Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 81 (2016) 120–128

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 17 February 2016

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/

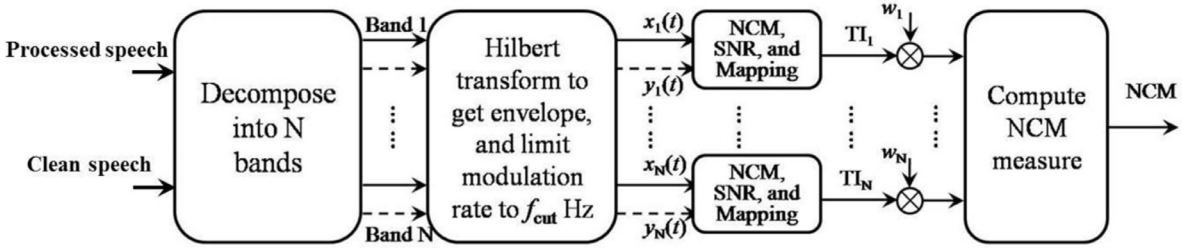

Fig. 1. Signal processing steps involved in computing the NCM measure.

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