Checkerboard speech vs interrupted speech: Effects of spectrotemporal segmentation on intelligibility

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

Kawakami, Riina
Department of Acoustic Design, Kyushu University

Takeichi, Hiroshige
Computational Engineering Applications Unit, R&D, ISC, RIKEN

出版情報：JASA Express Letters. 1 (7), pp.075204-1-075204-7, 2021-07-21. Acoustical Society of America
バージョン：

JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Checkerboard speech vs interrupted speech: Effects of spectrotemporal segmentation on intelligibility

Kazuo Ueda,[1,a] **Riina Kawakami,**[2,b] **and Hiroshige Takeichi**[3,c]
[1]*Department of Human Science, Faculty of Design/Research Center for Applied Perceptual Science/Research and Development Center for Five-Sense Devices, Kyushu University, 4-9-1 Shiobaru, Minami-ku, Fukuoka 815-8540, Japan*
[2]*Department of Acoustic Design, Kyushu University, 4-9-1 Shiobaru, Minami-ku, Fukuoka 815-8540, Japan*
[3]*Computational Engineering Applications Unit, R&D, ISC, RIKEN, 2-1 Hirosawa, Wako 351-0198, Japan*
ueda@design.kyushu-u.ac.jp, riina0326@gmail.com, takeichi@riken.jp

Abstract: The intelligibility of interrupted speech (interrupted over time) and checkerboard speech (interrupted over time-by-frequency), both of which retained a half of the original speech, was examined. The intelligibility of interrupted speech stimuli decreased as segment duration increased. 20-band checkerboard speech stimuli brought nearly 100% intelligibility irrespective of segment duration, whereas, with 2 and 4 frequency bands, a trough of 35%–40% appeared at the 160-ms segment duration. Mosaic speech stimuli (power was averaged over a time-frequency unit) yielded generally poor intelligibility (⩽10%). The results revealed the limitations of underlying auditory organization for speech cues scattered in a time-frequency domain.

[Editor: Qian-Jie Fu] https://doi.org/10.1121/10.0005600

Received: 20 April 2021 **Accepted:** 23 June 2021 **Published Online:** 21 July 2021

1. Introduction

Speech perception has been investigated by degrading speech signals in time and frequency. Periodic interruption [Fig 1(b)], introduced systematically by Miller and Licklider (1950), has been one of the typical techniques that degrades speech in the time domain [e.g., Powers and Wilcox (1977), Shafiro *et al.* (2016), and Shafiro *et al.* (2018)], along with local time-reversal [e.g., Steffen and Werani (1994), Saberi and Perrott (1999), Matsuo *et al.* (2020), Ueda *et al.* (2017), Ueda *et al.* (2019), Ueda and Ciocca (2021), and Ueda and Matsuo (2021)]. In frequency, many researchers have used filtering [e.g., French and Steinberg (1947), Miller and Nicely (1955), Studebaker *et al.* (1987), Warren *et al.* (2005), and Humes and Kidd (2016)]. The third type of techniques focuses on spectrotemporal modulations (Elliott and Theunissen, 2009; Venezia *et al.*, 2016; Flinker *et al.*, 2019), and thus this type of techniques smears modulations to degrade speech [e.g., ter Keurs *et al.* (1992, 1993), Shannon *et al.* (1995), and Venezia *et al.* (2016)]. A group of techniques in this type employs small units of time and frequency, and transforms the signal properties within each unit. *Pointillistic speech* (Kidd *et al.*, 2009), *mosaic speech* (Nakajima *et al.*, 2018; Santi *et al.*, 2020), and *pixelated speech* (Schlittenlacher *et al.*, 2019) come into this group.

