Effect of band power weighting on understanding sentences synthesized with temporal information

Fuqiang Ye, Dingchang Zheng, and Fei Chen

Citation: The Journal of the Acoustical Society of America **145**, EL168 (2019); doi: 10.1121/1.5091757
View online: https://doi.org/10.1121/1.5091757
View Table of Contents: https://asa.scitation.org/toc/jas/145/2
Published by the Acoustical Society of America

Effect of band power weighting on understanding sentences synthesized with temporal information

Fuqiang Ye
Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China
11749176@mail.sustc.edu.cn

Dingchang Zheng
Health and Wellbeing Academy, Faculty of Medical Science, Anglia Ruskin University, Chelmsford, United Kingdom
dingchang.zheng@anglia.ac.uk

Fei Chen[a)]
Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China
fchen@sustc.edu.cn

Abstract: This work examined the effect of band power weighting on understanding stimuli synthesized with temporal envelope or Hilbert-fine-structure (HFS) waveforms. The power of modulated carrier in a vocoder model or HFS waveform was level-matched to that of the bandpass filtered signal (matched condition) or equalized across bands (flat condition). The processed stimuli were played to normal-hearing listeners to recognize. For both vocoded and HFS stimuli, there was no significant performance difference between the matched and flat power-weighting conditions, suggesting that band power weighting did not notably influence the intelligibility of stimuli synthesized with temporal information from a few bands.

[DDO]
Date Received: November 2, 2018 **Date Accepted:** February 4, 2019

1. Introduction

Temporal envelope and fine-structure are two important temporal cues for speech perception. A vocoder model has been long used to assess the perceptual contribution of temporal envelope and how that contribution varies with parameter changes in the vocoding process (e.g., Shannon *et al.*, 1995; Dorman *et al.*, 1997). Vocoder processing involves decomposition of the input speech signal into multiple analysis bands, followed by envelope extraction and modulation of a carrier signal. There are two common types of carrier signals used in vocoder processing, i.e., pure-tone and white noise, yielding tone-vocoded and noise-vocoded stimuli, respectively (e.g., Dorman *et al.*, 1997; Chen *et al.*, 2017). Early studies have reported that many vocoder processing parameters affect the intelligibility of vocoded stimuli, including the number of bands (e.g., Shannon *et al.*, 1995; Dorman *et al.*, 1997), the cutoff frequency used to extract the envelope waveform (e.g., Shannon *et al.*, 1995; Dorman *et al.*, 1997), spectral holes (e.g., Shannon *et al.*, 2002; Kasturi *et al.*, 2002), and carrier signal type (i.e., pure-tone or noise) (e.g., Whitmal *et al.*, 2007; Chen *et al.*, 2017). Vocoder models preserve only temporal envelope information in each band, while discarding underlying fine-structure (or phase) information. Most of the existing cochlear implant (CI) speech processors capture multi-band temporal envelope waveforms of sound inputs, and then generate electric stimulations that excite CI users' residual auditory nerves directly. As vocoder processing aims to transfer only those acoustic cues that are present for CI users, it simulates the signal processing of a CI (Chen and Loizou, 2011). In contrast to vocoded stimuli, temporal fine-structure based stimuli contain only fast-varying Hilbert fine-structure (HFS) information (commonly extracted with Hilbert transform), while removing slow-varying amplitude variation information (e.g., Smith *et al.*, 2002; Gilbert and Lorenzi, 2006; Lorenzi *et al.*, 2006). Studies have showed that stimuli synthesized with multiple fine-structure waveforms of all analysis bands (discarding the envelope waveform) also carried intelligibility information. Notably, Smith *et al.*

[a)]Author to whom correspondence should be addressed.

(2002) showed that speech processed in 1–2 bands to contain only temporal fine-structure cues was highly intelligible. Additionally, Gilbert and Lorenzi (2006) reported that the contribution of the recovered envelope cues in the HFS stimuli diminished when more than eight bands were used. Fine-structure also carries important information for lexical tone identification, music appreciation and other tasks requiring pitch perception (e.g., Smith *et al.*, 2002). The analysis of temporal envelope and fine-structure largely simulates the sound processing by human auditory system. Auditory neurons respond to temporal envelope and fine-structure via their fluctuations in the short-term rate of firing and by the synchronization of nerve spikes to a specific phase of temporal fine-structure, respectively (Lorenzi *et al.*, 2006).

