modeling speech intelligibility and improving the prediction power of the conventional NCM-based intelligibility index.

Keywords: Speech intelligibility; Temporal fine structure; Recovered envelope; Normalized covariance measure.

1. Introduction

A number of intelligibility indices have been developed to objectively model the intelligibility of the processed (e.g., by noise corruption or noise suppression) speech (e.g., Steeneken and Houtgast, 1980; Holube and Kollmeier, 1996; Goldsworthy and Greenberg, 2004; Kates and Arehart, 2005; Jørgensen and Dau, 2013; Chen et al., 2013; Mamun et al., 2015). Temporal envelope and fine structure have long been identified as two acoustic cues important for speech perception (e.g., Rosen, 1992; Smith et al., 2002; Zeng et al., 2005). The most straightforward mathematical definition of temporal envelope and fine structure stems from the decomposition of a band-passed signal into its envelope and fine structure components using the Hilbert transform (Smith et al., 2002). The temporal envelope carries slow-varying amplitude fluctuation information in time, whereas the temporal fine structure (TFS) component mostly captures the rapid oscillations occurring at a rate close to the center frequency of the band. The relative contributions of temporal envelope and fine structure for speech perception have been extensively assessed in a number of studies (Shannon et al., 1995; Smith et al., 2002; Zeng et al., 2005; Gilbert and Lorenzi, 2006; Lorenzi et al., 2006; Moore, 2008; Chen and Guan, 2013). For instance, it was found that the envelope waveforms extracted from up to four channels carry sufficient intelligibility information in a quiet environment (Shannon et al., 1995). In addition, many speech intelligibility indices were developed primarily based on envelope information, such as the speech-based speech transition index (STI) (Houtgast and Steeneken, 1980).

The STI metric originally used an artificial signal as a probe signal to measure the reduction of signal modulation

* Corresponding author. Tel.: +86 755 88015878.
E-mail address: fchen@sustc.edu.cn (F. Chen).

http://dx.doi.org/10.1016/j.specom.2016.01.006
0167-6393/

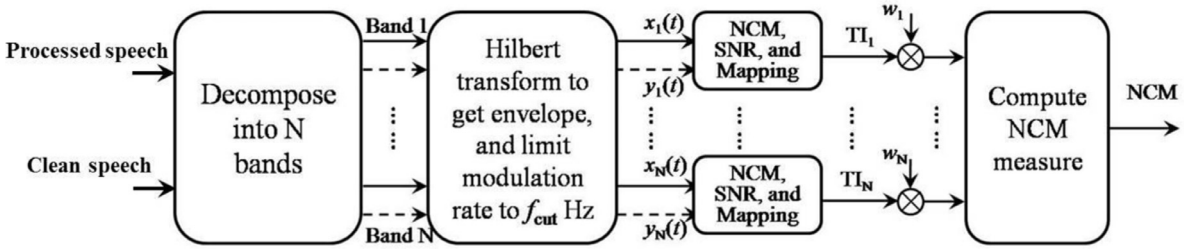

Fig. 1. Signal processing steps involved in computing the NCM measure.

from a number of frequency bands and a range of modulation frequencies (e.g., 0.6–12.5 Hz) that carry important information for speech intelligibility (Houtgast and Steeneken, 1971). Recently, many modifications have been proposed, for instance, to use speech signals as probe signals in computing the STI metric. Among them, one successful example is the speech-based normalized covariance measure (NCM) (Holube and Kollmeier, 1996; Goldsworthy and Greenberg, 2004). The computation of the NCM measure discards the fine structure from analysis bands (see more in Fig. 1 and Section 2.2) because earlier studies have demonstrated that the information contained in low-frequency (<16 Hz) envelope modulations is sufficient for speech perception (e.g., Drullman et al., 1994a, 1994b). In other words, the conventional NCM measure is envelope-centric and based on the primary role of envelope information to speech intelligibility. Many studies have shown the efficiencies of the NCM measure in predicting the intelligibility of speech in noise, in reverberation, or processed by vocoder (e.g., Goldsworthy and Greenberg, 2004; Chen and Loizou, 2011). Although the envelope-based NCM measures are able to account for the linear distortions introduced by filtering and additive noise, when speech is subjected to nonlinear processing (e.g., noise suppression), they fail to successfully predict speech intelligibility (e.g., van Buuren et al., 1999; Goldsworthy and Greenberg, 2004). For instance, some noise suppression algorithms (e.g., the spectral subtractive algorithm in Gustafsson et al., 2001) can introduce nonlinear distortions in the noise-suppressed signal and unduly increase the level of modulation in the temporal envelope that would be incorrectly interpreted as increased signal-to-noise ratio (SNR) by the envelope-based measure (e.g., Goldsworthy and Greenberg, 2004).

However, a number of recent studies have shown that listeners can recognize, with high accuracy, speech synthesized to contain only multiband TFS information. The TFS stimulus was synthesized by splitting a wideband speech signal into multiple bands, extracting the TFS waveform (e.g., via Hilbert transform) in each band, and summing root-mean-square (RMS) weighted TFS waveforms from all bands (see more in Section 2.3 on the process of TFS stimulus synthesis, and see the review by Moore, 2008). Smith et al. (2002) showed, for instance, that when speech was synthesized using the Hilbert-derived TFS waveforms from a smaller number of frequency bands, speech intelligibility was generally good. Studies also suggested that the recovered envelope from the TFS- or phase-based stimulus accounts for the intelligibility of the TFS- or phase-based stimulus (e.g., Gilbert and Lorenzi, 2006; Chen and Guan, 2013). Hence, given the importance of the TFS cue to speech perception, the first motivation of this study is to assess whether the recovered envelope from the TFS stimulus can be used to predict the intelligibility of noise-distorted and noise-suppressed speech and compare its performance with the conventional envelope-based intelligibility index (i.e., NCM) using envelope information extracted from the wideband speech signal. As mentioned earlier, the conventional NCM measure discards fine structure information in its computation. The fine structure waveforms will be summed to synthesize the TFS stimulus in this study. We hypothesize that the recovered envelope from the synthesized TFS stimulus would also well predict speech intelligibility; however, the prediction power may be influenced by several factors used in synthesizing the TFS stimulus, e.g., number of TFS channels. Studies have found that using a large number of channels in synthesizing the TFS stimulus yields a reduced amount of intelligibility information contained in the TFS stimulus (Smith et al., 2002; Lorenzi et al., 2006; Gilbert and Lorenzi, 2006). In addition, studies have also suggested the usage of a high modulation frequency (i.e., the low-pass cutoff frequency used to extract the temporal envelope waveform) for modeling the intelligibility of speech with diminished acoustic cues (e.g., Chen and Loizou, 2011). When the modulation rate was improved from 12.5 Hz to 100 Hz, Chen and Loizou found that the extracted temporal envelope information captured more intelligibility information, i.e., the correlation coefficient between the envelope-based NCM measures and subjective intelligibility scores was increased from 0.85 to 0.92 (Chen and Loizou, 2011). The present work will examine how these two factors (i.e., number of TFS channels and modulation rate) would influence the prediction performance of the recovered-envelope-based intelligibility index.

