INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Factors Affecting the Intelligibility of Low-pass Filtered Speech

Lei Wang, Fei Chen

Department of Electrical and Electronic Engineering, Southern University of Science and Technology, Shenzhen, China

fchen@sustc.edu.cn

Abstract

Frequency compression is an effective alternative to conventional hearing aids amplification for patients with severe-to-profound middle- and high-frequency hearing loss and with some low-frequency residual hearing. In order to develop novel frequency compression strategy, it is important to first understand the mechanism for recognizing low-pass filtered speech, which simulates high-frequency hearing loss. The present work investigated three factors affecting the intelligibility of low-pass filtered speech, i.e., vowels, temporal fine-structure, and fundamental frequency (F0) contour. Mandarin sentences were processed to generate three types (i.e., vowel-only, fine-structure-only, and F0-contour-flattened) of low-pass filtered stimuli. Listening experiments with normal-hearing listeners showed that among the three factors assessed, the vowel-only low-pass filtered speech was the most intelligible, which was followed by the fine-structure-based low-pass filtered speech. Flattening F0-contour significantly deteriorated the intelligibility of low-pass filtered speech.

Index Terms: Speech intelligibility, low-pass filtered speech.

1. Introduction

When performing hearing aid (HA) fitting, it has always been challenging to fit patients with sloping severe-to-profound high-frequency (HF) hearing loss, notwithstanding that their hearing ranges from normal to moderate loss at low frequencies (LFs). Soft HF sounds often remain inaudible because the amount of gain needed is unreachable by conventional hearing aids. Studies have suggested that HF amplification may not benefit and may even degrade speech intelligibility of patients with HF hearing loss [e.g., 1-2]. Recently, frequency compression (FC) techniques have emerged as a trend in clinical HA fitting. FC delivers the inaudible information from unaidable HF regions to the aidable LF regions [3]. It aims to help patients make use of their residual hearing in LF regions to hear HF information after this information is converted and presented in lower frequencies. Therefore, hearing-impaired patients' low-frequency residual hearing or their ability to understand LF-dominated speech is important for the performance of the frequency-compression-based hearing aids.

In order to develop novel FC strategy, it is important to first understand the mechanism for speech perception with low-frequency residual hearing [e.g., 4-5], which is the main objective of this work. The contribution of low-frequency residual hearing for speech recognition was also demonstrated in studies of combined electric-and-acoustic hearing, where hearing-impaired patients were implanted with electrodes to restore their HF hearing, and also utilized their LF residual hearing (typically 20 to 60 dB HL up to 750 Hz, and severe-to-profound hearing loss at 1000 Hz and above). A number of studies have shown that this combined electric-and-acoustic hearing provides hearing-impaired listeners with much better speech understanding performance, particularly in noise, than electric hearing in conventional cochlear implantation alone [e.g., 6].

Although the importance of low-frequency hearing has been widely identified, we still lack the knowledge to account for the mechanism for understanding the LF-based speech (or low-pass filtered speech). Bhargava and Başkent assessed the combined effect of low-pass filtering (i.e., LPF, cut-off frequencies between 500 and 3000 Hz) and periodic interruptions on speech intelligibility, and found that the intelligibility of combined manipulations was lower than each manipulation alone, even in conditions where there was no effect from a single manipulation [7]. Zhang and McPherson showed that low-frequency cuts in hearing aid settings may negatively impact on vowel recognition and Mandarin tone recognition in adverse noise conditions in simulation studies with normal-hearing (NH) listeners [8]. It is well-known that low-frequency regions contain rich acoustic information for speech intelligibility, including vowels (characterized by long duration, formant structure and low frequency), temporal fine-structure (FS), and fundamental frequency (F0) contour (which is important for tonal language understanding). Many studies have examined how these acoustic cues affected speech intelligibility [e.g., 9-11]. For instance, Fogerty and Kewley-Port [12] and Chen et al. [13] showed that the vowel-only sentences (i.e., preserving vowel segments and replacing consonant segments with silence or white noise) were more intelligible than the consonant-only sentences (i.e., vowels replaced by silence or white noise) in both English and Mandarin. Fogerty and Chen compared the perceptual contributions of vowels in Mandarin and English, and found that vowels played a more important role in Mandarin speech recognition than in English [14]. Hilbert transform decomposes a band-passed signal into its envelope (slowly-varying modulations of amplitude in time) and temporal fine-structure (rapid oscillations occurring at a rate close to the center frequency of the band) components. Studies showed that listeners could perfectly understand speech synthesized to only contain Hilbert fine-structure information [15-18]. Chen et al. recently showed that while F0 contour was important for tonal language perception in noise, it was a redundant acoustic cue for sentence understanding in quiet [9].

