Intelligibility of chimeric locally time-reversed speech

Matsuo, Ikuo
Department of Information Science, Tohoku Gakuin University

Ueda, Kazuo
Department of Human Science, Faculty of Design, Kyusyu University

Nakajima, Yoshitaka
Department of Human Science, Faculty of Design, Kyusyu University

出版情報：Journal of the Acoustical Society of America. 147 (6), pp.EL523-EL528, 2020-06-19. Acoustical Society of America
バージョン：

JASA EXPRESS LETTERS

https://doi.org/10.1121/10.0001414

Intelligibility of chimeric locally time-reversed speech

Ikuo Matsuo,[1,a)] Kazuo Ueda,[2] and Yoshitaka Nakajima[2]

[1]*Department of Information Science, Tohoku Gakuin University, 2-1-1 Tenjinzawa, Izumi-ku, Sendai, 981-3193, Japan*

[2]*Department of Human Science/Research Center for Applied Perceptual Science/Research and Development Center for Five-Sense Devices, Kyusyu University, 4-9-1 Shiobaru Minami-ku, Fukuoka 815-8540, Japan*

matsuo@mail.tohoku-gakuin.ac.jp, ueda@design.kyushu-u.ac.jp, yoshitaka.nakajima@100years.life

Abstract: The intelligibility of chimeric locally time-reversed speech was investigated. Both (1) the boundary frequency between the temporally degraded band and the non-degraded band and (2) the segment duration were varied. Japanese mora accuracy decreased if the width of the degraded band or the segment duration increased. Nevertheless, the chimeric stimuli were more intelligible than the locally time-reversed controls. The results imply that the auditory system can use both temporally degraded speech information and undamaged speech information over different frequency regions in the processing of the speech signal, if the amplitude envelope in the frequency range of 840–1600 Hz was preserved.

[Editor: Douglas D. O'Shaughnessy] Pages: EL523–EL528

Received: 16 February 2020 **Accepted:** 27 May 2020 **Published Online:** 19 June 2020

1. Introduction

Most of the locally time-reversed speech (LTR) speech stimuli used in previous studies [e.g., Ishida *et al.* (2018), Saberi and Perrott (1999), Greenberg and Arai (2004), Stilp *et al.* (2010), Ueda *et al.* (2017), and Ueda *et al.* (2019)] were produced by segmenting original speech stimuli, reversing each segment in time, and concatenating the reversed segments in the original order without any filtering. One recent study (Ueda *et al.*, 2017) showed that the segment duration primarily determined the intelligibility of LTR speech, irrespective of language, if the segment duration was normalized by the speech rates of individual speakers in each language. With segment durations less than 40 ms, a performance ceiling almost always appeared, whereas with segment durations greater than approximately 100 ms, a performance floor became evident. The precipitous decline in intelligibility with segment duration greater than 40 ms is considered to reflect the progressive decline in both the magnitude and phase of the modulation spectrum distributed across the frequency spectrum (Greenberg and Arai, 2004).

In contrast to the majority of stimuli, the first LTR speech (Steffen and Werani, 1994) consisted of two frequency bands: One of the frequency bands contained low-pass-filtered speech with a 300 Hz cutoff to preserve the fundamental frequency component, whereas the other frequency band contained high-pass-filtered speech with the same cutoff, which was subsequently locally time-reversed. They fixed the cutoff frequency and examined only the effect of segment duration on intelligibility. Thus, the effect of cutoff frequency has been largely unknown. It is predictable that intelligibility should go up generally when the cutoff frequency increases (Greenberg and Arai, 2004); however, nobody knows how far the auditory system can integrate, or pick up, speech information in such chimeric stimuli. In addition, Steffen and Werani counted the number of participants who reported correctly for two fixed sentences; these sentences were repeatedly presented from the stimulus with the longest segment duration to the shortest. Thus, the results provided by Steffen and Werani are difficult to be compared with the recent results with more rigorous measurements of intelligibility, although the present authors fully respect their pioneering work. Another issue is that a switched combination—i.e., the low-pass-filtered speech is locally time-reversed, while the high-pass-filtered speech remains unchanged—was not employed in their experiment.