Some early work studied the perceptual importance of spectral bands only containing temporal envelope waveforms (e.g., Shannon *et al.*, 2002; Kasturi and Loizou, 2002; Apoux and Bacon, 2004). Using the spectral-hole method to assess the intelligibility of speech signals with a single hole in each of several spectral bands, Kasturi and Loizou (2002) found uneven weighting across bands for vowel identification and further found that the generated band-weighting function was relatively flat for consonant identification. Also using the spectral-hole method, Apoux and Bacon (2004) showed that all bands contributed equally to consonant identification in the absence of noise. When temporal information from all bands is present in the synthesis of vocoded and HFS stimuli, it is essential to take into account the impact of power weighting across bands on intelligibility. Regarding the synthesis of HFS stimuli, band power weighting has been commonly used (e.g., Gilbert and Lorenzi, 2006; Lorenzi *et al.*, 2006), because HFS waveforms do not have temporal amplitude variation. Vocoder processing in several studies adjusted the level of vocoded output bands to match those of the bandpass filtered signals (processed by analysis filters) (e.g., Whitmal III *et al.*, 2007; Rosen *et al.*, 2015). However, early vocoder research did not include substantial assessments of the effects of power weighting across bands on intelligibility, and in some studies the powers of vocoded output bands were not matched to those of the bandpass filtered signals processed by analysis filters (e.g., Shannon, *et al.*, 1995; Dorman *et al.*, 1997; Qin and Oxenham, 2005).

The aim of this work was to evaluate how power weighting across bands affects the intelligibility of three types of stimuli synthesized with temporal information, including vocoded (both tone- and noise-vocoded) and HFS stimuli. Two band power weighting conditions were implemented in the synthesis of vocoded and HFS stimuli: (1) matched weighting, wherein the level of modulated carrier or temporal fine-structure waveform in each band was matched to that of the bandpass filtered signal processed by the analysis filter and (2) flat weighting, wherein the levels of all modulated carriers or temporal fine-structure waveforms were equalized. The matched and flat weighting conditions preserved and removed the original power weighting across bands, respectively. We further examined whether removing the original power weighting across bands would deteriorate the understanding of vocoded and HFS sentences.

2. Methods

2.1 Subjects

Twelve (7 males and 5 females) native-Mandarin-Chinese listeners (18–23 years old) participated in this study. All participants were undergraduate students at Southern University of Science and Technology, and were paid for their participation. All subjects had normal-hearing with pure-tone thresholds better or equal to a 20 dB hearing level from 250 to 8000 Hz. The study protocol was approved by the Human Research Ethics Committee for Non-clinical Faculties of Southern University of Science and Technology.

2.2 Materials

The speech material consisted of sentences taken from the Mandarin Hearing in Noise Test (MHINT) database (Wong *et al.*, 2007), which includes 24 lists of 10 sentences, with each sentence containing ten key words. All of the sentences were produced by a male speaker with a fundamental frequency range of 75–180 Hz. A steady-state speech-shaped noise was used to corrupt the clean sentences before they were processed by the vocoder or HFS synthesizer. A random noise segment of the same length as the clean speech signal was cut out of the noise recordings, scaled to the desired input signal-to-noise (SNR) level, and added to the speech signals at a 3 dB (for tone-vocoding and HFS processing) or 5 dB (for noise-vocoding processing) input SNR level. Input SNR levels were chosen based on performance in a pilot study to avoid ceiling/floor effects.