However, so far most of the above-mentioned studies on the perceptual importance of acoustic cues were performed for wideband speech (i.e., covering the whole frequency range). This motivated the present work to investigate their

http://dx.doi.org/10.21437/Interspeech.2017-2

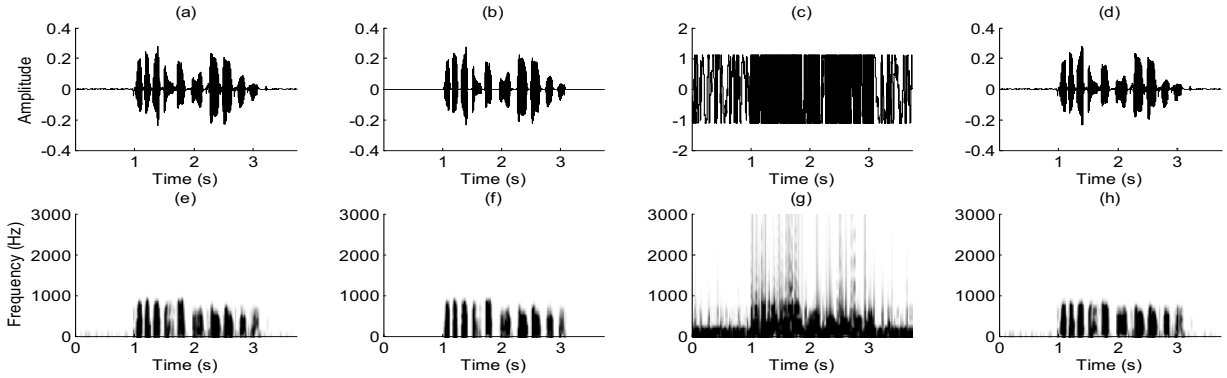

Figure 1. *The waveforms of the (a) LPF-, (b) LPF_vowel-, (c) LPF_FS- and (d) LPF_F0-processed sentences. Panels (e)-(h) show the spectrograms of the (e) LPF-, (f) LPF_vowel-, (g) LPF_FS- and (h) LPF_F0-processed sentences. The low-pass filtering cut-off frequency is 750 Hz.*

contributions to the intelligibility of low-pass filtered speech, which simulates the speech perception of hearing-impaired listeners with low-frequency residual hearing. More specifically, this study generated three types of low-pass filtered stimuli with different types of acoustic cues, i.e., 1) preserving vowels segments and replacing consonant segments with silence, 2) preserving temporal fine-structure and discarding temporal envelope, and 3) containing flattened F0-contour. Listening experiments with NH listeners were conducted to examine the intelligibility of these three types of stimuli and assess the perceptual contributions of vowels, temporal fine-structure and F0-contour to the intelligibility of the LPF-processed sentences

2. Methods

2.1. Subjects and materials

Thirteen NH (i.e., pure-tone thresholds better than 20 dB HL at octave frequencies from 125 to 8000 Hz in both ears) listeners participated in this experiment. All subjects were native-speakers of Mandarin Chinese, and were paid for their participation. The speech material consisted of sentences extracted from the Mandarin Hearing in Noise Test (MHINT) database [19]. MHINT corpus has a total of 24 lists, and each list has 10 sentences (10 keywords in each sentence). All the sentences were spoken by a male native Mandarin-Chinese speaker having a fundamental frequency of 75 to 180 Hz, which was recorded at a sampling rate fs=16 kHz.

2.2. Signal processing

This study synthesized four types of low-pass filtered stimuli, i.e., 1) condition LPF which only applied low-pass filtering processing, 2) condition LPF_vowel which only preserved the vowel segments of the LPF-processed speech, 3) condition LPF_FS only containing the temporal fine-structure of the LPF-processed speech, and 4) condition LPF_F0 where the LPF-processed speech contained flattened F0-contour.

The low-pass filtering in all four signal processing conditions was implemented by using a linear-phase FIR filter with filter order n=10×fs/fcut, where fs is the sampling rate (i.e., 16 kHz) and fcut is the LPF cut-off frequency (500, 750 and 1000 Hz in this study).

