Hearing Research 309 (2014) 75–83

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Research paper

The effects of noise vocoding on speech quality perception

Melinda C. Anderson*, Kathryn H. Arehart, James M. Kates

University of Colorado, Speech Language, Hearing Sciences, 2501 Kittredge Loop Road, 409 UCB, Boulder, CO 80309, USA

ARTICLE INFO

Article history:
Received 23 July 2012
Received in revised form
22 November 2013
Accepted 25 November 2013
Available online 11 December 2013

ABSTRACT

Speech perception depends on access to spectral and temporal acoustic cues. Temporal cues include slowly varying amplitude changes (i.e. temporal envelope, TE) and quickly varying amplitude changes associated with the center frequency of the auditory filter (i.e. temporal fine structure, TFS). This study quantifies the effects of TFS randomization through noise vocoding on the perception of speech quality by parametrically varying the amount of original TFS available above 1500 Hz. The two research aims were: 1) to establish the role of TFS in quality perception, and 2) to determine if the role of TFS in quality perception differs between subjects with normal hearing and subjects with sensorineural hearing loss. Ratings were obtained from 20 subjects (10 with normal hearing and 10 with hearing loss) using an 11-point quality scale. Stimuli were processed in three different ways: 1) A 32-channel noise-excited vocoder with random envelope fluctuations in the noise carrier, 2) a 32-channel noise-excited vocoder with the noise-carrier envelope smoothed, and 3) removal of high-frequency bands. Stimuli were presented in quiet and in babble noise at 18 dB and 12 dB signal-to-noise ratios. TFS randomization had a measurable detrimental effect on quality ratings for speech in quiet and a smaller effect for speech in background babble. Subjects with normal hearing and subjects with sensorineural hearing loss provided similar quality ratings for noise-vocoded speech.

1. Introduction

Many of the approximately 35 million Americans with hearing loss are candidates for hearing aids (Kochkin, 2010). While recent clinical trials document the benefit of hearing aids (e.g., Larson et al., 2000), only 20–40% of individuals who are candidates actually own them (Dubno et al., 2008; Kochkin, 2010). Of those who own hearing aids, approximately 65–80% are satisfied with their instruments (Dubno et al., 2008; Kochkin, 2010). Sound quality, along with speech intelligibility, is correlated with overall user satisfaction with hearing aids (Kochkin, 2010). Modifications to the signal caused by environmental noise and/or by nonlinear and linear hearing aid signal processing can affect both speech intelligibility and speech quality (e.g., Moore and Tan, 2003; Arehart et al., 2007; Davies-Venn et al., 2007; Tan and Moore, 2008; Anderson et al., 2009; Arehart et al., 2010). These modifications affect speech in both the spectral and temporal domains.

A complex signal such as speech can be separated into multiple frequency bands. The temporal information in each band can be divided into two components. The temporal envelope (TE) is the slowly varying amplitude modulation. Temporal fine structure (TFS) is the more rapidly varying carrier signal (Shannon et al., 1995). In recent years, researchers have developed several speech quality indices to predict perceptual effects caused by changes in one or more signal characteristics. However, these indices may not accurately reflect the impact on speech quality of modifications to the TFS of the signal. The Perceptual Evaluation of Speech Quality (PESQ) index (Beerends et al., 2002) focuses on the change in excitation patterns introduced by the signal modifications, and will be affected by signal modifications only to the degree that the modifications change the average power in each band. The PEMO-Q quality index (Huber and Kollmeier, 2006) measures the change in the signal envelope modulation. The Hearing Aid Speech Quality Index (HASQI) (Kates and Arehart, 2010) measures the change in envelope time-frequency modulation and the change in the signal long-term spectrum. Neither PEMO-Q nor HASQI directly measures the change in TFS, although both indices will be indirectly affected by how changes in TFS are reflected in changes to the envelope modulation. For example, additive noise will randomize the TFS and also reduce the depth of the envelope modulation. Another quality model (Moore et al., 2004; Moore and Tan, 2004; Tan et al.,

Abbreviations: TFS, temporal fine structure; TE, temporal envelope; TSNR, signal-to-noise ratio; BC, band cutoff

* Corresponding author. Permanent address: University of Colorado Hospital, 1635 Aurora Court Suite 6200, Mail Stop F736, Aurora, CO 80045, USA. Tel.: +1 720 848 7218; fax: +1 720 848 2857.

E-mail addresses: melinda.anderson@uch.edu (M.C. Anderson), kathryn.arehart@colorado.edu (K.H. Arehart), james.kates@colorado.edu (J.M. Kates).

0378-5955/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.heares.2013.11.011

2004) uses the normalized cross-correlation between the output and input signals after the signal has been filtered into bands to estimate the effect of noise and nonlinear distortion on the signal. The signals are divided into 30-ms segments and the cross-correlation between the input and output is computed. The cross-correlation value is normalized by the signal energy in the segments, and a level weighting function is applied to reduce the importance of low-intensity segments. The normalized and weighted cross-correlations are then averaged within each frequency band. The cross-correlation directly measures changes in the TFS, but assumes the same sensitivity to TFS modification at all frequencies despite the reduction in neural phase locking at frequencies above 1500 Hz (Johnson, 1980).

In summary, current models of speech quality perception focus primarily on TE modifications without accurately quantifying the effects of TFS modifications. The data presented here provide information regarding the effects of TFS modifications on speech quality perception. This information is needed, in part, for improvements in models of speech quality perception to more accurately predict the effects of hearing aid signal processing.