Suppose that the "on" and "off" durations are equal, an interrupted speech stimulus [Fig. 1(b)] consists of a half of the original speech signal [Fig. 1(a)]. In this case, it is well-known that the intelligibility of interrupted speech can be 80% or higher at segment durations shorter than 100 ms for phonetically balanced monosyllabic word lists (Miller and Licklider, 1950), and still exceeds 50% even for an extended segment duration, up to 650 ms, for meaningful sentences (Powers and Wilcox, 1977). The simultaneous onsets and offsets along the frequency axis (Bregman, 1990), or temporal coherence in a spectrum, may provide strong perceptual cues to connect interrupted segments of speech stimuli, yielding such robust speech perception without any filling noise [cf. Ueda and Ciocca (2021)]. On the other hand, it is possible to create a variety of stimulus conditions, in which speech is interrupted in an incoherent manner across frequency, while the proportion of reduction from the original speech is kept constant. We name an extreme case for this type of degraded speech *checkerboard speech* [Fig. 1(c)], after *checkerboard noise* (Howard-Jones and Rosen, 1993). Checkerboard speech stimuli enable us to examine how auditory organization for spectrotemporal segments affects intelligibility; nevertheless, to

[a]Author to whom correspondence should be addressed, ORCID: 0000-0002-1885-0463.
[b]Present address: Rion Co., Ltd., 3-20-41 Higashimotomachi, Kokubunji, Tokyo 185-8533, Japan.
[c]Present address: Open Systems Information Science Team, Advanced Data Science Project (ADSP), RIKEN Information R&D and Strategy Headquarters (R-IH), RIKEN, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan, ORCID: 0000-0002-8545-417X.

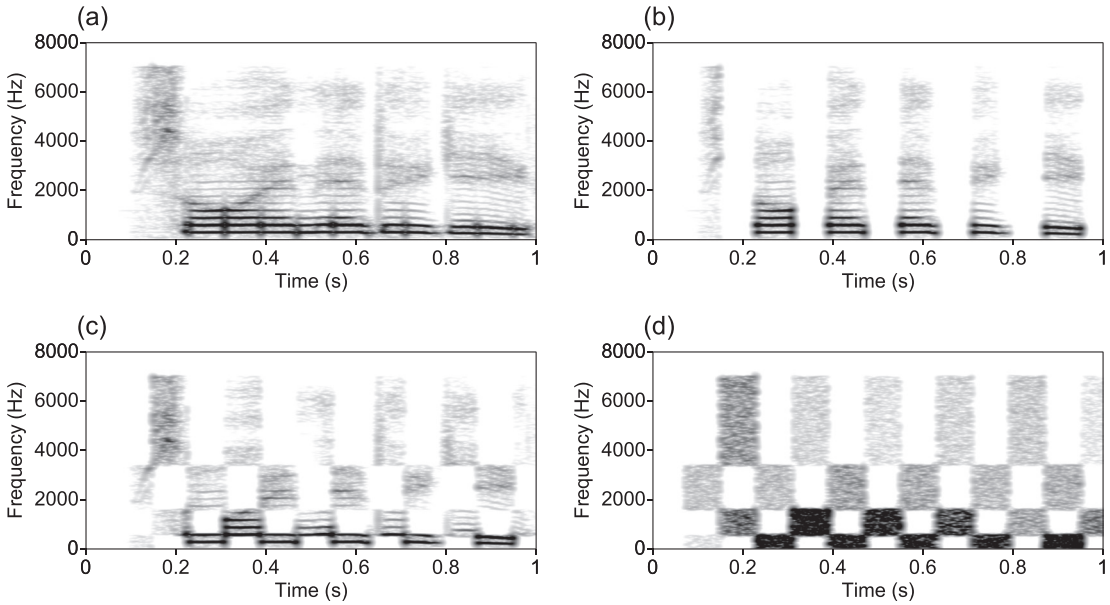

Fig. 1. Examples of narrowband spectrograms for degraded speech, produced from the same fragment of an original spoken sentence by a female talker. (a) Original speech, reconstructed with bandpass filtering into 4 frequency bands (passbands of 50–570, 570–1600, 1600–3400, 3400–7000 Hz) and segmenting with an 80-ms time window including 5-ms root-of-raised-cosine ramps in amplitude, (b) interrupted speech with the 80-ms segment duration, (c) checkerboard speech with the 4 frequency bands and 80-ms segment duration, and (d) checkered mosaic speech with the same time-frequency segmentation.

date, no investigation has attempted this approach. Thus, to compare the intelligibility between interrupted speech stimuli and checkerboard speech stimuli with systematically manipulated segment duration and frequency division, maintaining 50% reduction from the original speech, is the primary purpose for the present investigation.