1] Vowel-consonant boundaries for vowels and consonants (see [10] for more on vowel and consonant classification) in MHINT sentences were labeled manually by an experienced phonetician, and later verified by another experienced phonetician. The vowel segments of the LPF-processed speech were preserved, while the consonant segments were replaced by silence, yielding the vowel-only LPF-processed speech (i.e., condition LPF_vowel).

2] To generate the LPF_FS-processed stimuli, the temporal fine-structure waveform of the LPF-processed speech was extracted by Hilbert transform and preserved, while the temporal envelope of the LPF-processed speech was discarded, i.e., using a constant value (i.e., 1) to replace the temporal envelope waveform. Note that Hilbert transform was applied to the LPF-processed speech, and the LPF-processed speech was not processed by additional band-decomposition.

3] To generate the LPF_F0-processed stimuli, the dynamic F0-contour of sentence materials were extracted and subsequently replaced by a flattened F0 at the mean value of the utterance. 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 Mar 01 2017]. The F0-contour-flattened (wideband) stimuli were further processed by low-pass filtering to generate the LPF_F0-processed stimuli.

Figure 1 shows the waveforms and spectrograms of a sentence processed by the four signal processing conditions of LPF, LPF_vowel, LPF_FS and LPF_F0.

2.3. Procedure

The experiment was conducted in a sound-proof booth, and test stimuli were played to the participants binaurally through a circumaural headphone at a comfortable listening level. There was a practice session of 40 sample sentences (10 sentences per signal processing condition at LPF cut-off frequency 750 Hz, and different from those used in the testing session) before the experiment session. During the experiment session, participants orally repeated all the keywords they could recognize, and each sentence could be repeated twice. Each participant attended a total of 12 [=4 signal processing conditions (i.e., LPF, LPF_vowel, LPF_FS and LPF_F0) × 3 LPF cut-off frequencies (i.e., 500, 750 and 1000 Hz)] conditions. The test condition order was randomized across subjects, and subjects were given a 5-min break every 30 minutes of testing. The intelligibility score for each condition was computed as the ratio between the number of the correctly

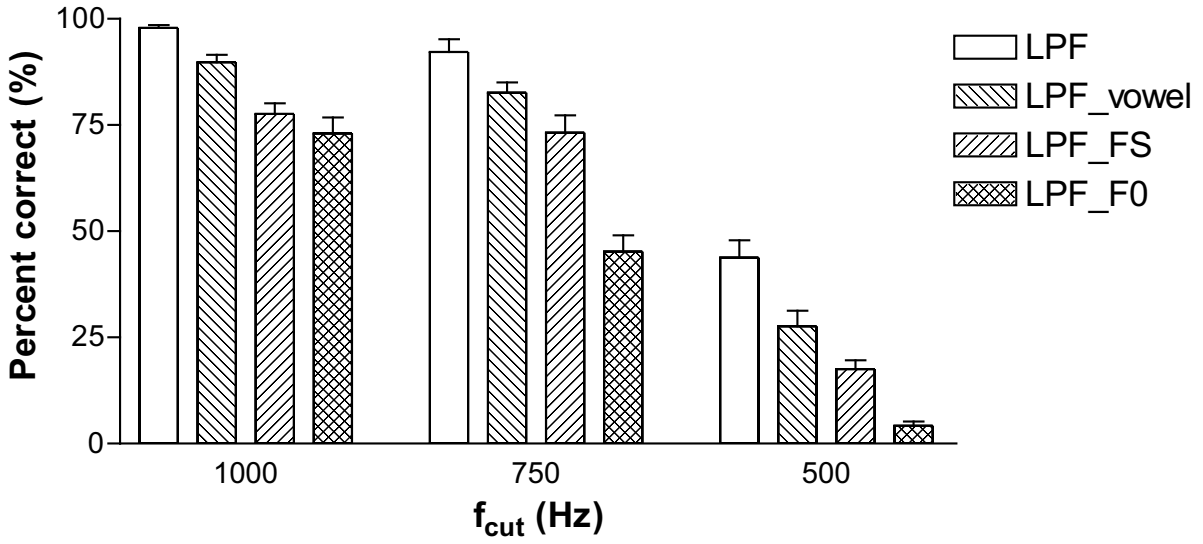

Figure 2. *Mean sentence recognition scores for all conditions. The error bars denote ±1 standard error of the mean.*

recognized words and the total number of words contained in each condition.