Traditional hearing aid processing, such as dynamic-range compression or noise suppression, directly modifies the signal envelope (Anderson et al., 2009). However, recently developed hearing-aid signal processing algorithms directly modify the TFS of the signal. Hearing aid processing algorithms are being developed that replace the high frequencies in the input speech signal with noise modulated by the speech envelope (Kates, 2011a; Ma et al., 2011). The envelope at high frequencies is preserved, while the speech TFS is replaced by that of the random noise. Because the original TFS has been replaced, the processed high-frequency output signal is uncorrelated with the input. The accuracy of the feedback path estimation in feedback cancellation increases as the cross-correlation of the input and output decreases, so the TFS replacement improves the performance of the adaptive feedback cancellation implemented in the device (Kates, 2011a; Ma et al., 2011). While these techniques improve stability, the impact of this type of signal processing on speech quality has not been determined.

Other types of hearing aid processing modify the TFS even though the processing objective is to change the signal envelope or spectrum. An example of this involves shifting the high frequency content of a signal. This shift may be implemented in multiple ways: 1) by moving a block of frequencies to a lower frequency region (Korhonen and Kuk, 2008), 2) by proportionally reducing the frequencies of the signal components above a cutoff frequency (Aguilera-Muñoz et al., 1999; Simpson et al., 2005; Souza et al., 2013), or 3) by shifting the frequencies towards the center of each frequency band in a multi-band system (Kulkarni et al., 2012). These frequency-shifting strategies reduce the correlation between the TFS of the processed signal and that of the original unprocessed version, and it is important to understand the impact of these TFS changes on speech quality.

The frequency modification algorithms described above are designed to maximize speech understanding and usable gain for hearing aid users. While maintaining high levels of speech intelligibility is important for user satisfaction with hearing aids, it is possible to have high levels of intelligibility combined with poor sound quality (e.g., Preminger and Van Tasell, 1995; Souza et al., 2013). To date, no literature explicitly explores the effects of TFS manipulation on speech quality perception. The focus of the present study was to determine speech quality with parametric variation of TFS randomization in specific frequency regions for situations in which speech intelligibility remains at high levels. This study used noise vocoding to explore the effects of TFS randomization on speech quality perception.

Vocoding has been used to study the separate effects of TE and TFS on speech perception (e.g., Dudley, 1939; Shannon et al., 1995). To vocode a signal, it is filtered into a number of bands, and the envelope of each band is used to modulate a carrier signal (either noise or sine waves). For the noise vocoder, all frequencies within a band receive the same modulation. The uniform modulation causes a nearly constant amplitude across the band, resulting in a stair-step spectrum shape (Stone and Moore, 2003). As the number of bands increases, the spectrum becomes smoother and more of the envelope time-frequency modulation remains intact (Kates, 2011b). Each band is re-filtered (using the same filter bank) to remove any out-of-band components and the bands are combined. The resulting signal includes the modified TE and limited portions of the original TFS (dependent on specific envelope filter cutoff frequencies and whether noise or tone carriers are used). In the vocoding process, the TFS is modified, not removed. The TFS of the vocoded output comprises two components. The first is the residual speech TFS, and the second is the TFS associated with the vocoder carrier. The amount of each type of TFS is dependent on the signal processing configuration of the vocoder.

The Gaussian noise traditionally used in noise vocoders has intrinsic random amplitude fluctuations over time, meaning that at any given point in time, the noise has its own random envelope. This intrinsic noise envelope may have a detrimental impact on speech understanding when combined with the temporal envelope of the speech (Whitmal et al., 2007; Stone et al., 2008; Souza and Rosen, 2009). It is possible to remove a substantial portion of the envelope from Gaussian noise. Both noise-envelope-intact vocoding noise and noise-envelope-removed vocoding noise have been described in the literature, with improved speech intelligibility for noise-envelope removed vocoding (Whitmal et al., 2007; Kates, 2011b).

Regardless of the type of carrier used in the vocoding process, a signal processing confound exists (Kates, 2011b). Although vocoding is designed to remove original TFS cues, it also affects the TE. The results of Kates (2011b) show that vocoding may not accurately reproduce envelope behavior across frequency bands. Each TFS modification technique considered in Kates (2011b) resulted in a loss in the accuracy of the envelope time-frequency modulation reproduction. In addition, while vocoding removes original TFS, TFS is still present in the vocoded signal and may show resemblance to the original TFS. Even with this limitation, vocoding is still a valuable signal processing tool because it provides a consistent method of TFS modification and allows for the study of TFS cues in speech perception.

While the role of TFS in speech quality perception is unclear, recent studies have examined how TFS influences speech intelligibility (e.g., Shannon et al., 1995; Qin and Oxenham, 2003; Lorenzi et al., 2006; Başkent, 2006; Hopkins et al., 2008; Hopkins and Moore, 2009, 2010). For a single talker in quiet, speech with limited original TFS is highly intelligible for subjects with normal hearing and for subjects with mild to moderate hearing loss due to cochlear damage (Shannon et al., 1995; Başkent, 2006). However, when listening to speech in the presence of competition, original TFS plays a more important role (Qin & Oxenham, 2003; Lorenzi et al., 2006; Başkent, 2006; Hopkins et al., 2008; Hopkins and Moore, 2009, 2010). When presented with a competing sound, speech with primarily TE cues is insufficient for high speech intelligibility for both subjects with normal hearing and subjects with hearing loss. The inclusion of original TFS for speech in the presence of noise improves speech understanding to differing degrees for subjects with normal hearing and subjects with sensorineural hearing loss. Subjects with normal hearing achieve better understanding of speech in noise from inclusion of original TFS up to about 5000 Hz, while subjects with sensorineural hearing loss