Irrelevant sound effects with locally time-reversed speech: Native vs. non-native language

Ueda, Kazuo
Unit of Perceptual Psychology, Department of Human Science, Kyushu University

Nakajima, Yoshitaka
Unit of Perceptual Psychology, Department of Human Science, Kyushu University

Kattner, Florian
Institut für Psychologie, Technische Universität Darmstadt

Ellermeier, Wolfgang
Institut für Psychologie, Technische Universität Darmstadt

出版情報：Journal of the Acoustical Society of America. 145 (6), pp.3686-3694, 2019-06-25. Acoustical Society of America
バージョン：

Irrelevant speech effects with locally time-reversed speech: Native vs non-native language[a)]

Kazuo Ueda,[1,b)] Yoshitaka Nakajima,[1] Florian Kattner,[2] and Wolfgang Ellermeier[2]
[1]*Unit of Perceptual Psychology, Department of Human Science/Research Center for Applied Perceptual Science, Kyushu University, 4-9-1 Shiobaru, Minami-ku, Fukuoka 815-8540, Japan*
[2]*Institut für Psychologie, Technische Universität Darmstadt, Alexanderstraße 10, D-64283 Darmstadt, Germany*

(Received 16 October 2018; revised 30 May 2019; accepted 3 June 2019; published online 25 June 2019)

Irrelevant speech is known to interfere with short-term memory of visually presented items. Here, this irrelevant speech effect was studied with a factorial combination of three variables: the participants' native language, the language the irrelevant speech was derived from, and the playback direction of the irrelevant speech. We used locally time-reversed speech as well to disentangle the contributions of local and global integrity. German and Japanese speech was presented to German ($n = 79$) and Japanese ($n = 81$) participants while participants were performing a serial-recall task. In both groups, any kind of irrelevant speech impaired recall accuracy as compared to a pink-noise control condition. When the participants' native language was presented, normal speech and locally time-reversed speech with short segment duration, preserving intelligibility, was the most disruptive. Locally time-reversed speech with longer segment durations and normal or locally time-reversed speech played entirely backward, both lacking intelligibility, was less disruptive. When the unfamiliar, incomprehensible signal was presented as irrelevant speech, no significant difference was found between locally time-reversed speech and its globally inverted version, suggesting that the effect of global inversion depends on the familiarity of the language.
 https://doi.org/10.1121/1.5112774

[VMR] Pages: 3686–3694

I. INTRODUCTION

Overhearing irrelevant background speech is known to interfere with the maintenance of information in short-term memory, i.e., the *irrelevant speech effect*, ISE (Banbury *et al.*, 2001; Ellermeier and Zimmer, 2014). It has been shown that ISEs of comparable magnitude can be observed with both forward and reversed speech (Jones *et al.*, 1990; Röer *et al.*, 2014; Röer *et al.*, 2017; Surprenant *et al.*, 2007). However, previous investigations do not completely agree on whether forward speech and reversed speech are equipotent in producing ISEs. Therefore, to replicate and extend the earlier findings, the present investigation addressed this issue by factorially combining the forward/backward manipulation with a familiar vs unfamiliar language from which the irrelevant speech material was derived, and we did so in two subject populations (native German vs native Japanese speakers) for whom the respective other language was totally incomprehensible. Besides, we employed locally time-reversed speech—speech that is cut into short segments, reversed in time, and concatenated in the original order—and its global reversal, i.e., playing backward the entire locally time-reversed speech, to see the effects of integrity of local and global features of speech on magnitude of the ISE. The experiment was run with both German and Japanese participants. The stimuli, both in German and Japanese, and experimental procedure were precisely matched in two sites, one in Germany and the other in Japan. Substantial ISEs were observed in all conditions employing speech, confirming that semantics is not a crucial factor in ISEs. However, while the effect of global playing direction in locally time-reversed speech clearly appeared in native speakers of the language, the effect was nonsignificant in participants who did not understand the respective language, suggesting that the effect is related to linguistic comprehensibility, i.e., the language in question is perfectly understood under optimal (undegraded) listening conditions.