Our second purpose was to check whether or not periodicity (about 50–500 Hz) and temporal fine structure (about 600–10 000 Hz) in speech (Rosen, 1992) play a crucial role in organizing the perceptual cues scattered in time and frequency. Mosaicking (Nakajima *et al.*, 2018) was employed to remove periodicity and temporal fine structure from interrupted speech and checkerboard speech [Fig. 1(d)]. It is a kind of noise-vocoding (Shannon *et al.*, 1995), in which band noises are modulated with amplitude envelopes of filtered speech, whereas in mosaicking, the amplitude envelopes are stepwise functions, reflecting the average power for each time-frequency unit. Similarly, as in noise-vocoded speech, the periodicity and temporal fine structure in the original speech signal are lost; nevertheless, mosaic speech is almost perfectly intelligible (>95% intelligibility) with the 20 frequency bands (critical bands) and 40-ms segment duration. Extending the segment duration to 80 ms drops the intelligibility to about 45% (Nakajima *et al.*, 2018).

2. Methods

2.1 Listeners

A total of 22 Japanese native listeners (ages 20–25) participated in this study with 2 unpaid listeners participating in experiment 1 and 20 paid listeners participating in experiment 2. None of the listeners participated in both experiments. Their normal hearing within the frequency range of 250–8000 Hz was ensured by the tests with an audiometer (Rion AA-56, Rion, Kokubunji, Japan). The research was conducted with prior approval of the Ethics Committee of Kyushu University (approval ID: 70).

2.2 Stimuli and conditions

A total of 180 (experiment 1) or 150 (experiment 2) Japanese sentences spoken by a female talker were extracted from the "Multilingual Speech Database 2002" (NTT Advanced Technology Corp., Kawasaki, Japan; 16 000-Hz sampling, 16-bit linear quantization). The sentences in the database were based on articles published in newspapers and magazines. The average duration per sentence was 2.5 s ($SD = 0.40$ and 0.39 for experiments 1 and 2, respectively), and the average number of morae (a mora is a syllable-like unit in Japanese) per sentence was 18 ($SD = 2.9$ and 2.8, respectively). The extracted spoken sentences were edited to eliminate unnecessary blanks and noises. The edited speech samples were converted into 44 100-Hz sampling, with 16-bit linear quantization with PRAAT (Boersma and Weenink, 2020) prior to the subsequent processing.

The frequency range from 50 to 7000 Hz was divided into 2, 4, and 20 frequency bands with bandpass filter banks in both experiments 1 and 2. The passbands of the filters for the two frequency band condition were 50–1600 and 1600–7000 Hz, whereas for the four frequency band condition, the passbands were 50–570, 570–1600, 1600–3400, and 3400–7000 Hz. The passbands for the 20 frequency band condition were determined according to the critical bandwidths (see the filter bank B in Ueda and Nakajima, 2017). We adopted the four frequency bands determined by Ueda and Nakajima (2017). The four frequency bands were the basis for the two frequency bands. The resulted cutoff frequency in the middle, i.e., 1600 Hz, was quite similar to the typical crossover frequency of intelligibility curves obtained in highpass and lowpass filtering experiments [e.g., French and Steinberg (1947), Miller and Nicely (1955), and Studebaker *et al.* (1987)]. Each filter was constructed as a concatenated convolution of an upward frequency glide and its temporal reversal (an FIR filter). The frequency characteristics of the filters showed transition regions of 100-Hz wide, with out-of-band attenuations of 50–60 dB.