Poeppel and his colleagues (Chait *et al.*, 2015; Giraud and Poeppel, 2012) proposed a multiple time window model of speech perception. The model assumes that a short time window of ~20–30 ms (corresponding to segmental information) and a long time-window of ~200 ms (global cues from syllable-sized units) work together in the auditory system. Ueda *et al.* (2019) proposed that LTR speech is intelligible when the segment duration is shorter than 60 ms, because the global cues provided by the long time-window override the scrambled local cues

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

coming from the short time window. It would be interesting to examine to what extent this hypothesized mechanism works with chimeric locally time-reversed (CLTR) speech.

Thus, the present investigation focuses on the recognition of chimeric speech in which two frequency bands, a temporally degraded band and an intact band, are combined. As a method of temporal degradation, we employed local time reversal, i.e., segmenting speech periodically, reversing each segment, and then joining the reversed segments. We tried to examine the combined effects of the following three variables on intelligibility: segment duration, cutoff frequency, and the frequency band that was locally time-reversed. With these manipulations, we planned to obtain sets of speech stimuli that had the same long-term spectrum but were different in their degree of temporal degradation and the frequency regions in which the degradation occurred. Herein, we report that CLTR speech became less intelligible when either the segment duration or the frequency range of the locally time-reversed portion increased. At the same time, the intelligibility of the CLTR speech always exceeded that of conventional LTR speech, in which the whole frequency range of the original speech stimulus was locally time-reversed without any filtering. The results of this study imply that preserving the amplitude envelope in the frequency range of 840–1600 Hz should be mandatory to keep the chimeric stimuli perfectly intelligible. Moreover, the results are in line with the idea that the amplitude envelope in frequency band between 540 and 1700 Hz is closely related to syllable formation (Nakajima *et al.*, 2017), and speech rhythm formation (Yamashita *et al.*, 2013).

2. Methods

2.1 Participants

Twelve native speakers of Japanese, nine females and three males (age, 20–22 years; median, 21 years) participated in the experiments performed at Tohoku Gakuin University. All of them passed a hearing test at the regular medical checkup. Informed consent was obtained from each participant before they participated in the experiment. The research was conducted with prior approval of the Ethics Committee of Kyushu University. All methods employed in the present study were in accordance with the guidelines published by the Japanese Psychological Association (JPA).

2.2 Stimuli

Eighty sentences in Japanese, spoken by a male and a female speaker, were extracted from the NTT-AT Multilingual Speech Database 2002 (NTT-AT, Kawasaki, Japan; recorded with a 16-kHz sampling rate and 16-bit linear quantization). Japanese is a mora-timed language (Ladefoged and Johnson, 2011). A mora often corresponds to a syllable, but sometimes comprises a part of a syllable. For example, *ojisan* (/o-ji-sa-N/, a middle-aged man or an uncle) has four morae, and *ojīsan* (/o-ji-i-sa-N/, an old man or a grandfather) has five morae in Japanese