Colle and Welsh (1976) showed that incomprehensible foreign speech interferes with serial recall performance of visually presented letters. Then, phonological similarity between visually presented items and auditory backgrounds, rather than semantic similarity (cf. Marsh and Jones, 2010), was brought into focus (Baddeley, 1986; Salamé and

[a)]Portions of this work were presented in "Irrelevant sound effects with locally time-reversed speech" and "Irrelevant sound effects with locally time-reversed speech: Real performance difference between German and Japanese native speakers?," Fechner Day 2017: the 33rd Annual Meeting of the International Society for Psychophysics, Fukuoka, Japan, October 2017; "Irrelevant sound effects with locally time-reversed speech: Effects of native language," Spring Meeting of the Acoustical Society of Japan, Saitama, Japan, March 2018; "Irrelevant speech effects with locally time-reversed speech: Language familiarity," the 41st Perceptual Frontier Seminar, Fukuoka, Japan, August 2018; "Irrelevant sound effects with locally time-reversed speech: Speech reversal and language familiarity," the 82nd Annual Convention of the Japanese Psychological Association, Sendai, Japan, September 2018; and "Irrelevant speech effects with locally time-reversed speech: Native vs non-native language," the 176th Meeting of the Acoustical Society of America and 2018 Acoustics Week in Canada in Victoria, BC, Canada, November 2018.

[b)]Electronic mail: ueda@design.kyushu-u.ac.jp

Baddeley, 1982), supporting a working-memory model involving a phonological loop (Baddeley, 1986), because the model—retaining short-term memory items in a loop after converting visual items into auditory items—seemed to explain the ISE with (foreign) speech. However, the effect of phonological similarity between the irrelevant speech and the to-be-remembered items found by Salamé and Baddeley (1982) was not replicated (Jones and Macken, 1995; Larsen *et al.*, 2000; LeCompte and Shaibe, 1997). Many other empirical findings also go against the model (see Caplan *et al.*, 2012, for review). Several alternative models or theories for the ISE have been proposed, e.g., accounts related to perceptual organization or *auditory scene analysis* (Bregman, 1990) such as the O-OER model with changing-state (Jones, 1993), perceptual-motor interference (Hughes and Marsh, 2017), and interference-by-process (Marsh *et al.*, 2018), or other accounts with temporal distinctiveness theory (LeCompte *et al.*, 1997), an extended feature model (Neath, 2000), attention (Bell *et al.*, 2019; Lange, 2005), and psychoacoustical or acoustical parameters (Schlittmeier *et al.*, 2012; Senan *et al.*, 2018a,b), seeking a better account of a wide range of findings.

The ISE itself is a robust phenomenon: Speech is always the most disruptive background sound, compared to non-speech sounds. Even with severely degraded speech, e.g., noise-vocoded speech, which conveys only amplitude envelope patterns in several frequency bands (Shannon *et al.*, 1995), ISEs of substantial magnitude have been obtained, provided sufficient spectral resolution is preserved (Dorsi, 2013; Ellermeier *et al.*, 2015; Senan *et al.*, 2018a,b). In this case, temporal changes in several frequency bands are necessary to produce the ISE. Therefore, both the temporal and spectral integrity of the distractor have to be considered to account for the particularly disruptive nature of irrelevant speech.

Temporal reversal (Kellogg, 1939; Meyer-Eppler, 1950), i.e., playing long utterances from end to beginning, is a traditional technique to make natural speech unintelligible without changing the spectrum. This procedure maintains the overall spectrum and the kinds of spectral changes occurring, but destroys linguistic meaning completely (Licklider and Miller, 1951). To show the semantics of the irrelevant stream are not crucial to obtain the ISE, some researchers have used such *reversed speech* (Jones *et al.*, 1990; LeCompte *et al.*, 1997; Röer *et al.*, 2014, 2017; Surprenant *et al.*, 2007). Typically, ISEs of comparable magnitude as with forward speech were obtained with reversed speech. We found two exceptions in the literature: a marginally significant difference in ISE ($p = 0.08$ with $n = 79$) was found between forward and reversed words in LeCompte *et al.* (1997), and no significant difference between a white-noise control condition and a reversed speech condition ($p = 0.196$ with $n = 29$) was observed in Viswanathan *et al.* (2014). However, the difference, if any, was very small (2%) in LeCompte *et al.;* in Viswanathan *et al.*, the evidence is indirect and inconclusive because no direct comparison between forward and reversed speech was included in their experiments. To clarify the issue, we performed an experiment with two comparable sizes of participant groups of vastly different native languages, i.e., German and Japanese, with a symmetrical experimental design in receiving both one's native and an incomprehensible language as irrelevant speech, and with both forward and reversed speech being presented to each group of participants.