Because both envelope waveforms (i.e., from wideband speech and from synthesized TFS stimulus) carry important information for modeling speech intelligibility, the second aim of the present work is to further improve the intelligibility prediction power of the conventional NCM measure originally based on the wideband speech signal. The envelopes computed from the wideband speech and synthesized TFS stimulus will be integrated into a new intelligibility index to improve the performance of envelope-based speech

Please cite this article as: F. Chen et al., Modeling speech intelligibility with recovered envelope from temporal fine structure stimulus, Speech Communication (2016), http://dx.doi.org/10.1016/j.specom.2016.01.006

Table 1
The center frequencies and bandwidths of the N Butterworth filters used in this study.

N	Center frequencies (Hz)	Bandwidth (Hz)
2	711, 2261	822, 2277
4	454, 865, 1549, 2688	308, 513, 854, 1422
8	367, 521, 720, 977, 1309, 1736, 2288, 2999	134, 173, 224, 289, 373, 481, 621, 801
16	331, 398, 475, 562, 661, 773, 900, 1045, 1209, 1396, 1608, 1849, 2122, 2433, 2786, 3186	63, 71, 81, 92, 105, 119, 135, 153, 174, 198, 225, 256, 290, 330, 375, 426
20	324, 377, 435, 500, 571, 650, 737, 834, 941, 1060, 1191, 1336, 1498, 1676, 1873, 2092, 2334, 2602, 2898, 3227	49, 55, 61, 67, 74, 82, 91, 101, 112, 124, 138, 152, 169, 187, 207, 229, 254, 281, 311, 345

intelligibility prediction. This is based on the hypothesis that the performance of intelligibility prediction will be improved when more relevant acoustic information is integrated into a model to predict intelligibility.

2. Methodology

2.1. Temporal envelope and fine structure

The Hilbert transform is commonly used to separate a signal into two components of temporal envelope and fine structure. The Hilbert envelope and fine structure are derived from the following (complex-valued) analytic signal $s(t)$:

$$s(t) = s_{\mathrm{r}}(t) + i \cdot s_{\mathrm{i}}(t), \tag{1}$$

where $s_{\mathrm{r}}(t)$ is the real-valued signal, and $s_{\mathrm{i}}(t)$ representing the imaginary part is the Hilbert transform of $s_{\mathrm{r}}(t)$ (Baher, 2001). The signal in Eq. (1) can also be represented as:

$$s_{\mathrm{r}}(t) = A(t) \cdot \cos(\phi(t)), \tag{2}$$

where $A(t)$ denotes the Hilbert envelope as:

$$A(t) = \sqrt{s_{\mathrm{r}}^2(t) + s_{\mathrm{i}}^2(t)}, \tag{3}$$

and $\cos(\phi(t))$ denotes the Hilbert fine structure, and the instantaneous phase $\phi(t)$ is computed as:

$$\phi(t) = \mathrm{atan}\left(\frac{s_{\mathrm{i}}(t)}{s_{\mathrm{r}}(t)}\right). \tag{4}$$

Note that the signal $A(t)$ in Eq. (3) represents the slowly varying component of the signal (or amplitude fluctuation), whereas the signal $\cos(\phi(t))$ contains the more rapidly varying component of the signal (or frequency modulation).

2.2. The envelope-based NCM index

The normalized covariance measure, which is largely seen as an STI-based measure, is computed as follows (see Fig. 1). The speech signal is first decomposed into N bands spanning the signal bandwidth (300–3400 Hz in this study). Table 1 shows the center frequencies and bandwidths of the N filters used in this study. Note that the bandwidth covaries when the number of bands is increased. The speech decomposition is implemented with a series of fourth-order Butterworth filters, whose cutoff frequencies space the cochlear frequency map with equal steps and are computed according to the cochlear frequency–position function (Greenwood, 1990). The envelope of each band is computed using the Hilbert transform and then downsampled to $2f_{\mathrm{cut}}$ Hz, thereby limiting the envelope modulation rate to f_{cut} Hz. Note that the envelope modulation rate could be understood as the cutoff frequency when using low-pass filtering to extract the slowly varying envelope waveform. Let x_i(t) and y_i(t) be the downsampled envelope of the clean and processed signals, respectively, in the ith band. The normalized covariance in the ith band is computed as:

$$\rho_i = \frac{\sum_t (x_i(t) - \bar{x}_i)(y_i(t) - \bar{y}_i)}{\sqrt{\sum_t (x_i(t) - \bar{x}_i)^2}\sqrt{\sum_t (y_i(t) - \bar{y}_i)^2}}, \tag{5}$$

where $\bar{x}_i$ and $\bar{y}_i$ are the mean values of x_i(t) and y_i(t), respectively. The SNR in the ith band is defined as:

$$\mathrm{SNR}_i = 10\log_{10}\left(\frac{\rho_i^2}{1-\rho_i^2}\right), \tag{6}$$