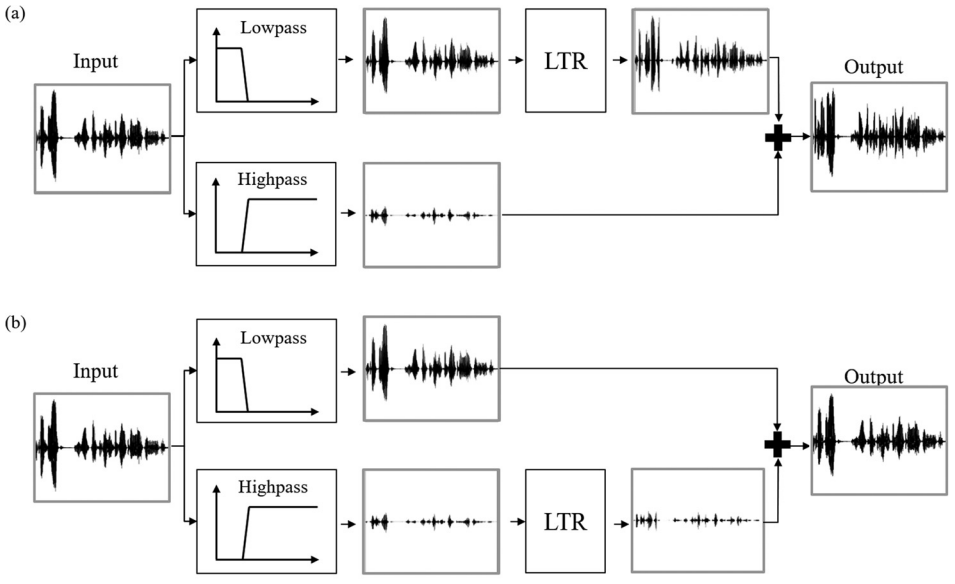

Fig. 1. A schematic diagram of signal processing employed in the present study. LTR: Local time reversal, in which an original speech signal is periodically segmented, each segment is reversed in time, and the reversed segments are then connected in the original order. In (a), only the lower frequency band was locally time-reversed, whereas in (b), only the higher frequency band was locally time-reversed. In both cases, the input signal was passed through a bandpass filter with a passband of 50–7000 Hz before signal processing.

https://doi.org/10.1121/10.0001414

(Ueda *et al.*, 2017). (The overbar indicates a long vowel, in which the length corresponds to two morae in Japanese; the hyphens in the phonemic descriptions indicate mora boundaries.) Figure 1 depicts the signal processing procedure. The speech sentences were bandpass-filtered with a passband of 50–7000 Hz with the slope of 30 dB per 100 Hz. The bandpass-filtered signals were split into two contiguous frequency bands using a set of low-pass and high-pass filters with the same six cutoff frequencies: 570, 840, 1170, 1600, 2150, and 2900 Hz. These cutoff frequencies were spaced at two-Bark intervals (Zwicker, 1961; Zwicker and Terhardt, 1990). Either the low-passed or high-passed signal was locally time-reversed. Three segment duration steps, 70, 95, and 120 ms, including 2.5 ms cosine ramps, were used in the process. These segment durations were chosen because the previous study in which the same speech database was used (Ueda *et al.*, 2017) showed that intelligibility for LTR speech stimuli was around 50% at the segment duration of 70 ms, whereas it reached to the floor performance at around 120 ms. Finally, the locally time-reversed part and its counterpart were combined to produce a chimeric stimulus (CLTR; Fig. 2). CLTR_H denotes the CLTR stimuli in which the higher frequency band was locally time-reversed, whereas CLTR_L denotes the CLTR stimuli in which the lower frequency band was locally time-reversed. In addition, three control conditions (LTR), in which the whole frequency range was locally time-reversed over the three segment duration levels, and one control condition, in which the original speech was presented, were included for each speaker gender. In total, eight control conditions were defined. The signal processing was run with a software written in MATLAB.

To summarize, there were six cutoff frequencies, three segment duration levels, two frequency bands in which local time reversal took place, and two levels for the gender of the speakers, yielding 72 experimental conditions. A total of 80 conditions, including the eight control conditions, were prepared for each participant. Eighty sentences were randomly assigned to the 80 conditions for each participant; no sentence was repeated across different conditions for any one participant.

2.3 Apparatus

The stimuli were passed through a USB audio processor (Onkyo, SE-U55GX, Osaka, Japan), and a headphone amplifier (STAX SRM-323S, Fujimi, Japan). The stimuli were then presented to the participants diotically through headphones (STAX SR-307). The A-weighted sound