Filtered signals were segmented at every 20, 80, and 320 ms in experiment 1. In experiment 2, they were segmented at every 20, 40, 80, 160, and 320 ms. All segment durations included 5-ms rise and fall root-of-raised-cosine ramps in amplitude. To produce interrupted speech stimuli, every other segment in a filtered output was replaced with a silent segment of the same length. Whereas, to produce checkerboard speech stimuli, the "on" and "off" phases of interruption were switched across adjacent frequency bands. Then, the segments were concatenated, and summed up across frequency. In experiment 1, mosaicked versions for both interrupted speech and checkerboard speech stimuli were prepared, too. To produce mosaicked stimuli, an average of power in each segment in a bandpass filtered output (a frequency band) was calculated and reflected in an amplitude envelope. Band-pass noise of the same bandwidth as the filtered speech signal was modulated with the amplitude envelope in each frequency band and summed up across frequency.

In summary, in experiment 1, three steps of segment duration (20, 80, and 320 ms), three steps of frequency bands (2, 4, and 20), two types of reduction (interruption and checkerboard), and mosaic processing (non-mosaicking and mosaicking) were combined to yield 36 conditions in total. In experiment 2, five steps of segment duration (20, 40, 80, 160, and 320 ms), three steps of frequency bands (2, 4, and 20), and two types of reduction (interruption and checkerboard) were combined to produce 30 conditions in total. A block of trials consisted of a set of full conditions. Five blocks of trials, in which the order of conditions was randomized in each block, were constructed. Sentences were randomly allotted to one of the trials for individual participants.

2.3 Procedures

The stimuli were presented to participants diotically through headphones (Beyerdynamic DT 990 PRO, Beyerdynamic GmbH, Heilbronn, Germany) in a sound-attenuated booth (Music cabin SC3, Takahashi Kensetsu, Kawasaki, Japan). The headphones were driven with an optical interface (USB interface, Roland UA-4FX, Roland Corp., Shizuoka, Japan) and a headphone amplifier with a built-in D/A converter (Audiotechnica AT-DHA 3000, Audiotechnica, Machida, Japan). The sound pressure level of the original speech was adjusted to about 73 dB (A), using a 1000-Hz calibration tone provided with the speech database. The sound pressure level was measured with an artificial ear (Brüel & Kjær type 4153, Brüel & Kjær Sound & Vibration Measurement A/S, Nærum, Denmark), a condenser microphone (Brüel & Kjær type 4192), and a sound level meter (Brüel & Kjær type 2250).

Participants were instructed to write down exactly what they heard with hiragana or katakana (sets of symbols that are used to represent Japanese morae) without guessing. They were instructed to write down the morae that they could immediately recognize, and not to fill blanks afterwards from the context. Each mora was examined whether it was correct or incorrect. The number of correct morae was counted for each sentence. A blank response was counted as incorrect. Homophone errors were permitted. Percentages of correct morae were calculated for summarizing and displaying the data, whereas statistical analysis was performed on the binomial (correct or incorrect) results.

3. Results

3.1 Experiment 1

The details of the results are provided in supplementary material.[1] Summarizing briefly, the results showed (1) the intelligibility (measured as mora accuracy) for the interrupted speech stimuli decreased monotonically from close to perfect (99%) at the 20-ms segment duration to 55% at the 320-ms segment duration, (2) the intelligibility for the 20-band checkerboard speech stimuli was close or equal to 100%, irrespective of segment duration, (3) the intelligibility for the 2- and 4-band checkerboard speech stimuli varied drastically from more than 94% at the 20-ms segment duration, via less than 46% at the 80-ms segment duration, to about 60% at the 320-ms segment duration, and (4) for the mosaicked speech stimuli, the intelligibility was generally poor (less than 10%) irrespective of the types of reduction (i.e., interrupted or checkered), except for the stimuli with the finest division (with the 20 frequency bands and 20-ms segment duration), which brought more than 85% intelligibility.