2.3 Signal processing

Three types of signal processing were used in this work: tone-vocoding, noise-vocoding, and HFS synthesis. To implement the tone-vocoding processing, speech signals (3 dB SNR) were processed through a pre-emphasis filter (first-order high-pass filter with a 1200-Hz cutoff frequency). Then, the signals were bandpass-filtered into $N = 4$ or $N = 8$ frequency bands between 80 and 6000 Hz with sixth-order Butterworth filters. The cut-off frequencies for the band allocation of bandpass filters were (in Hz): {80, 426, 1158, 2710, 6000} for the $N = 4$ experiment and {80, 221, 426, 724, 1158, 1790, 2710, 4050, and 6000} for the $N = 8$ experiment (Greenwood, 1990). The envelope was extracted from each band by half-wave rectification and low-pass filtering with a 160-Hz cutoff frequency by way of a fourth-order Butterworth filter. Sine waves at the center frequencies of the bandpass filters were generated with amplitudes modulated by the extracted envelopes. All amplitude-modulated sine waves from the resultant set of bands were weighted by two conditions: (1) flat weighting with all waves adjusted to have the same root-mean-square (RMS) energy and (2) matched weighting with band-specific temporal waveform adjusted to have an RMS energy that was the same as that of the bandpass filtered signal. All weighted waves were summed to generate a tone-vocoded stimulus, whose RMS energy scaling was performed with respect to the input speech signal.

Implementation of the noise-vocoding processing was similar to that of the tone vocoder, except that a white noise was used as the carrier signal instead of a sine wave and amplitude-modulated by the extracted envelope. The output from each band was further band-limited with the same bandpass filter at that band. All amplitude-modulated noises (with band-limiting processing) were subjected to flat or matched weighting, and summed to generate a noise-vocoded stimulus, with its amplitude adjusted to match the RMS power of the original input speech signal.

The signal processing for implementing the HFS synthesis was conducted as described by Lorenzi *et al.* (2006). Input speech signals (3 dB SNR) were first split into $N = 4$, 6, 8, or 10 frequency bands. The cutoff frequencies for the band allocation of bandpass filters were (in Hz): {80, 426, 1158, 2710, 6000} for $N = 4$, {80, 281, 612, 1158, 2060, 3547, 6000} for $N = 6$, {80, 221, 426, 724, 1158, 1790, 2710, 4050, and 6000} for $N = 8$, and {80, 188, 334, 532, 798, 1158, 1644, 2301, 3197, 4384, 6000} for $N = 10$ (Greenwood, 1990). The Hilbert transform was applied to the band-passed signals to obtain HFS waveforms. The envelope components were discarded, while the N-band HFS components were subjected to flat or matched weighting, summed, and adjusted to match the RMS level of the original input speech signal. The examples of all types of stimuli will be provided upon request.

2.4 Procedure

Stimuli were played to listeners diotically through an HD 650 circumaural headphone (Sennheiser, Germany) set at a comfortable listening level in a sound booth. Before testing, each subject participated in a 10-min training session and was given three lists of ten MHINT sentences. The training session familiarized the subjects with the test procedure and conditions. During training, the subjects were allowed to read transcripts of the training sentences while listening to them. Three testing conditions (tone-vocoded, noise-vocoded, and HFS synthesis; implemented with $N = 4$ bands and matched weighting) were used in training.

In the testing session, the order of testing conditions was randomized across subjects, and the subjects were asked to repeat orally all of the words they heard. In addition, the lists were randomized across listeners. The sentences used during testing were not the same as any of the training sentences. Each subject participated in a total of 16 conditions [4 tone-vocoded conditions ($N = 4$ and 8, with 2 weighting conditions) + 4 noise-vocoded conditions ($N = 4$ and 8, with 2 weighting conditions) + 8 tone-vocoded conditions ($N = 4$, 6, 8, and 10, with 2 weighting conditions)]. One list of ten Mandarin sentences was used per testing condition, and none of the sentences was repeated across conditions. Subjects were allowed to listen to each stimulus a maximum of three times, and were asked to repeat as many words as they could recognize. The participants used a simple custom software interface designed for the listening experiment to control the auditory delivery of the processed stimuli. An investigator accompanied each participant and scored his/her responses. The oral responses were also digitally audio-recorded for the purpose of later verification. A 5-min break was given every 30 min to avoid listening fatigue. The intelligibility score for each condition was computed as the ratio between the number of correctly recognized words and the total number of words contained in each MHINT list.