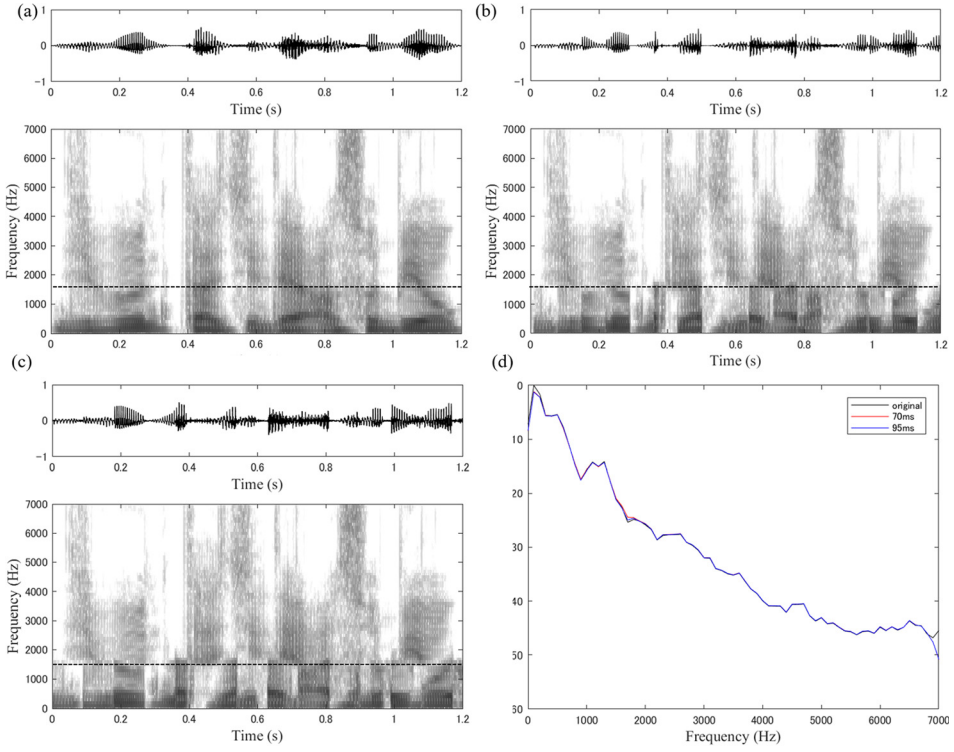

Fig. 2. (Color online) Examples of waveforms (upper panels) and spectrograms (lower panels). (a) an original speech signal, (b) a stimulus in which only the lower frequency band (below 1600 Hz) was locally time-reversed with a segment duration of 70 ms, (c) a stimulus in which only the lower frequency band (below 1600 Hz) was locally time-reversed with a segment duration of 95 ms, and (d) long-term average spectra (LTAS), which were computed from these waveforms.

JASA EXPRESS LETTERS

pressure level of the stimuli was adjusted to 72 dBA, measured with a precision sound level meter (RION, NL-52, Kokubunji, Japan) and a condenser microphone (RION, UC-59) mounted with an artificial ear made according to the standard IEC 60318-1.

2.4 Procedure

Forty stimuli produced by a male and a female speaker were randomly presented to each participant with no break. The stimuli were presented diotically to the participants through the headphones in a quiet room where the background noise level was approximately 30–40 dBA.

The participants initiated a trial by clicking "Enter Key" on a computer screen. Each stimulus was presented three times successively, with interstimulus intervals of 1 s (Ueda *et al.*, 2017). The participants were instructed to write down on a sheet of paper the morae he/she had heard, in Japanese hiragana/katakana script. They were instructed not to guess what they did not recognize. Each listener finished the experiment within 15 min.

3. Results

The percentage for mora accuracy in the original speech condition was 99%. Thus, the experimental method and the baseline performance of the participants as native speakers were validated. Figure 3 shows the percent of correct morae as a function of the boundary frequency between the locally time-reversed band and the intact band. The three panels show the results for segment durations 70, 95, and 120 ms in this order. The horizontal dotted line in each panel shows the mean percentage for mora accuracy for the LTR control stimuli (49%, 11%, and 5% for the 70, 95, and 120-ms segments, respectively). The control performance was comparable to previous results under similar conditions (Ueda *et al.*, 2017). As shown in Fig. 3, the mora accuracy was dropped with frequency range over which the local time reversal was applied at each segment duration. For example, the mora accuracy was decreased from 98.1% to 47.3% and 99.3% to 33.5% at CLTR_H and CLTR_L conditions with 120-ms segment duration, respectively. Conversely, the accuracy was decreased only a little from 99.5% to 88.5% and 100% to

