Revised submission

Understanding low-pass-filtered Mandarin sentences: Effects of fundamental frequency contour and single-channel noise suppression

Lei Wang

Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Xueyuan Road 1088#, Xili, Nanshan District, Shenzhen 518055, China

wangl6@mail.sustc.edu.cn

Dingchang Zheng

Health and Wellbeing Academy, Faculty of Medical Science, Anglia Ruskin University, Chelmsford, UK

dingchang.zheng@anglia.ac.uk

Fei Chen [a)]

Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Xueyuan Road 1088#, Xili, Nanshan District, Shenzhen 518055, China

fchen@sustc.edu.cn

Saturday, February 24, 2018

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

PACS number: 43.71.Gv, 43.71.Es

Keywords: Low-pass filtering, fundamental frequency contour, noise suppression.

Running title: Intelligibility of low-pass-filtered sentences

Abstract: The present work assessed the effects of flattening the fundamental frequency (F0) contour and processing by single-channel noise suppression on the intelligibility of low-pass (LP)-filtered (LPF) sentences. The original F0 contour was replaced by an average flat F0 contour or treated by single-channel noise suppression, followed by application of LP filtering to Mandarin sentences. Processed stimuli were presented to normal-hearing listeners to recognize. Flattening the F0 contour significantly affected the understanding of LPF sentences. Noise suppression by existing single-channel algorithms did not improve the intelligibility of LPF sentences.

1. Introduction

Full-spectrum speech signal contains many acoustic cues that are important for speech perception. In some cases, the lack of a certain acoustic cue does not significantly affect the understanding of full-spectrum speech. One explanation for this observation relates to the contribution of the specific acoustic cue for speech intelligibility. For instance, in vocoder simulation, phase information is largely discarded, and only multichannel temporal envelope waveforms are preserved to synthesize vocoded speech. Under quiet conditions, normal-hearing (NH) listeners could almost perfectly understand vocoded speech with four envelope waveforms extracted from four channels (Shannon et al., 1995). This indicates that multichannel envelope waveforms alone carry important information for speech perception. A second explanation relates to the top-down process used by human listeners to compensate for distortions in speech signals. In other words, other cues (such as contextual cues or lexical, syntactic, or semantic information from speech) might be used to retrieve the meanings of speech containing waveform distortions. For instance, Fogerty and Kewley-Port (2009) studied the perception of temporally interrupted speech containing only vowel or consonant segments, with the rest replaced by white noise. Vowel sentences (in which consonants were replaced) were highly intelligible. Chen et al. (2014) recently examined the effects of fundamental frequency (F0) contour on Mandarin sentence intelligibility. Even when the original F0 contour was replaced by a flat F0 contour, NH listeners still achieved almost-perfect understanding of Mandarin sentences in quiet.

Numerous studies have employed low-pass (LP)-filtered (LPF) speech to simulate speech perception by listeners with high-frequency (HF) hearing impairment (e.g., Chen and Chan, 2016; Zhang and McPherson, 2008). In simulation studies with NH listeners, Zhang and McPherson (2008) showed that low-frequency (LF) cuts in hearing aid settings may negatively affect vowel recognition and Mandarin tone recognition under adverse noise conditions. Bhargava and Başkent (2012) assessed the combined effect of LPF with cutoff frequencies of 500–3000 Hz and periodic interruptions on speech intelligibility. The intelligibility of combined manipulations was lower than each manipulation alone, even when there was no effect from a single manipulation. The contribution of LF hearing for speech recognition was further demonstrated in studies of combined electric-and-acoustic hearing. In these studies, hearing-impaired (HI) patients were implanted with electrodes to restore their HF hearing and utilized their residual LF hearing (typically 20–60 dB hearing level up to 750 Hz, and severe-to-profound hearing loss at ≥1000 Hz). Combined electric-and-acoustic hearing provided HI listeners with much better speech understanding, particularly in noise, than did electric hearing in conventional cochlear implantation (e.g., Gantz and Turner, 2003). These studies have increased our knowledge on speech perception by HI listeners; however, the mechanism underlying the perception of LPF speech is still not well understood. The present work specially assessed the effects of two manipulations (i.e., flattening the F0 contour and processing by single-channel noise suppression) on the intelligibility of LPF sentences.