and is subsequently limited to the range of [−15,15] dB. The transmission index (TI) in each band is computed by linearly mapping the SNR values between 0 and 1, as:

$$\mathrm{TI}_i = (\mathrm{SNR}_i + 15)/30. \tag{7}$$

Finally, the transmission indices are averaged across all frequency bands to yield the NCM index, as:

$$\mathrm{NCM}_{\mathrm{oE}} = \frac{\sum_{i=1}^{N} \mathrm{TI}_i \times w_i}{\sum_{i=1}^{N} w_i}, \tag{8}$$

where $\boldsymbol{W} = (w_1 \ldots w_i \ldots w_N)^{\mathrm{T}}$ denotes a band importance function applied to the N transmission index TI_i. Among several methods for choosing the band importance function $\boldsymbol{W}$ in Eq. (8), this study uses the most common ANSI articulation index (AI) weights (ANSI, 1997). Note that we use subscript "oE" in $\mathrm{NCM}_{\mathrm{oE}}$ in Eq. (8) to denote the fact that the computation of this NCM measure uses the *original envelope* (or "oE") waveform extracted from the input wideband speech signal. It can be seen that this NCM (or $\mathrm{NCM}_{\mathrm{oE}}$) measure uses the temporal envelope waveforms extracted using the Hilbert transform in different frequency bands, and in this process, the TFS waveforms stemming from the other part of the Hilbert transform are discarded.

2.3. Temporal fine structure stimulus

As noted earlier, the TFS stimulus is synthesized with the information discarded in the computation of the conventional

Please cite this article as: F. Chen et al., Modeling speech intelligibility with recovered envelope from temporal fine structure stimulus, Speech Communication (2016), http://dx.doi.org/10.1016/j.specom.2016.01.006

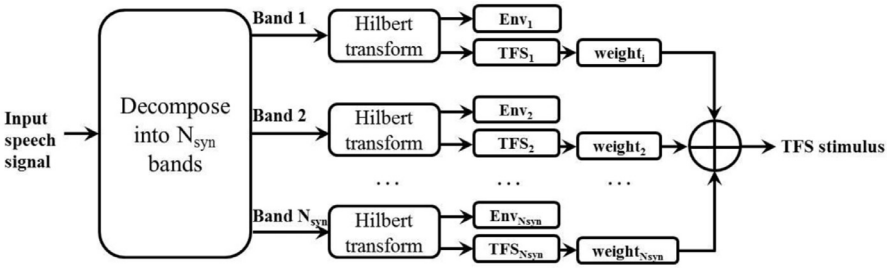

Fig. 2. Signal processing steps involved in synthesizing the TFS stimulus. "Env" and "TFS" denote the temporal envelope and fine structure, respectively.

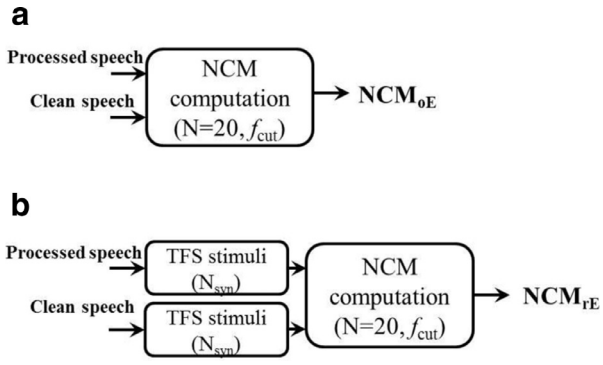

Fig. 3. Block diagrams to compute the (a) NCM_{oE} and (b) NCM_{rE} measures.

NCM measure in Fig. 1. More specifically, as shown in Fig. 2, the input speech signal is first split into N_{syn} frequency bands. The subscript "syn" indicates the number of channels used in synthesizing the TFS stimulus. The Hilbert transform is applied to the N_{syn} band-passed signals to obtain the TFS waveforms. Unlike the steps in computing the NCM measure, the envelope components are discarded, whereas the N_{syn}-channel TFS components are weighted to have the same RMS value as the band-passed signal, summed up, and finally adjusted to the RMS level of the original input speech signal. Note that both the processed and clean speech signals are used to generate the TFS stimuli. These two TFS stimuli are then used as input signals to compute the transmission index and NCM_{rE} measure, as shown in Fig. 3(b). We use subscript "rE" in NCM_{rE} to denote the fact that its computation uses the *recovered envelope* from the synthesized TFS stimulus in Fig. 2.

Fig. 3 compares the implementations of the NCM_{oE} and NCM_{rE} measures. Note that when the number of synthesis channels (N_{syn}) in Fig. 2 is identical to the number of channels (N) in Fig. 1, the TFS stimulus in Fig. 2 is synthesized exactly with the information (i.e., multiband TFS waveforms) discarded in the computation of the NCM_{oE} measure in Fig. 1.

2.4. Additive intelligibility model to integrate two envelopes

Two types of envelope waveforms are noted in this study, i.e., one from the original wideband speech signal (i.e., "oE") and the other from the synthesized TFS stimulus (i.e., "rE"). We expect that the combination of these two envelopes ought to improve the prediction of speech intelligibility. An additive intelligibility model is designed in the present study, which assumes that the contribution of the original and recovered envelopes is additive. The transmission index for band i is constructed as follows:

$$\mathrm{TI}_{\mathrm{oE+rE}_i} = a \cdot \mathrm{TI}_{\mathrm{rE}_i} + (1-a) \cdot \mathrm{TI}_{\mathrm{oE}_i}, \tag{9}$$

where $\mathrm{TI}_{\mathrm{oE}_i}$ and $\mathrm{TI}_{\mathrm{rE}_i}$ denote the transmission indices computed from the original and recovered envelopes, respectively, extracted in the ith band, and a is a weight taking values between 0 and 1. Accordingly, the integrated intelligibility index is denoted as $\mathrm{NCM}_{\mathrm{oE+rE}}$, where the subscript "oE + rE" represents the usage of two types of envelope waveforms. The model in Eq. (9) could provide valuable insights regarding the contribution of the original envelope and recovered envelope to intelligibility modeling. By varying the weight a in Eq. (9), we can assess the individual contribution of the original envelope and recovered envelope. A small value of weight a in Eq. (9) indicates a smaller contribution by the recovered envelope (or a larger contribution of the original envelope) in the additive intelligibility model, and vice versa.