3. Results

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 cut-off frequency and signal processing condition as the two within-subject factors. Two-way analysis of variance (ANOVA) with repeated measures indicated a significant effect (F[2, 24]=772.7, p<0.001) of LPF cut-off frequency, signal processing condition (F[3, 36]=122.6.0, p<0.001), and a significant interaction (F[6, 72]=3.8, p<0.01) between LPF cut-off frequency and signal processing condition. Post-hoc pairwise comparisons showed significant differences among all LPF cut-off frequencies (all ps<0.001) and significant differences among the four signal processing conditions (all ps<0.01).

4. Discussion and conclusions

The present work examined the perceptual contributions of three acoustic cues for understanding the LPF-processed speech, i.e., vowels, fine-structure and F0 contour contained in low frequencies. Earlier studies showed that the vowel-only, fine-structure-based (when using fine-structure from the whole band, as this study did) and F0-flattened wideband stimuli all had perfect recognition scores, i.e., almost 100% [e.g., 9-10, 15]. However, when recognizing the three types of acoustically-manipulated low-pass filtered stimuli (i.e., LPF_vowel, LPF_FS and LPF_F0) in this study, listeners showed degraded speech recognition performance relative to the LPF-processed speech (see Fig. 2). For instance, at LPF cut-off frequency 500 Hz, the recognition scores of the LPF_vowel-, LPF_FS- and LPF_F0-processed stimuli were 27.6%, 17.5% and 4.2%, respectively, while the score of the LPF-processed speech was 43.8%. This degraded recognition score may be attributed to the interaction of two manipulations, i.e., acoustic signal processing and low-pass filtering. Hence, although earlier work showed that the three types of acoustically-manipulated stimuli yielded perfect recognition performance, when acoustic cues were further processed in the context of low-pass filtering, the acoustically-manipulated LPF-processed stimuli caused reduced intelligibility scores relative to the LPF-processed speech.

Among the three acoustic cues assessed in this study, it is seen that the LPF_vowel-processed speech led to the largest intelligibility score (see Fig. 2). This indicates the perceptual importance of vowels for understanding the LPF-processed speech, and this finding is consistent with the importance of vowels in recognizing wideband speech reported in earlier studies [e.g., 12-13]. This result is not surprising. As vowels are characterized by their low frequencies and long durations, only preserving vowels in the LPF-processed speech causes a relatively small amount of intelligibility degradation relative to the LPF-processed speech. Hence, the intelligibility difference between the LPF_vowel- and LPF-processed speech is small, e.g., 8.1 and 9.6 percentage points at LPF cut-off frequencies 1000 and 750 Hz, respectively. However, when the LPF cut-off frequency is further reduced to 500 Hz, this difference is increased to 16.2 percentage point.

Many studies have suggested that F0 contour has a minimal effect on sentence intelligibility in quiet [9], or F0 contour could be viewed as a redundant cue for speech perception in quiet. However, the present work showed that F0 contour was not trivial for understanding the LPF-processed speech. For instance, when the LPF cut-off frequency was set to 750 Hz, flattening F0 contour caused a 47.0 percentage point intelligibility reduction relative to condition LPF (i.e., 45.2% vs. 92.2%). Considering that low-pass filtering removes many important cues (e.g., formant structure), F0 contour now plays an important role for understanding low-pass filtered speech, and may not be viewed as a redundant cue any more. This is different to findings in earlier studies evaluating the perceptual contribution of F0 contour in recognizing wideband speech [9, 11].

Figure 1 shows that the envelope of the LPF_F0-processed speech is very similar to that of the LPF-processed speech; however, due to the F0-contour-flattening manipulation, the intelligibility of the LPF_F0-processed speech was significantly lower than that of the LPF-processed speech, i.e., 4.2% vs. 43.8% at fcut=500 Hz in Fig. 2. This indicates that the underlying spectral detail may account for the intelligibility loss of the LPF_F0-processed speech. The present work also examined the performance of understanding LPF_FS-processed stimuli, which primarily contained temporal fine-structure cues. Fine-structure cues (of wideband speech) have been identified as carrying frequency modulation information, and contain much intelligibility information [e.g., 15]. This study further assessed the FS-based intelligibility information contained in the LPF-processed speech, and showed that the fine-structure stimuli synthesized from the LPF-processed speech were also intelligible.