3.2 Experiment 2

The percentages of mora accuracy are represented in Fig. 2(a). The results for the interrupted speech stimuli were averaged over the number of frequency bands. The intelligibility for the interrupted speech stimuli went down from almost perfect

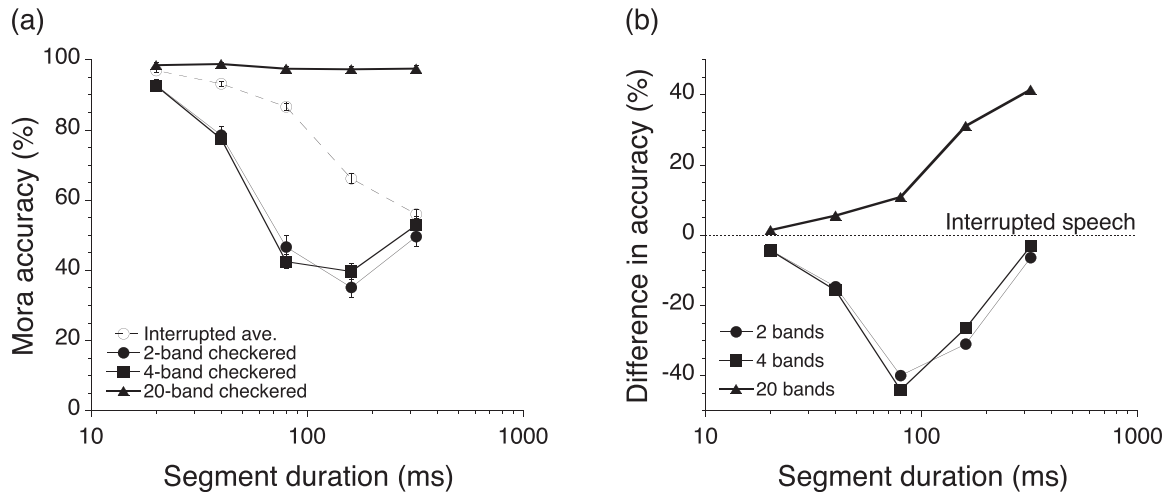

Fig. 2. Results of experiment 2. (a) Mean percentages of mora accuracy ($n = 20$) for the interrupted speech and checkerboard speech stimuli as a function of segment duration and number of frequency band. The data for the interrupted speech stimuli were averaged over the number of frequency bands. (b) Mora accuracy differences between the interrupted speech and checkerboard speech stimuli. The percentages of correct for the checkerboard speech stimuli are represented as deviations from those for the interrupted speech stimuli at corresponding segment duration. Error bars in panel (a) reflect standard error of the mean (SEM).

(97%) at the 20-ms segment duration to a little better than a half point (56%) at the 320-ms segment duration along with the segment duration, as expected from the previous investigations with meaningful speech sentences [e.g., Powers and Wilcox (1977)]. On the other hand, the intelligibility for the 20-band checkerboard speech stimuli stayed at ceiling (97%–99%) irrespective of segment duration, whereas the intelligibility for the 2- and 4-band checkerboard speech stimuli was highest (93%) at the 20-ms segment duration, lowest (35% or 40%) at the 160-ms segment duration, and moderate (50% or 53%) at the 320-ms segment duration.

These observations were supported by the analysis using a generalized linear mixed model (GLMM) with a logistic linking function as implemented in an add-in for JMP (SAS Institute Inc., 2018). The data were analyzed for fixed effects of segment duration (continuous predictor), number of frequency band (continuous predictor), type of reduction (categorical predictor), and their interactions, and for random effects of listener and sentence. This model revealed the p values smaller than 0.05 in all effects and interactions: segment duration ($\beta = -0.007$, SE < 0.001, $t = -25.94$, $p < 0.001$), number of frequency band ($\beta = 0.096$, SE $= 0.005$, $t = 18.92$, $p < 0.001$), type of reduction ($\beta = -0.226$, SE $= 0.029$, $t = -7.75$, $p < 0.001$), number of frequency band × type of reduction ($\beta = 0.105$, SE $= 0.005$, $t = 20.62$, $p < 0.001$), segment duration × number of frequency band ($\beta < 0.001$, SE < 0.001, $t = 2.40$, $p = 0.017$), segment duration × type of reduction ($\beta = 0.002$, SE < 0.001, $t = 8.17$, $p = 0.001$), and segment duration × number of frequency band × type of reduction ($\beta < 0.001$, SE < 0.001, $t = 2.46$, $p = 0.014$).