Moreover, rather than playing an entire speech utterance backwards, the present investigation focuses on *locally time-reversed speech*, i.e., temporally inverting successive segments (20–120 ms long) of a speech utterance. Essentially, this manipulation was introduced by Steffen and Werani (1994) to investigate the temporal resolution required for speech perception. Locally time-reversed speech is constructed by first cutting a speech utterance into short segments and subsequently reversing each segment in time. It is called *locally* time-reversed speech because the reversal takes place only within each segment, rather than reversing the entire utterance. By varying segment duration, the intelligibility of locally time-reversed speech can be manipulated systematically. Research from our laboratories (Ueda *et al.*, 2017) has shown that the intelligibility of locally time-reversed speech gradually decreased from perfectly intelligible (at 20 ms segment duration) to unintelligible (>100 ms), in very similar ways for four different languages (English, German, Mandarin Chinese, and Japanese).

Since the segmentation and reversal to obtain locally time-reversed speech is performed with no regard to phonemic cues in original speech [Fig. 1(a)], these cues are in most cases split up at some haphazard position and displaced in time [Fig. 1(b); cf. Nakajima *et al.*, 2018; Ueda *et al.*, 2017)]. The results of previous perceptual experiments (Greenberg and Arai, 2001; Ishida *et al.*, 2018; Kiss *et al.*, 2008; Meunier *et al.*, 2002; Nakajima *et al.*, 2018; Remez *et al.*, 2013; Saberi and Perrott, 1999; Steffen and Werani, 1994; Stilp *et al.*, 2010; Ueda *et al.*, 2017) indicate that the auditory system is capable of overriding this kind of degradation and retrieving plausible solutions to some extent, unless the reversed segment duration becomes too long. The restoration process, however, should certainly impose an extra processing load on the auditory system compared to the processing of normal speech. Thus, we may hypothesize that the process of reconstructing locally time-reversed speech uses *global* (across-segment) cues to seek plausible solutions despite the scrambled *local* (within-segment) cues. This process should run automatically because, for the auditory system, there is no way to know how the stimulus was made beforehand. This is consistent with the idea that two time windows of different time scales (~20–30 ms and ~200 ms) exist in our brain and that the two processes, i.e., phonemic and syllabic processes, run in parallel to some extent, which is supported by several works done by Poeppel and colleagues (Chait *et al.*, 2015; Giraud and Poeppel, 2012; Hickok and Poeppel, 2007; Poeppel, 2003; Sanders and Poeppel, 2007; Teng *et al.*, 2016). The theory implies that perception of locally time-reversed speech is possible when the segment duration is shorter than ~30 ms because the syllabic process can still work with integrating information dispersed over more than three segments and phonemes can be deduced from the syllabic information. Retrieval becomes difficult with longer segment duration.

Here we examine whether the global (across-segment) cues take precedence over the local (within-segment) cues in speech processing, or vice versa, and how these cues affect the size of the ISE. To be more specific, when locally time-reversed speech is played forward, global integrity is maintained, but degraded gradually as the segment duration becomes longer, whereas when locally time-reversed speech is played backward, each segment shows a fractional series of normal speech cues, and local integrity increases as the segment duration becomes longer.