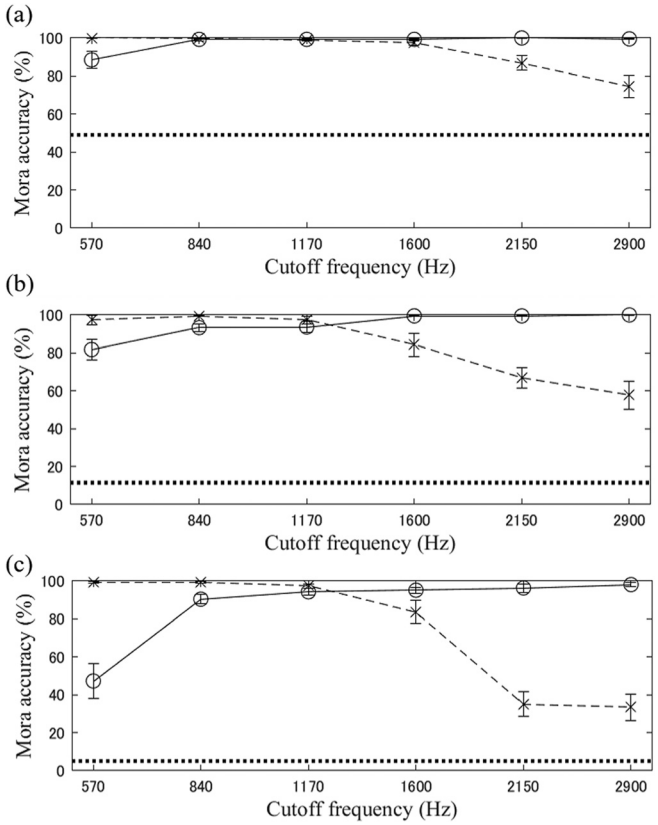

Fig. 3. Mora accuracy as a function of the boundary frequencies between the locally time-reversed band and the intact band. The reversed-speech segment duration was (a) 70 ms, (b) 95 ms, and (c) 120 ms. Circles represent the results in the CLTR_H conditions, and crosses represent the results in the CLTR_L conditions. Horizontal dotted lines show the results in the LTR control condition. Whether the cutoff frequency of degradation went beyond 1600 Hz for the CLTR_L stimuli or went below 840 Hz for the CLTR_H stimuli seemed crucial for maintaining high intelligibility. The error bars represent standard errors of mean.

https://doi.org/10.1121/10.0001414

74.6% at these conditions with 70-ms segment duration, respectively. That is, the segment duration was another crucial variable that affected mora accuracy. The accuracy decreased from 88.5% to 47.3% for the CLTR_H stimuli at the 570 Hz cutoff as the duration increased from 70 to 120 ms; nevertheless, the lowest accuracy, 33% for the CLTR_L stimuli with a 120 ms segment duration and the 2900-Hz cutoff [Fig. 3(c)], was still much higher than the 5% accuracy for the LTR control stimuli. In general, if the cut-off frequency of degradation went beyond 1600 Hz for the CLTR_L stimuli or went below 840 Hz for the CLTR_H stimuli, intelligibility started to decrease crucially.

To avoid the problems associated with the conventional analysis of variance that is performed on arcsine-transformed ratios (Warton and Hui, 2011), we employed beta-binomial regression models that could be applied to correct vs incorrect counts. The number of correct morae in the participants' responses and the total number of morae in the sentence were counted. A beta-binomial regression model was applied to the results for the LTR control and CLTR stimuli for each segment duration. The type of stimuli was the predictor variable (CLTR_H, CLTR_L, LTR). The area under the curve (AUC) was reported as an estimate of the effect size. All statistical effects had a p-level smaller than 0.001. The effects of stimulus type were moderate: for the segment duration of 70 ms, Wald $\chi^2(2) = 66.4$, AUC = 0.78; for 95 ms, Wald $\chi^2(2) = 85.3$, AUC = 0.78; for 120 ms, Wald $\chi^2(2) = 61.8$, AUC = 0.71. Multiple comparisons between the LTR condition and the CLTR_H and CLTR_L conditions using Dunnett's test supported the idea that the accuracies obtained in the CLTR_H and CLTR_L conditions were higher than the accuracies in the LTR conditions.