Figure 2(b) shows the differences in the mora accuracy between the interrupted speech and checkerboard speech stimuli. The average accuracy for the interrupted speech stimuli at each segment duration is represented as the baseline, that is, "0" of the vertical axis. The deviations of the intelligibility for the 2- and 4-band checkerboard speech stimuli, reached the maximum at the 80-ms segment duration (–40% and –44%, respectively). The monotonic increase observed for the 20-band checkerboard speech stimuli essentially reflected the decrement of intelligibility for the interrupted speech stimuli.

4. Discussion

A growing body of literature has revealed the relationship between speech intelligibility and the modulation power spectrum (MPS) (Singh and Theunissen, 2003; Elliott and Theunissen, 2009; Venezia *et al.*, 2016; Sohoglu and Davis, 2020; Flinker *et al.*, 2019). The majority of the previous investigations has identified the area confined to the low spectral (<2 cyc/kHz) and slow temporal (< 10 Hz) modulation rates as the area significantly contributing to intelligibility (Elliott and Theunissen, 2009; Venezia *et al.*, 2016; Sohoglu and Davis, 2020). A signal analysis with the MPS was performed for the current stimuli, followed by an analysis of variance (ANOVA). The analysis results shown in the supplementary material[1] were generally in a good agreement with the previous results.

Both the interrupted speech stimuli and checkerboard speech stimuli used in the present investigation retained a half of each original speech signal; nevertheless, in most of the cases, the intelligibility for these stimuli was not 50% [supplementary Fig. S1(a) and Fig. 2(a)]. With regards to the interrupted speech stimuli, it has been well-known that intelligibility is high, 80% or above, at the segment duration shorter than 100 ms, and then declines gradually as the segment

duration is extended, to about 50% at the segment duration close to or longer than the duration of a word, i.e., about 500 ms (Miller and Licklider, 1950). The present results are in line with this general trend.

By contrast, the intelligibility for the checkerboard speech stimuli was affected by both the number of frequency bands and segment duration, with a wider range of variation (35%–93%) than the range for the interrupted speech stimuli (56%–97%). Furthermore, segment duration was the primary determinant for the intelligibility for the checkerboard speech stimuli with a small number (2 or 4) of frequency bands; on the other hand, the intelligibility for the stimuli with 20 bands was close to 100% irrespective of segment duration. Therefore, the results suggest that speech cue integration across critical bands occurred for the 20-band stimuli irrespective of segment duration, whereas the integration across frequency bands was difficult for the 2- and 4-band stimuli, especially at the 80- and 160-ms segment duration. Thus, the auditory system exhibits limitations in organizing spectrotemporally scattered speech cues, if the number of frequency bands is small. Furthermore, the limitations should be responsible for the intelligibility difference between the interrupted speech stimuli and checkerboard speech stimuli with a small number of frequency bands, which is particularly apparent at the 80-ms segment duration [Fig. 2(b)].

The multiple time window model, proposed by Poeppel and his colleagues (Giraud and Poeppel, 2012; Chait *et al.*, 2015; Teng *et al.*, 2016; Teng and Poeppel, 2020), assumes that two temporal windows, i.e., a short (~20–30 ms) and a long (~200 ms) window, process speech (or non-speech) in parallel. The model successfully explains the almost perfect accuracy found in the current results for non-mosaicked stimuli at the 20-ms segment duration, because both windows work properly for the stimuli with this segment duration. Further investigation is warranted to integrate the effects of the spectrotemporal segmentation into the model.

Shafiro *et al.* (2018) showed that a minimum of word intelligibility appeared at the 250-ms segment duration for interrupted speech stimuli, when the stimuli were lowpass-filtered at 2000 Hz. They claimed that the segment duration at the dip could be estimated with a probability summation for how many words were sampled in each segment (mainly at low frequency of interruption) and how often a word was sampled (mainly at high frequency of interruption): At the minimum point, the proportion of words, i.e., words contained per each segment, is equal to the number of segments per word. The present authors introduced an analogous probability summation, taking a mora as a perceptual unit instead of a word, of which the average duration was about 140 ms (2500/18). The model predicts the segment duration at the minimum intelligibility to be about 100 ms, which points to the bottom of the U-shaped curves observed for the checkerboard speech stimuli with a small number of frequency bands [Fig. 2(a)]. The successful prediction implies that, for spectrally degraded (coarsely segmented) speech stimuli, the segment duration at the minimum intelligibility may be determined by the average duration of the perceptual unit. The conjecture needs further verification in future investigations.