The LF regions contain rich acoustic information for speech intelligibility, including vowels (characterized by long duration, formant structure, and low frequency) and the F0 contour. The F0 contour is an important acoustic cue for lexical tone identification and speech understanding in adverse conditions (e.g., Chen et al., 2014; Nie et al., 2005). Therefore, this study hypothesized that the F0 contour played an important role in the perception of LPF speech. More specifically, the first aim of this study (Experiment 1) was to assess if the F0 contour contributed to the intelligibility of LPF sentences.

Speech understanding in noise has been a longstanding challenge for both NH and HI listeners. Most existing single-channel noise suppression algorithms are unable to improve intelligibility for NH listeners (e.g., Li et al., 2011; Hu and Loizou 2007), but may improve speech understanding for HI listeners who are fitted with assistive hearing devices (e.g., Chen et al., 2015; Yang and Fu, 2005). Furthermore, it is unknown whether existing single-channel noise suppression algorithms can improve the intelligibility of LPF speech (i.e., simulating the speech perception of HI listeners with residual LF hearing). Therefore, the second aim of this study (Experiment 2) was to examine the effect of single-channel noise suppression, using four types of existing single-channel noise-suppression algorithms, on the intelligibility of LPF sentences.

2. Experiment 1: Effect of flattening the F0 contour on the intelligibility of LPF speech

2.1 Subjects and materials

This experiment involved 15 (9 males and 6 females) listeners with NH (pure-tone thresholds better than 20 dB HL at octave frequencies from 125 to 8000 Hz in both ears). All subjects were native speakers of Mandarin Chinese and were paid for their participation. The subjects' age ranged from 19 to 23 years (mean age = 21 years), with the majority being undergraduate students at Southern University of Science and Technology. Speech material comprised sentences extracted from the Mandarin Hearing in Noise Test (MHINT) database (Wong et al., 2007). The MHINT corpus includes 24 lists, each with 10 sentences and 10 keywords per sentence. All sentences were spoken by a male native Mandarin Chinese speaker having a fundamental frequency of 75–180 Hz, and were recorded at a sampling rate of 16 kHz.

2.2 Signal processing

Two types of stimuli were synthesized in this experiment, including 1) condition LPF, in which only LP filtering processing was applied, and 2) condition LPF_F0, in which the LPF speech contained a flat F0 contour. To generate the LPF_F0-processed stimulus, the dynamic F0 contour of each sentence was extracted and replaced by a flattened F0 contour at the mean value for each individual sentence. The PRAAT code to implement the F0 flattening processing is available at http://www.linguistics.ucla.edu/faciliti/facilities/acoustic/praat.html#noisespeech [Last viewed November 01 2017]. The (full-spectrum) F0-contour-flattened stimuli were further processed by LP filtering to generate the LPF_F0-processed stimuli. The LP filtering was implemented by using a linear-phase FIR filter with filter order of 10 × fs/fcut, where fs is the sampling rate (16 kHz) and fcut is the LPF cutoff frequency (500, 750, and 1000 Hz in this experiment).

2.3 Procedure

The experiment was conducted in a soundproof booth. Test stimuli were played to participants binaurally through a circumaural headphone at a comfortable listening level, which was controlled by the participant. Before the testing session, participants engaged in a practice session of 20 sample sentences (10 sentences per signal processing condition at an LPF cutoff frequency of 750 Hz). Different sentences were used between the practice and testing sessions. During the testing session, participants orally repeated all of the keywords that they could recognize. Each sentence could be repeated twice. Each participant attended 6 conditions [i.e., 2 signal processing conditions

(LPF and LPF_F0) × 3 LPF cutoff frequencies (500, 750, and 1000 Hz)]. The test condition had 10 sentences per condition. In addition, the test condition order was randomized across subjects, and subjects were given a 5-minute break every 30 minutes of testing. The intelligibility score for each condition was computed as the ratio between the number of all correctly recognized keywords and the total number of keywords contained in each condition.

