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

Phonetic Restoration of Temporally Reversed Speech

Shi-yu Wang, Fei Chen

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

fchen@sustc.edu.cn

Abstract

Early study showed that temporally reversed speech may still be very intelligible. The present work further assessed the role of acoustic cues accounting for the intelligibility of temporally reversed speech. Mandarin sentences were edited to be temporally reversed. Experiment 1 preserved the original consonant segments, and experiment 2 only preserved the temporally reversed fine-structure waveform. Experimental results with normal-hearing listeners showed that for Mandarin speech, listeners could still perfectly understand the temporally reversed speech with a reversion duration up to 50 ms. Preserving original consonant segments did not significantly improve the intelligibility of the temporally reversed speech, suggesting that the reversion processing applied to vowels largely affected the intelligibility of temporally reversed speech. When the local short-time envelope waveform was removed, listeners could still understand stimuli with primarily temporally reversed fine-structure waveform, suggesting the perceptual role of temporally reversed fine-structure to the intelligibility of temporally reversed speech.

Index Terms: speech intelligibility, local time reversion, fine-structure.

1. Introduction

Both temporal envelope and fine-structure cues contain important information for human speech perception [1]. Many early studies have assessed factors affecting the perceptual importance of the envelope cue, including the low-pass cut-off frequency limiting the envelope waveform [2-4], the number of subband to synthesize envelope-based stimuli [2-5], etc. To this end, vocoder simulation is a very effective tool to study factors accounting for the intelligibility of envelope-based stimuli [2, 5-6]. For instance, Shannon et al. showed that with envelope waveforms extracted from up to 4 subbands, normal-hearing (NH) listeners may perfectly understand speech in quiet. Temporal fine-structure waveform may be extracted from a band-limited speech by using the Hilbert transform [2]. Early work showed that temporal fine-structure waveform from in a large band (i.e., large than one ERB band) contained sufficient intelligibility information [7-8], which was attributed to the recovered envelope extracted from temporal fine-structure waveform.

Temporally reversed speech provides a new paradigm to understand the perceptual role of envelope and fine-structure cues. It was shown that temporally reversed speech with a short reversion duration did not significantly deteriorate speech understanding [9]. This suggested that temporally reversed speech still preserved a certain degree of perceptual cues to support speech understanding. Early work regarding the effect of temporal modulation on the intelligibility of phase-based speech also suggested that using a high temporal modulation rate, i.e., sampling phase variation with a short period, may increase the intelligibility of phase-based speech [10].

The main aim of this work is to further examine how local temporal reversion processing affects the amount of intelligibility information contained in temporally reversed speech. This work extended the earlier temporal reversion study in the following two aspects. First, this study assessed the segmental contribution to the intelligibility of temporally reversed speech. More specifically, we investigated the importance of vowels to understand temporally reversed speech. This was motivated by the perceptual role of vowels to speech intelligibility [11-12]. Vowels and consonants differ in many aspects, e.g., temporally and spectrally. Temporally speaking, vowels are longer than consonants, especially for the testing material of Mandarin Chinese [13]. Hence, the effect of local temporal reversion on vowels may account more for the intelligibility of temporally reversed speech.

Second, both temporally reversed envelope and fine-structure cues were contained in temporally reversed speech. While local time-reversion notably modified the temporal waveform and affected speech intelligibility, temporally reversed fine-structure might also contribute to the intelligibility of temporally reversed speech. Hence, this study hypothesizes that temporally reversed fine-structure stimuli may also contain a large amount of intelligibility information.

2. Experiment 1

The purpose of experiment 1 was to assess the intelligibility of temporally reversed Mandarin speech and the perceptual role of vowels to the intelligibility of temporally reversed Mandarin speech.

2.1. Subjects and materials

Eleven (eight male and three female) NH native Mandarin-Chinese listeners participated in the experiment. The subjects' ages ranged from 23 to 35 years, and the majority of subjects were graduate students at Southern University of Science and Technology. All subjects were paid for their participation. The sentence materials were adopted from the Mandarin version of the Hearing in Noise Test (MHINT) [14]. There were 24 lists from the MHINT database, and each list composed of 10 ten-syllable Mandarin sentences. All the sentences were produced by a male speaker, with F0 ranging from 75 to 180 Hz.