3. Speech intelligibility data

The speech intelligibility data were taken from the intelligibility evaluation of noise-corrupted speech processed through eight different noise suppression algorithms by a total of 40 NH listeners (Hu and Loizou, 2007). IEEE sentences (IEEE, 1969) were used as test material, and all sentences were produced by a male talker and downsampled to 8 kHz. The masker signals were taken from the AURORA database (Hirsch and Pearce, 2000), involving real-world recordings from four different places: babble, car, street, and train. The maskers additively corrupted the speech signals at 5 and 0 dB SNR levels. The processed speech sentence files, along with the noisy speech files, were presented monaurally to the listeners in a double-walled soundproof booth via Sennheiser HD 250 Linear II circumaural headphones at comfortable listening levels. Twenty IEEE sentences were used for each condition, and none of the sentences were repeated. The intelligibility scores were obtained from NH listeners in a total of 72 conditions (=4 maskers × 2 SNR levels × 8 algorithms + 4 maskers × 2 noisy references), including 8 noisy and 64 noise-suppressed conditions, or 36 noisy/noise-suppressed conditions at each SNR level. The percentage intelligibility score

Please cite this article as: F. Chen et al., Modeling speech intelligibility with recovered envelope from temporal fine structure stimulus, Speech Communication (2016), http://dx.doi.org/10.1016/j.specom.2016.01.006

Table 2
Correlation coefficients (r) obtained with the NCM_{oE} and NCM_{rE} measures for predicting sentence intelligibility scores as a function of modulation rate (f_{cut}) and the number of channels N_{syn} in synthesizing the TFS stimulus. The number of channels $N=20$ was used to compute the NCM_{oE} and NCM_{rE} measures. An asterisk (∗) indicates that the correlation difference between NCM_{oE} and NCM_{rE} is significant ($p<0.05$).

N_{syn}	NCM_{rE}			NCM_{oE}		
	$f_{cut}=12.5$ Hz	25 Hz	50 Hz	$f_{cut}=12.5$ Hz	25 Hz	50 Hz
2	0.82	0.82	0.83			
4	0.80	0.82	0.83			
8	0.80	0.82	0.85	0.79	0.80	0.78
16	0.75	0.80	0.83			
20	0.60*	0.71	0.81			

Table 3
Correlation coefficients (r) obtained with the NCM_{oE}, NCM_{rE}, and NCM_{oE+rE} measures for predicting sentence intelligibility scores. The number of synthesis channels N_{syn} was 20. The number of channels $N=20$ and modulation rate $f_{cut}=50$ Hz were used to compute the NCM_{oE} and NCM_{rE} measures. Note that "36" includes 32 noise-suppressed conditions plus 4 noise-corrupted conditions. Asterisk indicates that the correlation coefficients from the NCM_{oE+rE} measure are significantly ($p<0.05$) larger than the two correlation coefficients computed with the NCM_{oE} and NCM_{rE} measures.

Conditions	NCM_{oE}	NCM_{rE}	NCM_{oE+rE} (a)
All (72)	0.78	0.81	0.88* (0.2)
Noisy (8)	0.89	0.90	0.92 (0.2)
Noise-suppressed (64)	0.83	0.81	0.91* (0.3)
0 dB SNR (36)	0.64	0.71	0.78* (0.2)
5 dB SNR (36)	0.73	0.73	0.83* (0.3)

for each condition was calculated by dividing the number of words correctly identified by the total number of words in a particular testing condition. More details about the noise suppression algorithms and the procedure used to collect the intelligibility scores can be found in Hu and Loizou (2007) and Loizou (2007).

4. Results

The average sentence intelligibility scores obtained by NH listeners in Section 3 were subjected to correlation analysis with the corresponding values obtained by the NCM measure (i.e., NCM_{oE}, NCM_{rE}, or NCM_{oE+rE}). More specifically, correlation analysis was performed between the mean (across all 40 subjects) sentence intelligibility scores obtained in each of the 72 testing conditions and the corresponding mean (computed across the 20 sentences in each condition) intelligibility index values obtained in each condition. Pearson's correlation coefficient (r) was used to assess the performance of the intelligibility measures to predict intelligibility scores.

4.1. Intelligibility prediction with recovered-envelope-based NCM measure

Table 2 shows the correlation results of the speech intelligibility modeling by the NCM_{oE} and NCM_{rE} measures. The diagrams to compute the NCM_{oE} and NCM_{rE} values are shown in Fig. 3(a) and (b). Note that the NCM_{rE} measure used the TFS stimulus synthesized with N_{syn} channels (see Fig. 2); in computing the NCM_{rE} and NCM_{oE} measures, we used the number of channel $N=20$ and selected the modulation rate from 12.5 to 50 Hz (see Fig. 1). For the NCM_{rE} measure, a positive contribution of high modulation rate (particularly at a large number of synthesis channels N_{syn}) is seen in Table 2. When the number of synthesis channels N_{syn} is set to 20, the correlation coefficient is increased from $r=0.60$ at $f_{cut}=12.5$ Hz to $r=0.81$ at $f_{cut}=50$ Hz. This is not surprising, because the recovered envelope from the TFS stimulus is commonly believed to account for the intelligibility of the TFS stimulus, and the envelope waveform recovered with a high modulation rate may carry more intelligibility information. In addition, the negative effect of a large N_{syn} value (i.e., more channels in synthesizing the TFS stimulus) is observed from Table 2, i.e., correlation coefficient $r=0.82$ at $N=2$ to $r=0.60$ at $N=20$ and $f_{cut}=12.5$ Hz. This finding is consistent with previous speech perception results regarding the effect of the number of synthesis channels (N_{syn}) on the intelligibility of the TFS stimulus (e.g., Smith et al., 2002; Gilbert and Lorenzi, 2006; Chen and Guan, 2013), i.e., intelligibility deterioration from increasing the number of bands for TFS-based speech synthesis.