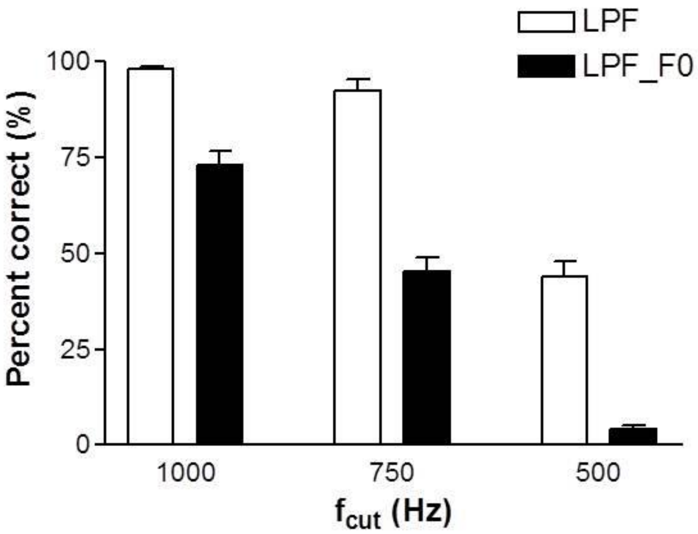

Fig. 1. Mean sentence recognition scores for all conditions in Experiment 1. The error bars denote ±1 standard error of the mean.

2.4 Results and discussion

Figure 1 shows the mean sentence recognition scores for all conditions. Statistical significance was determined by using the percent recognition score as the dependent variable, and the LPF cutoff frequency and signal processing condition as the two within-subject factors. Recognition scores were converted to rational arcsine units by using the rationalized arcsine transform (Studebaker, 1985). Two-way analysis of variance (ANOVA) with repeated measures indicated significant effects of LPF cutoff frequency ($F_{2,28} = 409.34$, $p < .001$) and signal processing condition ($F_{1,14} = 467.85$, $p < .001$), and a nonsignificant interaction ($F_{2,28} = 3.04$, $p = .06$) between LPF cutoff frequency and signal processing condition. Post-hoc pairwise comparisons showed significant differences among all LPF cutoff frequencies (all $p < .001$) and significant differences between the two signal processing conditions (all $p < .01$).

Many studies have suggested that, under quiet conditions, the F0 contour has a minimal effect on sentence intelligibility (e.g., Fogerty and Humes, 2012) or it could be viewed as a redundant cue for speech perception in quiet (e.g., Chen et al., 2014). However, in the present work, the F0 contour was not trivial for understanding LPF sentences. For instance, when the LPF cutoff frequency was set to 750 Hz, flattening the F0 contour caused a 48.5 percentage point intelligibility reduction relative to the LPF condition (i.e., 43.5% vs. 92.0%). Considering that LP filtering removes many important perceptual cues (e.g., formant structure), the F0 contour plays an important role for understanding LPF speech and may be redundant to the low-frequency information on the intelligibility of full-spectrum speech. This finding is different from those of earlier studies evaluating the contribution of the F0 contour in recognizing full-spectrum speech (e.g., Chen et al., 2014; Fogerty and Humes, 2012).

3. Experiment 2: Effect of single-channel noise suppression on the intelligibility of LPF speech

3.1 Subjects and materials

This experiment involved 15 (9 males and 6 females) NH listeners who were different from the participants in Experiment 1. All subjects were native speakers of Mandarin Chinese and were paid for their participation. The subjects' age ranged from 20 to 23 years (mean age = 21 years), with the majority being undergraduate students at Southern University of Science and Technology. Speech materials were the same as in Experiment 1. Steady-state speech-shaped noise (SSN) was used to corrupt the MHINT sentences. To generate the SSN masker, a finite impulse response filter was designed based on the average spectrum of the MHINT sentences. White noise was filtered and scaled to the same long-term average spectrum and level as the sentences. A noise segment of the same length as the clean intact (i.e., full-length) speech signal was randomly cut from the SSN masker, appropriately scaled to reach the desired signal-to-noise ratio (SNR), and added to the speech signals at –3 and –6 dB SNR. The SNR levels were carefully chosen to avoid ceiling/floor effects in understanding noise-corrupted/noise-suppressed LPF sentences in this experiment.

3.2 Signal processing

Noise-corrupted sentences were processed by four representative single-channel noise-suppression algorithms, including the generalized KLT approach, the log minimum mean square error (logMMSE) algorithm, the multiband spectral subtraction algorithm (MB), and the Wiener algorithm based on a priori SNR estimation (Wiener). These algorithms cover the four most-used types of single-channel noise-suppression methods (i.e., subspace, statistical-modeling, spectral-subtractive, and Wiener-filtering approaches), representing the state of the art. Parameters used to implement these algorithms were published previously. Detailed descriptions of the algorithms, including the exact parameters used in the current study, can be found in Hu and Loizou (2007) and Loizou (2007). The MATLAB code used to implement the four noise-suppression algorithms was obtained from Loizou (2007). Noise-corrupted/noise-suppressed sentences were processed by LP filters with cutoff frequencies of 1000 and 750 Hz. LP filtering was implemented as in Experiment 1.