In the following multiple beta-binomial regression analysis, all statistical effects had a p-level smaller than 0.001, unless reported otherwise. A multiple beta-binomial regression model was used to estimate the effect of introducing chimeric local time reversal in speech. This model comprised segment duration (70, 95, and 120 ms), cutoff frequency (570, 840, 1170, 1600, 2150, and 2900 Hz), stimulus type (CLTR_H and CLTR_L), and the interactions among these factors as predictors. The model showed a large effect size [AUC = 0.88; main effect of segment duration, Wald $\chi^2(1) = 17.8$; main effect of cutoff frequency, Wald $\chi^2(1) = 149.5$; main effect of stimulus type, Wald $\chi^2(1) = 10.4$, $p = 0.001$; segment duration by cutoff frequency interaction, Wald $\chi^2(1) = 4.5$, $p = 0.034$; cutoff frequency by stimulus type interaction, Wald $\chi^2(1) = 129.5$].

4. Discussion

In summary, the mora accuracy of the CLTR stimuli decreased with either an increase in the segment duration or a widening of the frequency range over which the local time reversal was applied. Nevertheless, the mora accuracy for the CLTR stimuli was much higher than the accuracy for the LTR control stimuli. Further, with the present experiment paradigm, it was revealed that the intelligibility for the CLTR_L stimuli started to decrease when the cutoff frequency went beyond 1600 Hz, and that the intelligibility for the CLTR_H stimuli started to decrease when the cutoff frequency went below 840 Hz. The results strongly suggest that preserving the amplitude envelope in the frequency range of 840–1600 Hz should be the key to make chimeric stimuli intelligible.

At the same time, the mora accuracy of the CLTR stimuli was much higher than that of the LTR control stimuli. Our present results show the robustness of speech perception against severe temporal degradation (Fig. 2), while the long-term average spectrum remained unchanged from the original. The results also suggest that the hypothesized multiple time window model applies to the chimeric stimuli.

It is interesting to observe which type of the stimuli, i.e., CLTR_L or CLTR_H, was more severely affected by the temporal degradation employed in the present investigation. Obviously, both the CLTR_L and CLTR_H stimuli were affected equally by the degradation. The pattern of the results looks similar to the patterns observed in some classical filtering experiments for speech, e.g., French and Steinberg (1947), Miller and Nicely (1955), and Studebaker *et al.* (1987). These studies showed two intelligibility curves, i.e., the curves of low-passed and high-passed speech, crossed over in the middle frequency range between 1200 and 1900 Hz (Studebaker *et al.*, 1987). The cross-over points in the present results roughly correspond to those in previous studies that focused on spectral distortion.

The present results should be generalizable to other languages as well. One reason is that it has been shown that the effects of segment duration in LTR speech on intelligibility are quite similar across four languages (English, German, Japanese, and Mandarin Chinese), if the speech rates of speakers are normalized (Ueda *et al.*, 2017). Yet another reason is that our research group (Ueda and Nakajima, 2017) found common spectral factors that divide the frequency range of speech into four frequency bands (50–540, 540–1700, 1700–3300, and above 3300 Hz) across eight languages/dialects: American English, British English, Cantonese, German, French,

Japanese, Mandarin Chinese, and Spanish. It was revealed that the frequency band between 540 and 1700 Hz is closely related to syllable formation (Nakajima *et al.*, 2017), and speech rhythm formation (Yamashita *et al.*, 2013). If the amplitude envelope information in this frequency range is removed, the intelligibility should be decreased crucially. The present results agree with this idea. Once the temporal reversal took place over the frequency range of 540–1700 Hz, intelligibility started to decrease sharply. This should be because the amplitude envelope in the frequency band conveys perceptual cues for *sonority*, which is indispensable for syllable formation (Nakajima *et al.*, 2017). This tendency appeared clearly, however, only when the temporal reversal window was as long as 120 ms, a typical syllable duration both in Japanese and English (Greenberg and Arai, 2004).