A paired comparison of correlation coefficients computed with the NCM_{oE} and NCM_{rE} measures at $N_{syn}=N=20$ and modulation rate $f_{cut}=12.5$ Hz shows that the correlation coefficient of the NCM_{oE} measure is significantly ($p<0.05$) larger than that of the NCM_{rE} measure (Steiger, 1980), i.e., $r=0.79$ vs. 0.60, which indicates that the envelope waveform extracted from wideband speech contains more intelligibility information correlated with speech intelligibility. However, at high modulation rate $f_{cut}=50$ Hz, the two correlation coefficients do not show a significant ($p>0.05$) difference, i.e., $r=0.78$ vs. 0.81.

To further compare the prediction powers of the NCM_{oE} and NCM_{rE} measures, we split the 72 conditions according to their signal processing approaches. As shown in Table 3, we computed the correlation coefficients from a subset of all 72 conditions. More specifically, we computed the correlation coefficients from 8 noisy conditions, 64 noise-suppressed conditions, 36 SNR = 0 dB conditions and 36 SNR = 5 dB conditions. The computation of the NCM_{oE} and NCM_{rE} measures used $N_{syn}=N=20$ and $f_{cut}=50$ Hz. It can be seen in Table 3 that all differences between the correlation coefficients computed with the NCM_{oE} and NCM_{rE} measures at the same sub-dataset are not significant ($p>0.05$).

4.2. Intelligibility prediction with combined-envelope-based NCM measure

This study also assessed the intelligibility prediction performance by using the additive model designed in Eq. (9), and the results are shown in Fig. 4. Each panel in Fig. 4 corresponds to one modulation rate used to compute the NCM_{oE} and NCM_{rE} measures, i.e., 12.5, 25, and 50 Hz in panels (a), (b), and (c), respectively. In each panel, the leftmost (at $a=0$)

Please cite this article as: F. Chen et al., Modeling speech intelligibility with recovered envelope from temporal fine structure stimulus, Speech Communication (2016), http://dx.doi.org/10.1016/j.specom.2016.01.006

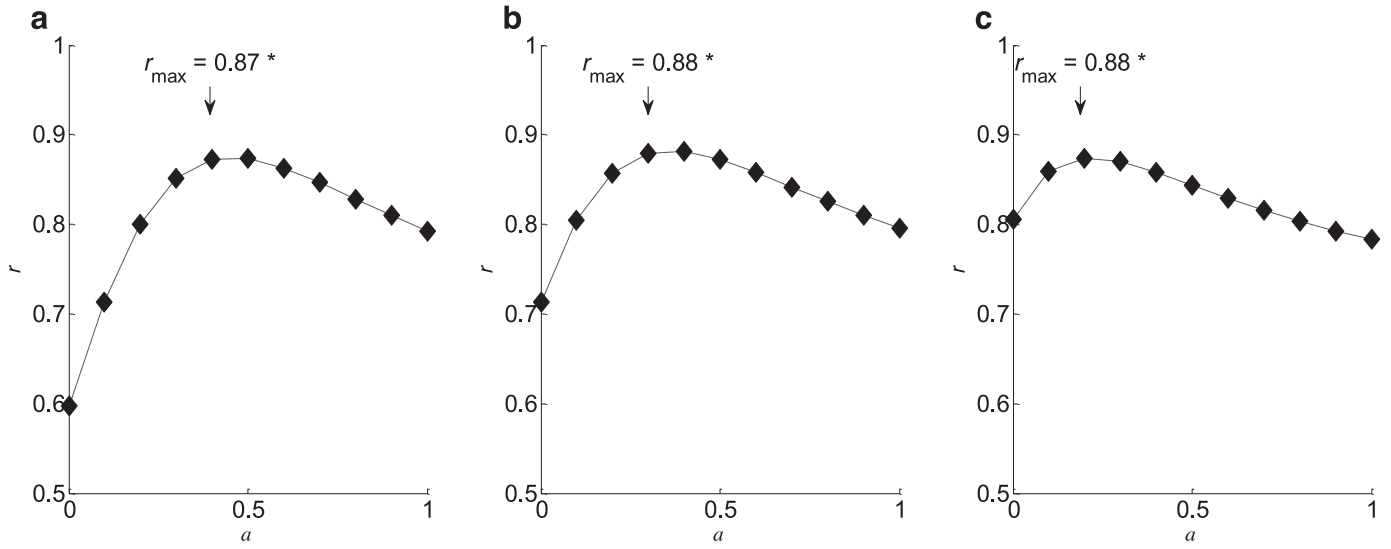

Fig. 4. The correlation coefficients (r) between the predicted NCM_{oE+rE} values and sentence recognition scores as a function of weight a in the additive intelligibility model and with modulation rate f_{cut} (a) 12.5, (b) 25, and (c) 50 Hz. The number of synthesis channels N_{syn} was 20. The number of channels $N=20$ was used to compute the NCM_{oE} and NCM_{rE} measures. Asterisks denote that the correlation coefficient r_{max} is significantly larger than those obtained from NCM_{oE} at $a=1$ and NCM_{rE} at $a=0$.