3.3 Procedure

The experimental procedure used in Experiment 2 was essentially the same as used in Experiment 1. In the training session, subjects were familiarized with the testing procedure and conditions. They were given 2 lists of 10 sentences (different from those used in the testing session) and read the transcriptions while listening to the sentences. However, in Experiment 2, each subject was exposed to a total of 20 conditions [i.e., 2 LP cutoff frequencies (1000 and 750 Hz) × 2 SNR levels (–3 and –6 dB) × 5 signal processing conditions (Noisy, Wiener, logMMSE, MB, and KLT)], which were randomized across subjects. As in Experiment 1, one list of 10 sentences was presented per condition, and none of the sentences was repeated across conditions.

3.4 Results and discussion

Figure 2 shows the mean sentence recognition scores for all conditions. Statistical significance was determined by using the percent recognition score as the dependent variable, and LPF cutoff frequency, SNR level, and signal processing condition as the three within-subject factors. Recognition scores were converted to rational arcsine units by using the rationalized arcsine transform (Studebaker, 1985). Three-way ANOVA with repeated measures indicated significant effects of LPF cutoff frequency ($F_{1,14} = 14.83$, $p < .005$), SNR level ($F_{1,14} = 192.71$, $p < .001$), and signal processing condition ($F_{4,56} = 15.26$, $p < .001$). Three-way ANOVA with repeated measures indicated nonsignificant interactions between LPF cutoff frequency and SNR level, between LPF cutoff frequency and signal processing condition, between SNR level and signal processing condition, and among LPF cutoff frequency, SNR level, and signal processing condition (all $p > .05$).

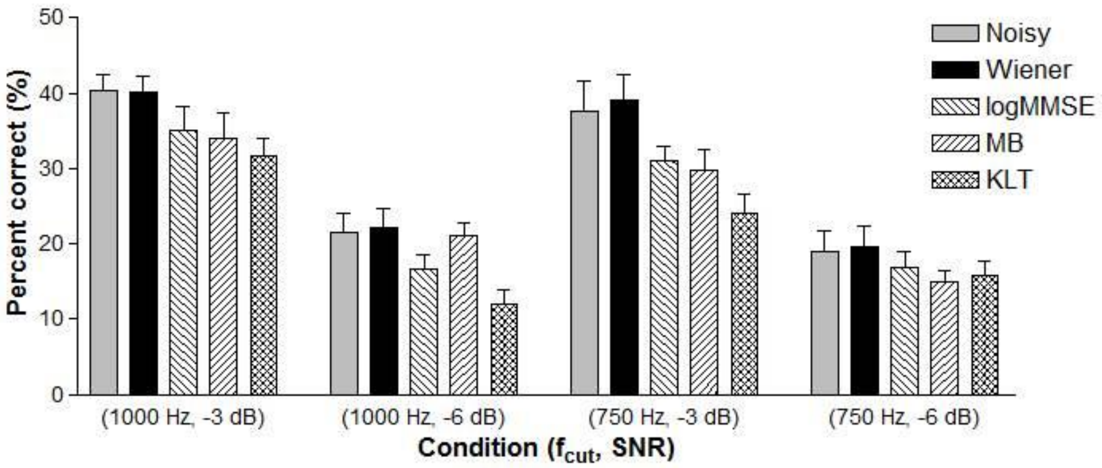

Fig. 2. Mean sentence recognition scores for all conditions in Experiment 2. The error bars denote ±1 standard error of the mean.

Post-hoc pairwise comparisons showed that at all tested conditions with the same LPF cutoff frequency and SNR level as in Fig. 2, the mean recognition scores of noise-suppressed sentences (i.e., Wiener, logMMSE, MB, and KLT in Fig. 2) were not significantly ($p > .05$) larger than those of noisy sentences (i.e., Noisy in Fig. 2). At some conditions, using noise suppression even significantly ($p < .05$) reduced the recognition score compared to that of noisy sentences. At the tested condition (750, –3 dB) in Fig. 2, the mean recognition score of Wiener-processed sentences was slightly larger than that of noisy sentences (i.e., 39.1% vs. 37.6%), but the difference between the mean scores was not significant ($p > 0.05$). In Fig. 2, at each tested condition, the score of Wiener-processed sentences was the largest among the four types of noise-suppression algorithms. This observation suggests that Wiener filtering might have had the least negative influence on the intelligibility of LPF speech among the noise-suppression algorithms examined.