2.2. Signal processing

To synthesize temporally reversed speech, the procedure originally proposed in [9] was used in this work. The speech

http://dx.doi.org/10.21437/Interspeech.2017-4

signal was cut into non-overlapping frames, and each frame lasted L ms. Each frame was temporally reversed to generate temporally reversed frame, and finally all temporally reversed frames were concatenated to generate the temporally reversed stimulus. This condition was noted as 'Full', as the local temporal reversion was applied to the full target speech.

This experiment also designed the condition 'Vowel', in which vowel segments were first marked from the original speech. Then the local temporal reversion (with length of L ms) was only applied to each vowel segment, while the intact consonant segments were preserved in the 'Vowel' condition. Vowel-consonant boundaries for vowels and consonants (see [11] for more on vowel and consonant classification) in MHINT sentences were labeled manually by an experienced phonetician, and later verified by another experienced phonetician.

2.3. Procedure

The experiment was conducted in a sound attenuating booth. Stimuli were played through a circumaural headphone binaurally at a comfortable listening level to listeners. Practice (i.e., with feedback) of 40 non-experimental sentences (in conditions L=50 and 70 ms) was given to listeners before the actual testing session. Each listener participated in a total of 12 experimental testing conditions [=6 local temporal reversion time (i.e., L=25, 50, 75, 100, 125, and 150 ms) × 2 segmental conditions (i.e., Full and Vowel], with each containing 2 lists or 20 MHINT sentences. Test order was varied randomly across listeners and no sentence was repeated across conditions. Participants were allowed to listen to each stimulus for a maximum of three times and required to repeat as many words as they could recognize. The intelligibility score for each condition was computed as the ratio between the number of the correctly recognized words and the total number of words contained in 20 MHINT sentences.

2.4. Results

Figure 1 shows the mean recognition scores for all conditions in Experiment 1. Statistical significance was determined by using the percent recognition score as the dependent variable, and the local temporal reversion duration (i.e., L) and segmentation method (i.e., Full and Vowel) as the two within-subject factors. The recognition scores were first converted to rational arcsine units (RAUs) using the rationalized arcsine transform [15]. Two-way analysis of variance (ANOVA) with repeated measures indicated a significant effect [$F(5, 50) = 448.03$, $p < 0.001$] of local temporal reversion duration, a non-significant effect [$F(1, 10) = 3.68$, $p = 0.84$] of segmentation method, and a significant interaction [$F(5, 50) = 3.23$, $p = 0.13$] between the local temporal reversion duration and segmentation method. The significant interaction appears to be due to the ceiling and floor effects of intelligibility scores. Paired comparisons were run between the intelligibility scores at the same local temporal reversion duration, and the statistical significance level was set at $p<0.05$ ($\alpha=0.05$). Results in Fig. 1 showed non-significant difference ($p >0.05$) for all paired comparisons.

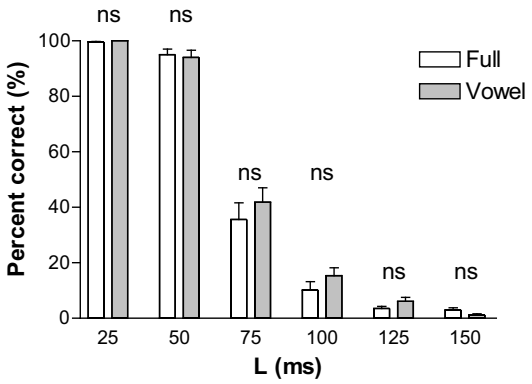

Figure 1. *Mean sentence recognition scores for all conditions in Experiment 1. The error bars denote ± 1 standard errors of the mean. 'ns' denotes that the difference between paired conditions is non-significant (p>0.05).*