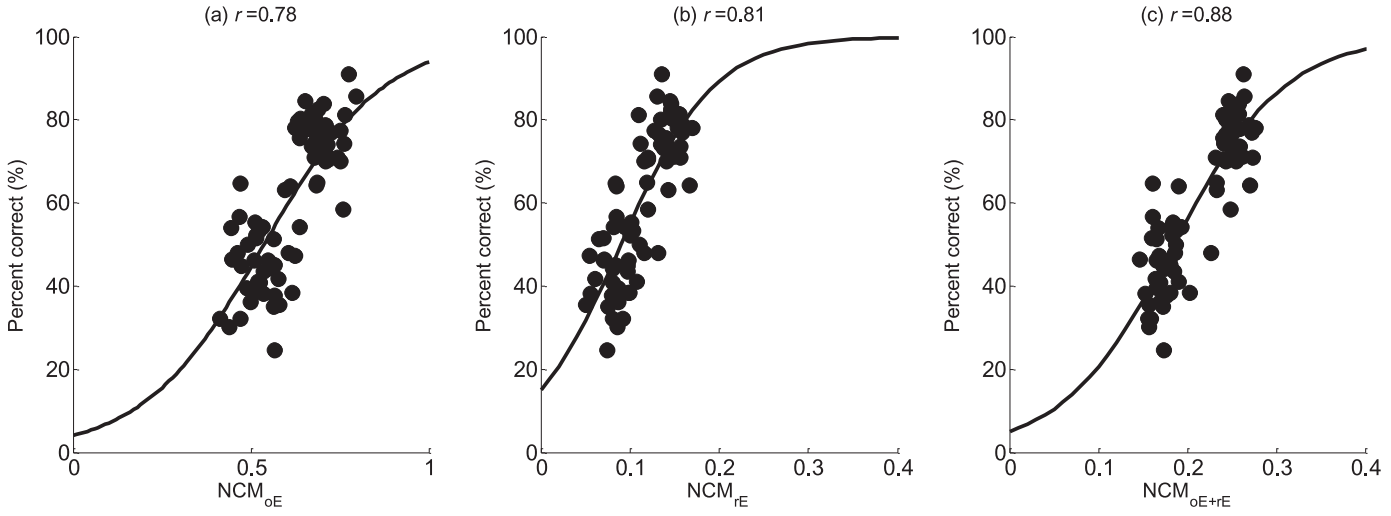

Fig. 5. Scatterplots of the predicted (a) NCM_{oE}, (b) NCM_{rE}, and (c) NCM_{oE+rE} values against sentence recognition scores. The number of synthesis channels N_{syn} was 20. The number of channels $N=20$ and modulation rate $f_{cut}=50$ Hz were used to compute the NCM_{oE} and NCM_{rE} measures.

and rightmost (at $a=1$) correlation coefficients are obtained with the NCM_{rE} and NCM_{oE} measures, respectively. It is seen that when using the additive model to integrate both envelopes from the wideband signal and synthesized TFS stimulus, the prediction correlation coefficient could be further improved to $r=0.87$ or 0.88, which is significantly ($p<0.05$) larger than the two correlation coefficients computed with the NCM_{oE} (i.e., $a=1$) and NCM_{rE} (i.e., $a=0$) measures. However, the value of the weight a yielding the maximal prediction correlation differs slightly in Fig. 4, i.e., $a=0.4$, 0.3, and 0.2 in panels (a), (b), and (c), respectively. Fig. 5 shows the scatterplots of the predicted NCM_{oE}, NCM_{rE}, and NCM_{oE+rE} values against sentence recognition scores.

Table 3 also shows the prediction correlation coefficients of the NCM_{oE+rE} measure when computed with various subdatasets. It is seen that, except for the analysis with 8 noisy conditions, the correlation coefficients from the NCM_{oE+rE} measure are significantly ($p<0.05$) larger than the two correlation coefficients computed with the NCM_{oE} and NCM_{rE} measures. In addition, the original envelope plays a larger contribution than the recovered envelope for the improved prediction correlation, as manifested by the small values of weight a shown in Table 3.

4.3. Cross-validation for the NCM_{oE+rE}-based intelligibility prediction

A cross-validation approach was used to assess the robustness of the proposed NCM_{rE} and NCM_{oE+rE} measures in modeling speech ineligibility. More precisely, the full dataset

Table 4
The correlation coefficients (r) obtained by the modified NCM measure based on various training–testing partitions of the dataset. The number of synthesis channels (N_{syn}) was 20. The number of channels $N=20$ and modulation rate $f_{cut}=50$ Hz were used to compute the NCM_{oE} and NCM_{rE} measures. An asterisk (∗) indicates that the correlation difference between NCM_{oE} (or NCM_{rE}) and NCM_{oE+rE} is significant ($p<0.05$).

Training–testing dataset partition	NCM_{oE}	NCM_{rE}	NCM_{oE+rE} (a)
50–50%	0.74*	0.79	0.85 (0.4)
33–67%	0.83	0.79*	0.90 (0.2)
25–75%	0.78*	0.78*	0.88 (0.3)
20–80%	0.80*	0.79*	0.90 (0.3)

(i.e., 72 conditions) was divided into a training sub-dataset, which was used to obtain the weight a for the two envelopes, and a testing sub-dataset that was used to assess the performance of the proposed NCM_{rE} and NCM_{oE+rE} measures. The partitions were made as follows. The complete set of conditions was first ordered according to their intelligibility scores. The training dataset was constructed by selecting one out of every two conditions, leading to a 50–50% partition of the training–testing datasets. Three additional training–testing dataset partitions were also implemented (including 33–67%, 25–75%, and 20–80%) by selecting one out of every 3, 4, and 5 conditions, respectively, from the complete dataset.

Table 4 shows several findings that are consistent with previous results. First, the NCM_{rE} measure performs equally well as the NCM_{oE} measure in predicting intelligibility scores at all training–testing conditions. The correlation coefficients obtained from the NCM_{oE} and NCM_{rE} measures are close, and their differences are insignificant ($p>0.05$). Second, when using the additive intelligibility model in Eq. (9) to predict intelligibility scores, the correlation coefficients are significantly ($p<0.05$) higher than those computed with the original envelope or recovered envelope. Third, the relative contribution from the original envelope is larger than that from the recovered envelope, as shown by the small values of weight a (i.e., 0.2–0.4) in Table 4.