Acknowledgments

The authors would like to thank Masumi Mikami and Eri Kanno for running the experiment. This work is supported by JSPS KAKENHI Grants Nos. 25242002, 17H06197, 17K18705, and 19H00630.

References and links

Chait, M., Greenberg, S., Arai, T., Simon, J. Z., and Poeppel, D. (**2015**). "Multi-time resolution analysis of speech: Evidence from psychophysics," Front. Neurosci. **9**, 1–10.

French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.

Giraud, A.-L., and Poeppel, D. (**2012**). "Cortical oscillations and speech processing: Emerging computational principles and operations," Nat. Neurosci. **15**, 511–517.

Greenberg, S., and Arai, T. (**2004**). "What are the essential cues for understanding spoken language?," IEICE Trans. Inf. Syst. **E87-D**, 1059–1070.

Ishida, M., Arai, T., and Kashino, M. (**2018**). "Perceptual restoration of temporally distorted speech in L1 vs. L2: Local time reversal and modulation filtering," Front. Psychol. **9**, 1–16.

Ladefoged, P., and Johnson, K. (**2011**). *A Course in Phonetics*, 6th ed. (Wadsworth, Canada), pp. 251–251.

Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.

Nakajima, Y., Ueda, K., Fujimaru, S., Motomura, H., and Ohsaka, Y. (**2017**). "English phonology and an acoustic language universal," Sci. Rep. **7**(46049), 1–6.

Saberi, K., and Perrott, D. R. (**1999**). "Cognitive restoration of reversed speech," Nature **398**, 760.

Steffen, A., and Werani, A. (**1994**). "Ein Experiment zur Zeitverarbeitung bei der Sprachwahrnehmung" ("An experiment on temporal processing in speech perception"), in *Sprechwissenschaft & Psycholinguistik* (*Speech Science and Psycholinguistics*), edited by G. Kegel, T. Arnhold, K. Dahlmeier, G. Schmid, and B. Tischer (Westdeutscher Verlag, Opladen), pp. 189–205.

Stilp, C. E., Kiefte, M., Alexander, J. M., and Kluender, K. R. (**2010**). "Cochlea-scaled spectral entropy predicts rate-invariant intelligibility of temporally distorted sentences," J. Acoust. Soc. Am. **128**, 2112–2126.

Studebaker, G. A., Pavlovic, C. V., and Sherbecoe, R. L. (**1987**). "A frequency importance function for continuous discourse," J. Acoust. Soc. Am. **81**(4), 1130–1138.

Ueda, K., and Nakajima, Y. (**2017**). "An acoustic key to eight languages/dialects: Factor analyses of critical-band-filtered speech," Sci. Rep. **7**(42468), 1–4.

Ueda, K., Nakajima, Y., Ellermeier, W., and Kattner, F. (**2017**). "Intelligibility of locally time-reversed speech: A multilingual comparison," Sci. Rep. **7**(1782), 1–8.

Ueda, K., Nakajima, Y., Kattner, F., and Ellermeier, W. (**2019**). "Irrelevant speech effects with locally time-reversed speech: Native vs non-native language," J. Acoust. Soc. Am. **145**, 3686–3694.

Warton, D. I., and Hui, F. K. C. (**2011**). "The arcsine is asinine: The analysis of proportions in ecology," Ecology **92**(1), 3–10.

Yamashita, Y., Nakajima, Y., Ueda, K., Shimada, Y., Hirsh, D., Seno, T., and Smith, B. A. (**2013**). "Acoustic analyses of speech sounds and rhythms in Japanese- and English-learning infants," Front. Psychol. **4**(57), 1–10.

Zwicker, E. (**1961**). "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," J. Acoust. Soc. Am. **33**, 248.

Zwicker, E., and Terhardt, E. (**1990**). "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," J. Acoust. Soc. Am. **68**, 1523–1525.