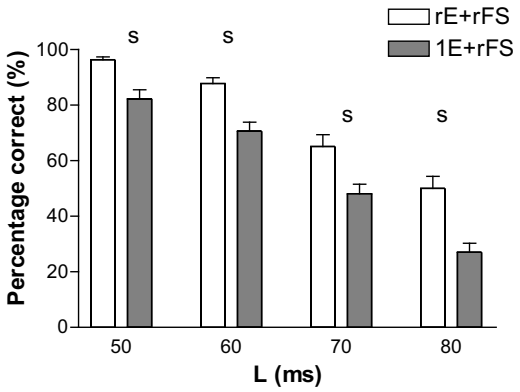

Figure 2. *Mean sentence recognition scores for all conditions in Experiment 2. The error bars denote ± 1 standard errors of the mean. 's' denotes that the difference between paired conditions is significant (p<0.05).*

3. Experiment 2

The purpose of experiment 2 was to assess the perceptual role of temporal fine-structure to the intelligibility of temporally reversed Mandarin speech.

3.1. Subjects and materials

Experiment 2 involved twelve new NH native Mandarin-Chinese listeners (nine males and three females, and ages ranged from 19 to 21 years). The same testing materials (i.e., MHINT sentences) used in Experiment 1 were used in this experiment.

3.2. Signal processing

As in Experiment 1, this experiment also synthesized temporally reversed speech. This condition was noted as condition 'rE+rFS', as for each temporally reversed frame, both temporally reversed envelop and fine-structure cues were preserved. This experiment also designed a condition '1E+rFS', which only preserved the fine structure of each temporally reversed frame, and used a constant amplitude (i.e., 1) to replace the amplitude fluctuation in envelope waveform. The level of the synthesized stimuli in condition '1E+rFS' was adjusted to have the same root-mean-square value as that of the original speech signal.

3.3. Procedure

The test procedure was the same as that used in Experiment 1. Each subject participated in 8 conditions, corresponding to 2 signal processing conditions (i.e., 'rE+rFS' and '1E+rFS') × 4 local temporal reversion time (i.e., L=50, 60, 70, and 80 ms). Twenty MHINT sentences were used per condition, and no sentence was repeated across conditions. The order of test conditions was randomized across subjects.

3.4. Results

Figure 2 shows the mean recognition scores for all conditions in Experiment 2. Statistical significance was determined by using the percent recognition score as the dependent variable, and the local temporal reversion duration (i.e., L) and envelope condition (i.e., rE+rFS and 1E+rFS) as the two within-subject factors. The recognition score scores were first converted to rational arcsine units (RAUs) using the rationalized arcsine transform. Two-way ANOVA with repeated measures indicated a significant effect [$F(3, 33) = 106.90$, $p < 0.001$] of local temporal reversion duration, a significant effect [$F(1, 11) = 120.55$, $p < 0.001$] of envelope condition, and a non-significant interaction [$F(3, 33) = 0.19$, $p = 0.90$] between local temporal reversion duration and envelope condition. Paired comparisons were run between the intelligibility scores at the same local temporal reversion duration, and the statistical significance level was set at $p<0.05$ ($\alpha=0.05$). Results in Fig. 2 showed significant difference ($p < 0.05$) for all paired comparisons.

4. Discussion and conclusions

Early studies showed that when the speech signal was temporally reversed at a short reversion duration, listeners could still perfectly understand temporally reversed speech [9]. From the point of view of amplitude modulation, a short temporal reversion still preserves the low-frequency amplitude fluctuation information. For instance, a temporal reversion duration of 50 ms may preserve the envelope information up to 20 Hz. Hence, listeners may still make use of this envelope information contained below 20 Hz for their speech perception. This result shows the importance of temporal modulation rate to speech intelligibility, and is consistent with established knowledge on the importance of envelope for speech understanding [2, 16]. The present work showed that for Mandarin sentences, when the local temporal reversion duration was set up to 50 ms, listeners could still have an almost perfect intelligibility score. This is consistent with the setting reported in [9], where a perfect understanding was maintained with local temporal reversion duration up to 50 ms.