Periodicity and temporal fine structure might be a basis for the sharp contrast observed between the interrupted and checkered stimuli in the non-mosaicked conditions, because the differences in the non-mosaicked conditions vanished away in the mosaicked conditions [experiment 1; supplementary Fig. S1(b)]. Periodicity and temporal fine structure may be crucial in organizing the scattered or interrupted perceptual cues in the interrupted and checkered speech stimuli, except for the stimuli with the finest division. It is indeed conceivable that continuity among interrupted segments in degraded speech affects intelligibility. This idea is supported by the drastic reduction in intelligibility (close to 60%) observed in the interrupted locally time-reversed speech for just 40-ms silent intervals and the recovery of intelligibility (about 30%) by filling the intervals with 10-dB noise (Ueda and Ciocca, 2021).

Our findings in the 20-band checkerboard speech stimuli might be considered to be comparable to the results obtained with the 8- and 16-band checkerboard masking noise of the 50-ms segment duration, made from pink noise, by Howard-Jones and Rosen (1993); they observed that the checkerboard noise had the same level of a masking effect as continuous pink noise. This seems to be comparable to the current results, in which the 20-band checkerboard speech stimuli always produced almost perfect intelligibility like original speech irrespective of segment duration. However, they found a nearly 10-dB difference in masked thresholds between their logarithmically scaled 2- and 4-band conditions, which can be contrastive to the current results. It seems likely that further investigations are warranted to fill the gap between the two studies.

In conclusion, the newly proposed experimental paradigm employing interrupted, checkered, and mosaicked speech stimuli looks promising in further exploring the mechanisms of spectrotemporal processing in speech perception.

Acknowledgments

This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI Grant No. JP19H00630. The authors would like to thank Yoshitaka Nakajima for providing J language software routines; Hikaru Eguchi for programming in LIVECODE; Takako Mitsudo and Alexandra Wolf for providing helpful suggestions to improve the early draft; Yoshitaka Nakajima, Gerard B. Remijn, Valter Ciocca, and Wolfgang Ellermeier for valuable discussion; and Josef Schlittenlacher and Valeriy Shafiro for providing insightful suggestions to improve the manuscript.

References and links

[1]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0005600 for the details of the results in Experiment 1 and the MPS analysis results of the stimuli.