5. Discussion and conclusions

Amplitude and phase are two primarily acoustic properties contained in speech signals. A number of studies have focused on making use of amplitude fluctuation information (e.g., temporal envelope) in speech perception; hence, many envelope-based speech intelligibility indices were developed and found to be effective in modeling speech intelligibility. Nevertheless, many recent studies have suggested that envelope information alone is not sufficient to provide better speech intelligibility in adverse conditions, e.g., by noise suppression processing. However, the importance of phase information has attracted increasing attention in speech recognition and intelligibility modeling. For instance, the TFS waveform, which contains phase variation information, was suggested to be important for recognizing speech in noise (Zeng et al., 2005), understanding tonal language (Nie et al., 2005), etc. New intelligibility indices based on TFS cues were also developed (e.g., Chen et al., 2013).

Although the mechanism of TFS- or phase-based speech recognition is not yet well understood, many studies have asserted that it may be largely attributed to the recovered envelope from the TFS stimulus (e.g., Zeng et al., 2004; Gilbert and Lorenzi, 2006). The recovered envelope can be simply extracted by the conventional signal processing involving band-passing, waveform rectification, and low-pass filtering, which simulates the signal transmission pathway in the periphery auditory system. As shown in Fig. 1, only temporal envelope information is used in the computation of the NCM measure, and the fine structure information complementary to the envelope information was discarded. The present study first assessed the performance using the discarded TFS information to predict speech intelligibility. The results showed that the two envelope waveforms (i.e., original and recovered) contained almost the same amount of intelligibility prediction information, as shown in Table 2. In addition, this result holds for all examined sub-datasets, as shown in Table 3. Nevertheless, it was found that the intelligibility information contained in the recovered envelope shows a different pattern compared with that in the original envelope. As observed in Table 2, the number of synthesis channels N_{syn} has a notable effect on the intelligibility predication performance. Synthesizing the TFS stimulus with more channels deteriorates the performance of intelligibility prediction, which is consistent with findings from speech perception experiments (e.g., Lorenzi et al., 2006; Gilbert and Lorenzi, 2006). Similarly, the usage of a high modulation rate leads to a better intelligibility prediction performance at a large number of channels (e.g., $N=20$). In contrast, the performance of the NCM_{oE}-based intelligibility prediction is not significantly influenced by the modulation rate (Ma et al., 2009).

Because the recovered envelope from the synthesized TFS stimulus also carries much information for intelligibility prediction, this study further used an additive model to integrate these two envelope waveforms into a composite intelligibility index, i.e., NCM_{oE+rE}. First, the results showed that the additive model could significantly ($p<0.05$) improve the prediction power of the conventional envelope-based intelligibility index, i.e., NCM_{oE}. The prediction correlation coefficient was improved to $r=0.88$, which is significantly ($p<0.05$) larger than the value computed with the NCM_{oE} or NCM_{rE} measures. This suggests the advantage of integrating two envelopes for intelligibility modeling. In addition, this result of improved prediction correlation indicates that the recovered envelope contains information complementary to the original envelope, at least for speech intelligibility modeling. The inclusion of this excess information from the recovered envelope accounts for the improved performance of intelligibility modeling. Second, the additive intelligibility model shows that the importance of the original envelope is larger than that of the recovered envelope in modeling speech intelligibility in this study, because the weight for the original envelope, i.e.,

Please cite this article as: F. Chen et al., Modeling speech intelligibility with recovered envelope from temporal fine structure stimulus, Speech Communication (2016), http://dx.doi.org/10.1016/j.specom.2016.01.006

(1–a), is larger than that for the recovered envelope, i.e., a, as shown in Fig. 1.

The noise-suppressed speech contains nonlinear distortion introduced by noise suppression algorithms. Many studies have attempted to improve the power of intelligibility prediction for the noise-suppressed speech (e.g., Ma et al., 2009; Chen and Loizou, 2012; Chen et al., 2013). For instance, Ma et al. (2009) used the same intelligibility dataset (i.e., the 72 conditions in this study) to evaluate the predictive power of the NCM measure. A signal-specific band-importance function was found to play critical roles in improving the performance of the examined NCM measure to predict the intelligibility of speech processed by noise suppression algorithms under noisy listening conditions (Ma et al., 2009). Chen and Loizou assessed the intelligibility prediction performance of the NCM measure by using selected (e.g., vowel–consonant transition) segments and obtained significantly improved intelligibility prediction performance. However, these approaches require additional (and complicated) signal processing to extract useful information to improve intelligibility prediction power. Some processing has a high computational load or is difficult to implement in real scenarios. For instance, Ma et al. (2009) computed the signal-dependent band-importance function, and Chen and Loizou (2012) required a priori information on the clean signal for segmenting speech signals into multiple regions. In contrast, the present work provides an alternative and simple method by using the fine structure information that is discarded in the computation of the NCM measure. The computation of the NCM_{rE} measure does not require additional knowledge of speech or noise signals and can be viewed as a byproduct of computing the conventional NCM_{oE} measure. However, this simple processing can significantly improve the prediction power of the NCM_{oE} measure, as shown in Fig. 2.

In conclusion, this study presented a detailed analysis of an intelligibility index that is computed with the recovered envelope from the fine structure stimulus. The computation of the NCM_{rE} measure is rather simple, i.e., using the synthesized TFS stimulus (discarded in the computation of the conventional NCM measure) as an input signal. Two analyses (i.e., its independent performance and assistive power to the conventional NCM measure) were carried out to investigate its performance for modeling speech intelligibility. The following conclusions can be drawn from the present study:

(1) A high correlation (up to $r = 0.81$) can be obtained with the modified NCM_{rE} measure by using the recovered envelope from the TFS stimulus. This correlation value is at the same level as that computed with the conventional NCM measure in all test conditions (see Table 3), suggesting that the TFS stimulus contains much important information for modeling speech intelligibility.
(2) Consistent with earlier findings on the effect of the TFS stimulus on speech recognition, the TFS stimulus synthesized with more channels contains less information for intelligibility prediction. However, this declined intelligibility prediction performance can be compensated by using a high modulation rate to recover the envelope waveform from the TFS stimulus.
(3) Higher correlations can be achieved when combining the two envelopes (i.e., original envelope from the wideband signal and recovered envelope from the TFS stimulus) to predict speech intelligibility. This may be attributed to the fact that the two indices (i.e., NCM_{oE} and NCM_{rE}) provide complementary information about speech. The original envelope captures intelligibility information that signifies the slow-varying amplitude modulation, which is lacking in the TFS stimulus. In contrast, the recovered envelope captures information related to modulating frequency (or phase) variation, which is discarded in the computation of the NCM measure.
(4) The original envelope from the wideband speech signal has a relatively larger contribution than the recovered envelope from the synthesized TFS stimulus in the additive intelligibility model [i.e., Eq. (9)] to model speech intelligibility.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant no. 61571213). This work was also supported by the Ministry of Science and Technology of Taiwan under Project MOST 104-2221-E-001-026-MY2. The authors thank the Associate Editor, Dr. Yannis Stylianou, and two anonymous reviewers for their constructive and helpful comments.