In addition, this work showed that temporally reversed fine-structure stimulus was also highly intelligible. The perception of fine-structure stimuli (i.e., discarding temporal envelope waveform) has been studied in a number of work [e.g., 7-8]. It has been commonly accepted that fine-structure stimuli, when synthesized with a limited number of bands (e.g., one band in this work), contain sufficient amount of intelligibility information. The underlying mechanism for understanding fine-structure stimuli is attributed to the recovered envelope contained in fine-stimuli stimuli. Hence, it is reasonable to hypothesize that both bandwidth-limited envelope and recovered envelope (from temporal fine-structure) contribute to the intelligibility of temporally reversed speech. Further work is needed to compare the perceptional importance of envelope and fine-structure cues contained in temporally reversed speech.

This study also showed that the main contributing factor to the intelligibility of temporally reversed speech was the reversion processing applied to vowel segments, but not to consonant segments. To some extent, this suggests a vowel importance for speech intelligibility in the context of locally time reversion processing. The importance of vowels over consonants to speech intelligibility has been shown in many studies, including both English and Mandarin Chinese [11-12]. Vowels have more important acoustic cues (e.g., fundamental frequency, formants, harmonic structure) than consonants. Also vowels have a longer duration than consonants. For instance, for Mandarin sentences, vowels may occupy about 66% of the whole sentence duration [11]. Hence, it is not surprising that more consecutive vowel segments would be reversed, causing a more detrimental influence to speech intelligibility.

In conclusion, the present work examined factors affecting the intelligibility of temporally reversed Mandarin sentences, including reversion duration, vowel segment, and temporally reversed fine-structure waveform. Results in this work provided additional knowledge on recognizing temporally reversed speech. It was found that the intelligibility of temporally reversed Mandarin speech was largely affected by the reversion processing applied to vowel segments, as preserving the original consonant segments did not improve the intelligibility of temporally reversed speech. In addition, temporally reversed fine-structure waveform also contained a large amount of intelligibility information, suggesting that the intelligibility of temporally reversed speech could not be fully attributed to the amplitude modulation contained in temporal envelope waveform.

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

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

[3] Xu, L., Thompson, C. S., and Pfingst, B. E., "Relative contributions of spectral and temporal cues for phoneme recognition," J. Acoust. Soc. Am., 117, 3255–3267, 2005.

[4] Rosen, S. "Temporal information in speech: Acoustic, auditory and linguistic aspects," Philos Trans R Soc Lond B Biol Sci, 336, 367–373, 1992.

[5] Dorman, M. F., Loizou, P. C., and Rainey, D., "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, 2043–2411, 1997.

[6] Chen, F. and Loizou, P. C., "Analysis of a simplified normalized covariance measure based on binary weighting functions for predicting the intelligibility of noise-suppressed speech," J. Acoust. Soc. Am., 128, 3715–3723, 2010.

[7] Zeng, F. G., Nie, K. B., Liu, S., Stickney, G., Del Rio, E., Kong, Y. Y., and Chen, H. B., "On the dichotomy in auditory perception between temporal envelope and fine structure cues (L)," J. Acoust. Soc. Am., 116, 1351–1354, 2004.

[8] 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.

[9] Saberi, K. and Perrott, D. R., “Cognitive restoration of reversed speech,” Nature, 398, 760–760, 1999.
[10] Chen, F. and Guan, T., “Effect of temporal modulation rate on the intelligibility of phase-based speech,” J. Acoust. Soc. Am., 134, EL520–EL526, 2013.
[11] Chen, F., Wong, L.L., and Wong, E.Y., “Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility,” J. Acoust. Soc. Am., 134, EL178–EL184, 2013.
[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] Fogerty, D. and Chen, F., “Vowel spectral contributions to English and Mandarin sentence intelligibility,” in Proceedings of 15th Annual Conference of the International Speech Communication Association (InterSpeech), Singapore, 2014, pp. 499–503, 2014.
[14] Wong, L.L., Soli, S., Liu, S., Han, N., and Huang, M., “Development of the Mandarin hearing in noise test (MHINT),” Ear Hear., 28, 70S–74S, 2007.
[15] Studebaker, G. A., “A ‘rationalized’ arcsine transform,” J. Speech Lang. Hear. Res., 28, 455–462, 1985.
[16] Chen, F., and Loizou, P. C. “Predicting the intelligibility of vocoded and wideband Mandarin Chinese,” J. Acoust. Soc. Am. 129, 3281–3290, 2011.