Results in this experiment showed that existing commonly used single-channel noise-suppression algorithms did not improve the intelligibility of LPF speech. Note that because most noise-suppression algorithms were originally designed for full-spectrum speech, this experiment first employed noise-suppression algorithms to process full-spectrum speech, and then applied LPF processing. Alternatively, noise-suppression algorithms could be used for LPF speech. It warrants further work to examine how the order of signal processing (i.e., noise-suppression and LPF) affects the intelligibility of LPF speech. Early work has shown that single-channel noise-suppression algorithms could not improve the understanding of full-spectrum speech in noise (e.g., Li et al., 2011; Hu and Loizou, 2007). Taken together, this work suggested that existing single-channel noise-suppression algorithms may not improve the intelligibility of both full-spectrum and LPF speech.

4. Conclusions

The present work studied the perceptional contribution of the F0 contour and the effect of single-channel noise suppression in understanding LPF Mandarin sentences. The F0 contour plays an important role in understanding LPF sentences and may be redundant to the low-frequency information on the intelligibility of full-spectrum speech. This finding differs from the results of full-spectrum speech perception studies (e.g., Chen et al., 2014; Fogerty and Humes, 2012). This difference may be largely attributed to the increased perceptual weight of the F0 contour under adverse (i.e., low-pass filtering in this study) listening conditions. Although the LF regions possess high local SNR level, which is favorable to noise estimation in noise-suppression processing, use of single-channel noise-suppression processing did not improve the intelligibility of LPF sentences in noise (e.g., in steady-state noise in this work). This deficit in intelligibility performance is consistent with findings observed in full-spectrum speech perception studies (e.g., Li et al., 2011; Hu and Loizou, 2007).

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61571213), and the Basic Research Foundation of Shenzhen (Grant No.

JCYJ20160429191402782).

References and links

Bhargava, P., and Başkent, D. (**2012**). "Effects of low-pass filtering on intelligibility of periodically interrupted speech," J. Acoust. Soc. Am. **131**, EL87–EL92.

Chen, F., and Chan, Fiona W. S. (**2016**). "Understanding frequency-compressed Mandarin sentences: Role of vowels," J. Acoust. Soc. Am. **139**, 1204–1213.

Chen, F., Wong, L. L. N., and Hu, Y., (**2014**). "Effects of lexical tone contour on Mandarin sentence intelligibility," J. Speech Lang. Hear. Res. **57**, 338–34.

Chen, F., Hu, Y., and Yuan, M. (**2015**). "Evaluation of noise reduction methods for speech recognition by Mandarin-speaking cochlear implant listeners," Ear Hear. **36**, 61–71.

Fogerty, D., and Kewley-Port, D. (**2009**). "Perceptual contributions of the consonant-vowel boundary to sentence intelligibility," J. Acoust. Soc. Am. **126**, 847–857.

Fogerty, D., and Humes, L. E. (**2012**). "The role of vowel and consonant fundamental frequency, envelope, and temporal fine structure cues to the intelligibility of words and sentences," J. Acoust. Soc. Am. **131**, 1490–1501.

Gantz, B. J., and Turner, C. (**2003**). "Combining acoustic and electric hearing," Laryngoscope **113**, 1726–1730.

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

Li, J., Yang, L., Zhang, J., Yan, Y., Hu, Y., Akagi, M., and Loizou, P. C. (**2011**). "Comparative intelligibility investigation of single-channel noisereduction algorithms for Chinese, Japanese, and English," J. Acoust. Soc. Am. **129**, 3291–3301.

Loizou, P. C. (**2007**). Speech Enhancement: Theory and Practice (CRC Press, Boca Raton, FL), pp. 1–689.

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

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

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

Wong, L. L. N., 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.

Yang, L., and Fu, Q. J. (**2005**). "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise," J. Acoust. Soc. Am. **117**, 1001–1004.

Zhang, J., and McPherson, B. (**2008**). "Hearing aid low frequency cut: Effect on Mandarin tone and vowel perception in normal-hearing listeners," Folia Phoniatr Logop. **60**, 179–187.