References

ANSI, 1997. Methods for Calculation of the Speech Intelligibility Index, S3.5-1997. American National Standards Institute, New York.
Baher, H., 2001. Analog & Digital Signal Processing, 2nd ed. Wiley, Chichester, NY.
Chen, F., Guan, T., 2013. Effect of temporal modulation rate on the intelligibility of phase-based speech. J. Acoust. Soc. Am. 134 EL 520–EL526.
Chen, F., Wong, L.L.N., Hu, Y., 2013. A Hilbert-fine-structure-derived physical metric for predicting the intelligibility of noise-distorted and noise-suppressed speech. Speech Commun. 55, 1011–1020.
Chen, F., Loizou, P., 2012. Contributions of cochlea-scaled entropy and consonant-vowel boundaries to prediction of speech intelligibility in noise. J. Acoust. Soc. Am. 131, 4104–4113.
Chen, F., Loizou, P., 2011. Predicting the intelligibility of vocoded speech. Ear Hear. 32, 331–338.
Drullman, R., Festen, J., Plomp, R., 1994a. Effect of temporal envelope smearing on speech reception. J. Acoust. Soc. Am. 95, 1053–1064.
Drullman, R., Festen, J., Plomp, R., 1994b. Effect of reducing slow temporal modulations on speech reception. J. Acoust. Soc. Am. 95, 2670–2680.
Gilbert, G., Lorenzi, C., 2006. The ability of listeners to use recovered envelope cues from speech fine structure. J. Acoust. Soc. Am. 119, 2438–2444.
Goldsworthy, R., Greenberg, J., 2004. Analysis of speech-based speech transmission index methods with implications for nonlinear operations. J. Acoust. Soc. Am. 116, 3679–3689.
Greenwood, D.D., 1990. A cochlear frequency-position function for several species – 29 years later. J. Acoust. Soc. Am. 87, 2592–2605.
Gustafsson, H., Nordholm, S., Claesson, I., 2001. Spectral subtraction using reduced delay convolution and adaptive averaging. IEEE Trans. Speech Audio Process 9, 799–807.

Hirsch, H., Pearce, D., 2000. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proceedings of ISCA Tutorial Research Workshop, ASR2000. Paris, France.

Holube, I., Kollmeier, K., 1996. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. J. Acoust. Soc. Am. 100, 1703–1715.

Houtgast, T., Steeneken, H.J.M., 1971. Evaluation of speech transmission channels by using artificial signals. Acustica 25, 355–367.

Hu, Y., Loizou, P., 2007. A comparative intelligibility study of single-microphone noise reduction algorithms. J. Acoust. Soc. Am. 122, 1777–1786.

IEEE, 1969. IEEE recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust. 17, 225–246.

Jørgensen, S., Dau, T., 2013. Modelling speech intelligibility in adverse conditions. Adv. Exp. Med. Biol. 787, 343–351.

Kates, J., Arehart, K., 2005. Coherence the speech intelligibility index. J. Acoust. Soc. Am. 117, 2224–2237.

Loizou, P., 2007. Speech Enhancement: Theory Practice. CRC, Boca Raton, FL.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., Moore, B.C., 2006. Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. Proc. Natl. Acad. Sci. USA 103, 18866–18869.

Ma, J.F., Hu, Y., Loizou, P., 2009. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J. Acoust. Soc. Am. 125, 3387–3405.

Mamun, N., Jassim, W.A., Zilany, M., 2015. Prediction of speech intelligibility using a neurogram orthogonal polynomial measure (NOPM). IEEE Trans. Audio Speech Lang. Process. 23, 760–773.

Moore, B.C., 2008. 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. 9, 399–406.

Nie, K., Stickney, G., Zeng, F.G., 2005. Encoding frequency modulation to improve cochlear implant performance in noise. IEEE Trans. Biomed. Eng. 52, 64–73.

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

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270, 303–304.

Smith, Z.M., Delgutte, B., Oxenham, A.J., 2002. Chimaeric sounds reveal dichotomies in auditory perception. Nature 416, 87–90.

Steeneken, H., Houtgast, T., 1980. A physical method for measuring speech transmission quality. J. Acoust. Soc. Am. 67, 318–326.

Steiger, J.H., 1980. Tests for comparing elements of a correlation matrix. Psychol. Bull. 87, 245–251.

van Buuren, R., Festen, J., Houtgast, T., 1999. Compression and expansion of the temporal envelope: evaluation of speech intelligibility and sound quality. J. Acoust. Soc. Am. 105, 2903–2913.

Zeng, F.G., Nie, K., Stickney, G.S., Kong, Y.Y., Vongphoe, M., Bhargave, A., Wei, C., Cao, K., 2005. Speech recognition with amplitude and frequency modulations. Proc. Natl. Acad. Sci. USA 102, 2293–2298.

Zeng, F.G., Nie, K., Liu, S., Stickney, G., Del Rio, E., Kong, Y.Y., Chen, H., 2004. On the dichotomy in auditory perception between temporal envelope and fine structure cues. J. Acoust. Soc. Am. 116, 1351–1354.

Please cite this article as: F. Chen et al., Modeling speech intelligibility with recovered envelope from temporal fine structure stimulus, Speech Communication (2016), http://dx.doi.org/10.1016/j.specom.2016.01.006