To investigate whether the integrity of local or global cues in task-irrelevant speech might affect the magnitude of the ISE, we used (a) natural speech recordings played forward (abbreviated as F), (b) locally time-reversed speech (LTR-F), (c) reversed playback of the entire locally time-reversed sequence (LTR-R)—thereby concatenating recovered forward segments in an unnatural (reversed) order, and (d) reversed speech (R) as irrelevant speech distractors during a serial recall task (see Fig. 1 and Table I for depictions of these stimulus manipulations). Further, participants were tested with either their native language or a foreign language they did not understand. Specifically, sentences spoken in German and Japanese, taken from the same speech database, were used to test German-speaking participants and Japanese-speaking participants. This configuration allows us to check the possibility as to whether mastering or comprehending the background language affects the ISE. If comprehensibility played a role, our native-speakers would be more affected by manipulations of material derived from their own language than the non-native speakers to whom the language is incomprehensible.

About the contrast in the integrity of local and global cues, we are considering two alternative hypotheses: hypothesis H_L, which states that an ISE produced by locally time-reversed speech is primarily caused by the integrity of local (within-segment) cues, and H_G, which states that an ISE of locally time-reversed speech is primarily caused by the integrity of global (between-segment) cues, i.e., global order of cues in the spectro-temporal pattern. The following predictions are made regarding the experimental results. Under the H_L, LTR-R should exhibit a greater ISE than LTR-F, because the integrity of local speech cues is reinstated by global reversal. By the same token, in LTR-R conditions, increasing segment duration should increase the magnitude of the ISE. In contrast, under the H_G, LTR-R should produce smaller ISEs than LTR-F, since in the former, the global spectro-temporal pattern is degraded. Likewise, in LTR-F conditions, increasing segment duration should reduce ISEs.

II. METHOD

The experiment was performed in two laboratories simultaneously: at Technische Universität Darmstadt in Germany, and at Kyushu University in Fukuoka, Japan.