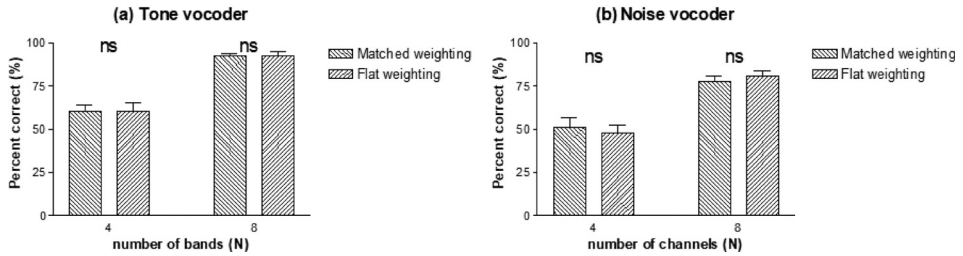

Fig. 1. Mean sentence recognition scores for all (a) tone-vocoded and (b) noise-vocoded conditions. The error bars denote ±1 standard error of the mean. "ns" denotes that the score difference between the two weighting conditions is non-significant.

2.5 Data analysis

Percentage correct scores were converted to rational arcsine units with the rationalized arcsine transform (Studebaker, 1985). Two-way repeated measures analyses of variance (rmANOVAs) were used to detect statistically significant variations in percentage correct score (dependent variable) with the number of bands and the weighting condition as within-subject factors. The paired t-test was used for pair-wise *post hoc* analyses.

3. Results

The mean recognition scores for tone-vocoded stimuli under all testing conditions are shown in Fig. 1(a). A two-way rmANOVA indicated a significant effect of the number of bands ($F_{1,11} = 95.997$, $p < 0.005$), but not weighting condition ($F_{1,11} = 0.185$, $p = 0.676$), and no significant interaction between these two factors ($F_{1,11} = 0.133$, $p = 0.723$) for tone-vocoded stimuli. The mean recognition scores for noise-vocoded stimuli under all testing conditions are shown in Fig. 1(b). A two-way rmANOVA revealed a significant effect of the number of bands ($F_{1,11} = 101.781$, $p < 0.005$), but not weighting condition ($F_{1,11} = 0.008$, $p = 0.929$), and no significant interaction between these two factors ($F_{1,11} = 2.147$, $p = 0.171$) for noise-vocoded stimuli. In both Figs. 1(a) and 1(b), *post hoc* analyses showed that for all paired comparisons (i.e., matched weighting vs flat weighting) with the same number of bands, there were no significant differences between the two types of weighting conditions ($p > 0.05$).

The mean recognition scores of HFS-based stimuli under all testing conditions are shown in Fig. 2. A two-way rmANOVA revealed a significant effect of the number of bands ($F_{1,11} = 21.776$, $p < 0.005$), but not weighting condition ($F_{1,11} = 0.007$, $p = 0.936$), and no significant interaction between the two factors ($F_{1,11} = 0.650$, $p = 0.589$). *Post hoc* analyses showed that for all paired comparisons (i.e., matched weighting vs flat weighting) with the same number of bands, there were no significant differences between the two types of weighting conditions ($p > 0.05$).

4. Discussion and conclusions

The present work extended those prior findings (e.g., Shannon *et al.*, 2002; Kasturi and Loizou, 2002; Apoux and Bacon, 2004) by further examining the perceptual importance of temporal information (i.e., temporal envelope and fine-structure) in various frequency regions for sentence recognition. In contrast to prior studies implementing a spectral hole in speech synthesis, this work modified the original band power weighting (i.e., matched weighting) to include a flat weighting condition, thereby removing the original band power weighting in the synthesis of temporal information (both envelope and fine-structure) based stimuli. We used flat power weighting across

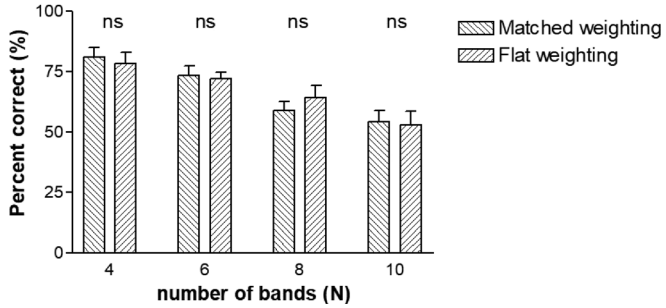

Fig. 2. Mean sentence recognition scores for all HFS-based conditions. The error bars denote ±1 standard error of the mean. "ns" denotes that the score difference between the two weighting conditions is non-significant.