The present work showed that at LPF cut-off frequency 500 Hz, the LPF-processed Mandarin sentences were still quite intelligible (i.e., recognition score 43.8%). This high recognition score with small LPF cut-off frequency suggests that for Mandarin speech recognition, low-frequency residual hearing may provide much intelligibility information. This advantage may be more significant than that from English speech [e.g., 14]. Hence we may foresee a language difference between Mandarin and English when evaluating the frequency-compression based HA speech processing strategies, which primarily use low-frequency residual hearing of hearing-impaired listeners to understand frequency-compressed speech. However, a lot of work is needed to support this performance difference between the two languages, as many factors may influence the performance of understanding the LPF-processed and frequency-compressed speech, e.g., LPF cut-off frequency, compression strategy, and manipulation of different acoustic cues.

Note that the present work has the following limitations. Firstly, the three signal processing conditions (i.e., preserving vowels, preserving temporal fine-structure, and flattening F0 contour) do not result in equal degrees of acoustic distortion as a result of the acoustic manipulation. Therefore, comparisons in this study are not only examining different cues, but different levels of degradation. The vowel conditions may result in the best performance because they also preserve TFS and F0 information and involve removing mainly HF consonant cues that are already removed due to the low-pass filtering [see Fig. 1 (f)]. In contrast, flattening F0 contour removes a cue that is predominantly present in the LF band. Secondly, the acoustic manipulations are not orthogonal, but overlap in the acoustic information provided. For instance, all three conditions preserve TFS cues and also vowel formant cues. Thirdly, some HF artifacts were seen when extracting temporal fine-structure from LPF-processed speech, as shown in Fig. 1 (g). These high frequency artifacts appear to potentially provide some cues to the temporal envelope, which warrants further investigation on their perceptual effects.

In conclusion, the present work investigated three factors affecting the intelligibility of low-pass filtered speech, i.e., vowels, temporal fine structure, and fundamental frequency contour. Listening experiments showed that among the three factors assessed, the vowel-only low-pass filtered speech carried the most intelligibility, and was followed by the fine-structure-based low-pass filtered speech. Flattening F0-contour significantly deteriorated the intelligibility of the LPF-processed speech. These results indicated the importance of the three acoustic cues in recognizing the LPF-processed speech.

5. Acknowledgements

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).

6. References

[1] Hogan, C. A., and Turner, C. W., "High frequency audibility: Benefits for hearing-impaired listeners," J. Acoust. Soc. Am. 104, 432–441, 1998.

[2] Simpson, A., "Frequency-lowering devices for managing high-frequency hearing loss: A review," Trends in Amplification 13, 87–106, 2009.

[3] Kuk, F., Keenan, D., Auriemmo, J., and Korhonen, P., "Re-evaluating the efficacy of frequency transposition," The ASHA Leader 14, 14–17, 2009.

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

[5] Vickers, D. A., Moore, B. C., and Baer, T., "Effects of low-pass filtering on the intelligibility of speech in quiet for people with and without dead regions at high frequencies," J. Acoust. Soc. Am. 110, 1164–1175, 2001.

[6] Gantz, B.J., and Turner, C., "Combining acoustic and electric hearing," Laryngoscope 113, 1726–1730, 2003.

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

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

[9] Chen, F., Wong, L.L.N., and Hu, Y., "Effects of lexical tone contour on Mandarin sentence intelligibility," J. Speech Lang. Hear. Res. 57, 338–345, 2014.

[10] Chen, F., Wong, S.W.K., and Wong, L.L.N., "Effect of spectral degradation to the intelligibility of vowel sentences," in Proc. of 15th Annual Conference of the International Speech Communication Association (InterSpeech), Singapore, pp. 2002–2005, 2014.

[11] Fogerty, D., and Humes, L. E., "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, 2012.

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

[13] Chen, F., Wong, L.L.N., and Wong, Y.W., "Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility," J. Acoust. Soc. Am. 134, EL178–EL184, 2013.

[14] Fogerty, D., and Chen, F. , "Vowel spectral contributions to English and Mandarin sentence intelligibility," in Proc. of 15th Annual Conference of the International Speech Communication Association (InterSpeech), Singapore, pp. 499–503, 2014.

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

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

[17] Heinz, M. G., and Swaminathan, J., "Quantifying envelope and fine-structure coding in auditory nerve responses to chimaeric speech," J. Asso. Res. Otolaryn. 10, 407–423, 2009.

[18] Ghitza, O., "On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception," J. Acoust. Soc. Am. 110, 1628–1640, 2001.

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