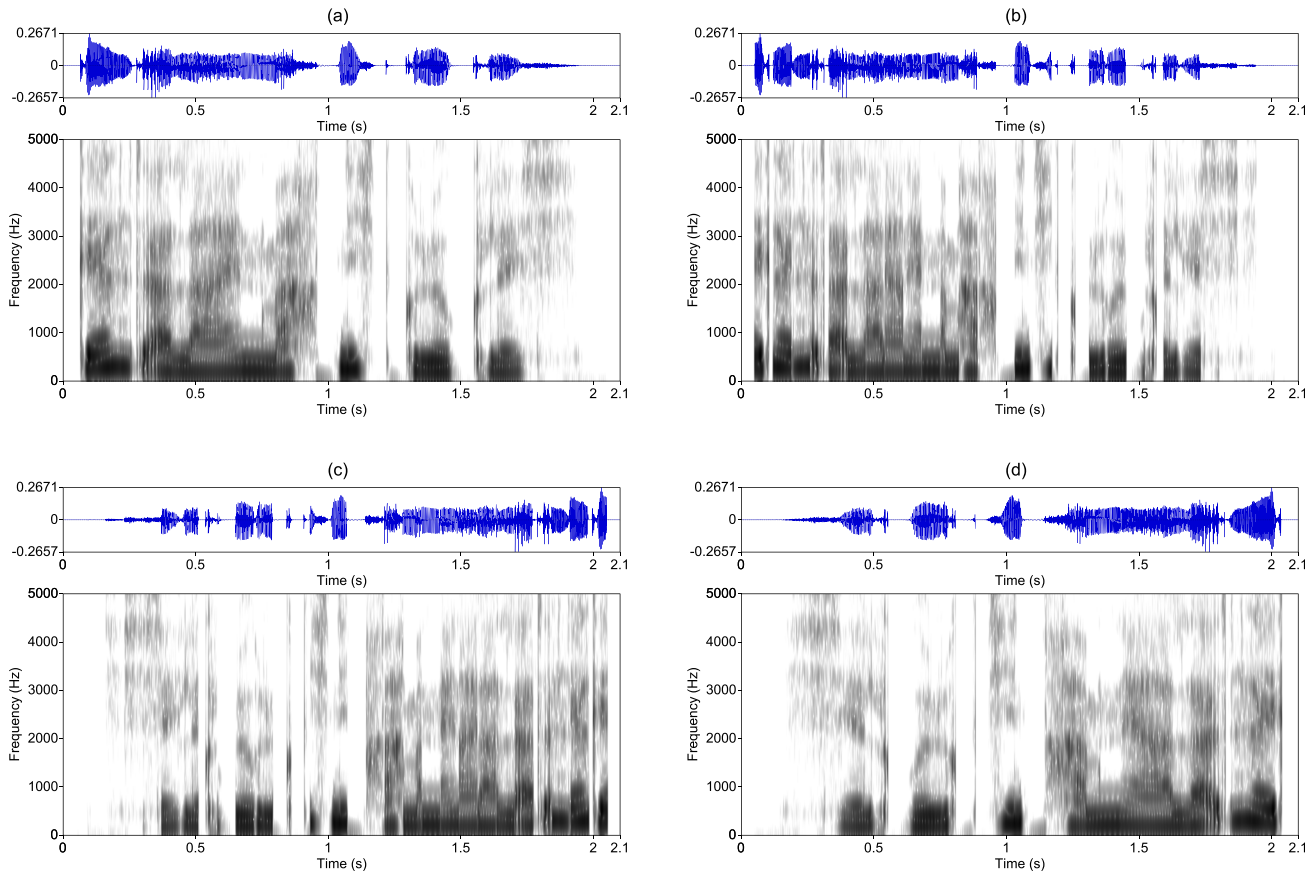

FIG. 1. (Color online) Examples of auditory stimuli represented as waveforms and spectrograms. (a) An original spoken sentence played forward, F, in German by a female talker, saying "Geld allein macht nicht glücklich." (Money alone does not make you happy.) (b) For a locally time-reversed version, LTR-F, the original speech is divided into segments of fixed duration (in this case, 70 ms) and each segment is played backwards. (c) LTR-F was globally inverted to generate LTR-R. (d) A temporally inverted version of the original, i.e., reversed speech, R. The original speech sample was extracted from the NTT-AT, Multi-lingual speech database 2002. The figure was drawn with Praat, developed by Boersma and Weenink (2016).

TABLE I. Types of task-irrelevant sound presented during serial recall. *Local integrity* refers to integrity within segments, whereas *global integrity* refers to integrity across segments. Intelligibility refers to the degree to which an utterance is understood under the given conditions. In (b), the global integrity decreases with segment duration getting longer, whereas in (c), the local integrity increases with segment duration getting longer.

	Type of auditory stimuli	Abbreviation	Local integrity	Global integrity	Intelligibility
(a)	Natural speech recordings played forward	F	(Yes)	Yes	Perfect
(b)	Locally time-reversed speech	LTR-F	No	Yes	Close to perfect at 20-ms segment duration, but almost none at 120 ms (Ueda *et al.*, 2017)
(c)	Reversed playback of locally time-reversed speech	LTR-R	Yes	No	Unintelligible, regardless of segment duration (<200 ms)
(d)	Reversed speech	R	(No)	No	None

A. Participants

1. German-native participants

Two samples of native speakers of German were recruited from the same population in two separate phases in Darmstadt: $n = 38$ (28 female and 10 male, age range 17–43 years, median = 20, the procedure carried out in fall 2016) were exposed to German background speech, and $n = 43$ (19 female and 24 male, age range 18–56 years, median = 23, the procedure carried out in spring 2017) to Japanese background speech. All of them declared not to understand Japanese. The majority consisted of university students participating for course credit; the remainder was paid a honorarium of 8 Euros. All participants claimed to have normal hearing and normal or corrected normal vision.

2. Japanese-native participants

A screening was administered to Japanese-native participants in Fukuoka to check whether a potential participant had ever learned German because students can take German classes. Only candidates who had not learned German were assigned to a group in which participants were exposed to German background speech ($n = 41$; 13 female and 28 male, age range 19–37 years, median = 22). Another group of participants were exposed to Japanese background speech ($n = 42$; 15 female and 27 male, age range 18–29 year, median = 22). All participants were student volunteers. All of them had normal hearing (tested with an audiometer, Rion, AA-56, Rion Co., Ltd., Kokubunji, Japan, to assure less than 25-dB hearing loss within the frequency range of 250–8000 Hz) and normal or corrected normal vision.