bands in vocoder processing and in HFS based stimulus synthesis. Use of a flat power weighting (band power not adjusted on an in-band basis) obviated the energy difference across the outputs of all bands, leaving perceptual cues to be carried largely by the temporal information contained in the vocoded or HFS waveform. Compared with the band-specific power weighting condition, removal of the original power weighting among bands did not alter speech recognition performance significantly, suggesting that the original power weighting across bands contributed little to the recognition of vocoded and HFS-based sentences in the scenarios of speech synthesis with multiple envelope or FHS waveforms from a few bands (e.g., 4 to 8).

The finding of this work has useful implication for experiments using vocoded stimuli as CI simulation. While the perceptual influences of many factors have been examined for vocoded stimuli, the impact of band power weighting in vocoder processing was not widely studied, hence rendering inconsistence (i.e., whether or not matching the powers of vocoded output bands to those of the bandpass filtered signals processed by analysis filters) in vocoder processing in literature. The present work indicated that the choice of band power weighting did not notably influence the understanding of vocoded (either tone- or noise-vocoded) stimuli in CI simulation.

In conclusion, the present work examined the effect of power weighting across all output bands on the intelligibility of noise- or tone-vocoded and HFS-based stimuli. Employing flat band power weighting did not degrade the understanding of vocoded or HFS sentences, suggesting that under a fixed number of spectral bands, removing the original band power weighting does not impact the understanding of stimuli synthesized with temporal information, including temporal envelope or fine-structure.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61828104 and 61571213), the Research Foundation of Department of Science and Technology of Guangdong Province (Grant No. 2018A050501001), and the Shenzhen High-level Overseas Talent Program (Grant No. KQJSCX201803193000018).

References and links

Apoux, F., and Bacon, S. P. (**2004**). "Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise," J. Acoust. Soc. Am. **116**, 1671–1680.

Chen, F., and Loizou, P. C. (**2011**). "Predicting the intelligibility of vocoded speech," Ear Hear. **32**, 331–338.

Chen, F., Zheng, D. C., and Tsao, Y. (**2017**). "Effects of noise suppression and envelope dynamic range compression to the intelligibility of vocoded sentences for a tonal language," J. Acoust. Soc. Am. **142**, 1157–1166.

Dorman, M. F., Loizou, P. C., and Rainey, D. (**1997**). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," J. Acoust. Soc. Am. **102**, 2403–2411.

Gilbert, G., and Lorenzi, C. (**2006**). "The ability of listeners to use recovered envelope cues from speech fine structure," J. Acoust. Soc. Am. **119**, 2438–2444.

Greenwood, D. D. (**1990**). "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605.

Kasturi, K., Loizou, P. C., Dorman, M., and Spahr, T. (**2002**). "The intelligibility of speech with 'holes' in the spectrum," J. Acoust. Soc. Am. **112**, 1102–1111.

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

Qin, M., and Oxenham, A. (**2005**). "Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers," Ear Hear. **26**, 451–460.

Rosen, S., Zhang, Y., and Speers, K. (**2015**). "Spectral density affects the intelligibility of tone-vocoded speech: Implications for cochlear implant simulations," J. Acoust. Soc. Am. **138**, EL318–EL323.

Shannon, R. V., Galvin, J. J., and Baskent, D. (**2002**). "Holes in hearing," J. Assoc. Res. Otolaryngol. **3**, 185–199.

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

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception," Nature (London) **416**, 87–90.

Studebaker, G. A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear Res. **28**, 455–462.

Whitmal, N. A. III, Poissant, S. F., Freyman, R. L., and Helfer, K. S. (**2007**). "Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience," J. Acoust. Soc. Am. **122**, 2376–2388.

Wong, L. L., Soli, S. D., Liu, S., Han, N., and Huang, M. W. (**2007**). "Development of the Mandarin Hearing in Noise Test (MHINT)," Ear Hear. **28**, 70S–74S.