Boersma, P., and Weenink, D. (**2020**). "Praat: Doing phonetics by computer, version 6.0.21 [computer program]," http://www.praat.org/ (Last viewed 27 February 2021).
Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).
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**(214), 1–10.
Elliott, T. M., and Theunissen, F. E. (**2009**). "The modulation transfer function for speech intelligibility," PLoS Comput. Biol. **5**(3), e1000302.
Flinker, A., Doyle, W. K., Mehta, A. D., Devinsky, O., and Poeppel, D. (**2019**). "Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries," Nat. Hum. Behav **3**(4), 393–405.
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**(4), 511–517.
Howard-Jones, P. A., and Rosen, S. (**1993**). "Uncomodulated glimpsing in 'checkerboard' noise," J. Acoust. Soc. Am. **93**(5), 2915–2922.
Humes, L. E., and Kidd, G. R. (**2016**). "Speech recognition for multiple bands: Implications for the speech intelligibility index," J. Acoust. Soc. Am. **140**(3), 2019–2026.
Kidd, G., Streeter, T. M., Ihlefeld, A., Maddox, R. K., and Mason, C. R. (**2009**). "The intelligibility of pointillistic speech," J. Acoust. Soc. Am. **126**(6), EL196–EL201.
Matsuo, I., Ueda, K., and Nakajima, Y. (**2020**). "Intelligibility of chimeric locally time-reversed speech," J. Acoust. Soc. Am. **147**(6), EL523–EL528.
Miller, G. A., and Licklider, J. C. R. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**(2), 167–173.
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., Matsuda, M., Ueda, K., and Remijn, G. B. (**2018**). "Temporal resolution needed for auditory communication: Measurement with mosaic speech," Front. Hum. Neurosci. **12**(149), 1–8.
Powers, G. L., and Wilcox, J. C. (**1977**). "Intelligibility of temporally interrupted speech with and without intervening noise," J. Acoust. Soc. Am. **61**(1), 195–199.
Rosen, S. (**1992**). "Temporal information in speech: Acoustic, auditory and linguistic aspects," Philos. Trans. R. Soc. London B **336**(1278), 367–373.
Saberi, K., and Perrott, D. R. (**1999**). "Cognitive restoration of reversed speech," Nature **398**(6730), 760–760.
Santi, Nakajima, Y., Ueda, K., and Remijn, G. B. (**2020**). "Intelligibility of English mosaic speech: Comparison between native and non-native speakers of English," Appl. Sci. **10**(19), 6920.
SAS Institute Inc. (**2018**). "JMP, version 14.3.0 [computer program]."
Schlittenlacher, J., Staab, K., Çelebi, Ö., Samel, A., and Ellermeier, W. (**2019**). "Determinants of the irrelevant speech effect: Changes in spectrum and envelope," J. Acoust. Soc. Am. **145**(6), 3625–3632.
Shafiro, V., Fogerty, D., Smith, K., and Sheft, S. (**2018**). "Perceptual organization of interrupted speech and text," J. Speech Lang. Hear. Res. **61**(10), 2578–2588.
Shafiro, V., Sheft, S., and Risley, R. (**2016**). "The intelligibility of interrupted and temporally altered speech: Effects of context, age, and hearing loss," J. Acoust. Soc. Am. **139**, 455–465.
Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Singh, N. C., and Theunissen, F. E. (**2003**). "Modulation spectra of natural sounds and ethological theories of auditory processing," J. Acoust. Soc. Am. **114**(6), 3394–3411.
Sohoglu, E., and Davis, M. H. (**2020**). "Rapid computations of spectrotemporal prediction error support perception of degraded speech," eLife **9**, 1–25.
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), Vol. 6, pp. 189–205.
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.
Teng, X., and Poeppel, D. (**2020**). "Theta and gamma bands encode acoustic dynamics over wide-ranging timescales," Cerebral Cortex **30**(4), 2600–2614.
Teng, X., Tian, X., and Poeppel, D. (**2016**). "Testing multi-scale processing in the auditory system," Sci. Rep. **6**(34390), 1–13.
ter Keurs, M., Festen, J. M., and Plomp, R. (**1992**). "Effect of spectral envelope smearing on speech reception. I," J. Acoust. Soc. Am. **91**, 2872–2880.
ter Keurs, M., Festen, J. M., and Plomp, R. (**1993**). "Effect of spectral envelope smearing on speech reception. II," J. Acoust. Soc. Am. **93**, 1547–1552.
Ueda, K., and Ciocca, V. (**2021**). "Phonemic restoration of interrupted locally time-reversed speech: Effects of segment duration and noise levels," Att. Percept. Psychophys. **83**(5), 1928–1934.
Ueda, K., and Matsuo, I. (**2021**). "Intelligibility of chimeric locally time-reversed speech: Relative contribution of four frequency bands," JASA Express Lett. **1**(6), 065201.
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**(6), 3686–3694.

Venezia, J. H., Hickok, G., and Richards, V. M. (**2016**). "Auditory 'bubbles': Efficient classification of the spectrotemporal modulations essential for speech intelligibility," J. Acoust. Soc. Am. **140**(2), 1072–1088.

Warren, R. M., Bashford, J. A. J., and Lenz, P. W. (**2005**). "Intelligibilities of 1-octave rectangular bands spanning the speech spectrum when heard separately and paired," J. Acoust. Soc. Am. **118**(5), 3261–3266.