B. Stimuli

1. Auditory stimuli

The speech material (German or Japanese) was extracted from a multilingual database of spoken sentences (NTT-AT, "Multi-lingual speech database 2002," NTT Advanced Technology Corp., Kawasaki, Japan) recorded with a 16-kHz sampling rate and 16-bit quantization. Recordings of 80 sentences in German and Japanese (only eight sentences in German and four sentences in Japanese contained single utterances of a numeral in the respective language), each of which was spoken by five female and five male native speakers, were extracted. Thus, 800 recordings were prepared for each language, eliminating irrelevant noise and silent periods before and after utterances by inspecting waveforms and listening to acoustic signals, leaving about 10-ms silent margins at the beginning and end of each recording. Average duration per sentence in German was 1.8 s [standard deviation (SD) = 0.38)] and in Japanese 2.6 s ($SD = 0.50$).

For each trial, eight sentences were concatenated to produce a desired 14-s irrelevant-speech stream. Different talkers were assigned to each sentence randomly without repetition in a trial. These were processed as illustrated in Fig. 1: They were either played back as such (original forward speech, F) or divided up into segments of 20, 70, or 120 ms duration, including 2.5-ms cosine ramps to fade in and out for further processing. Subsequently, each segment was reversed in time while maintaining the original order of segments, thus generating locally time-reversed speech (LTR-F) streams [see Fig. 1(b)]. Additional background speech conditions were produced by reversing these materials, i.e., locally-reversed [LTR-R; see Fig. 1(c)], or original speech signals [R; Fig. 1(d)]. For each of the resulting eight irrelevant-speech conditions, ten different exemplars were generated to present new acoustical material on each trial. Pink noise was generated for an additional non-speech control condition. Pink noise was faded in with a 10-ms cosine ramp. Both the pink noise and background speech streams were faded out with a cosine ramp during the last 1 s on each trial. The signal processing of the speech stimuli was performed with an in-house software written in J language (J Software, 2016), whereas pink noise was generated in MATLAB.

2. Visual stimuli

Nine digits from "1" to "9" were prepared in a Sans Serif font. The height of each digit on the screen was 20 mm during the exposure phase. The observation distance was about 0.45 m in Darmstadt, and 1.15 m in Kyushu. Thus, the visual angles of the stimuli were about 2.55 degrees in Darmstadt, and about 1.00 degree in Kyushu. The visual stimuli were generated with a software written in MATLAB.

C. Apparatus

Great care was taken to assure nearly identical auditory stimulation at both experimental sites, i.e., Technische Universität Darmstadt and Kyushu University laboratories, using headphones of the same make [Beyerdynamic DT 990 (Beyerdynamic GmbH, Heilbronn, Germany)] in a soundproof booth [Darmstadt: Industrial Acoustics Company

(Niederkrüchten, Germany); Kyushu: Music cabin SC3 (Takahashi Kensetsu, Kawasaki, Japan)]. The average sound pressure level (SPL) of the background auditory stimuli (except pink noise) at the headphones was adjusted to 74 dB SPL with a 1-kHz calibration tone, which was provided with the speech database, by using 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). The sound pressure level of pink noise was about 72 dB SPL. A D/A converter (a high-quality sound card) [RME multiface II (Audio AG, Haimhausen, Germany)] and a headphone amplifier [Behringer Pro 8 (Behringer, Zhongshan, China)] were used at Technische Universität Darmstadt, and 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)] were used at Kyushu University to drive the headphones.

The visual stimuli were presented on a computer screen, a TFT-LCD monitor with a resolution of 1280 × 1024 pixels [Darmstadt: Zalman, ZM-M190 (Seoul, Korea); Kyushu: Epson, LD1755S (Suwa, Japan)]. The screen was placed in the soundproof booth in Darmstadt, whereas it was placed outside of the booth and was visible through the double-glass window in Kyushu. The luminance of the screen was adjusted and fixed at a comfortable looking level in each site.

D. Procedure

Each trial was initiated by the participant mouse-clicking a button, which appeared on the screen. Then, a random permutation of eight digits (drawn from the set of 1 through 9 without repetition) was presented in the center of the screen. Digits were displayed for 1 s each without inter-stimulus intervals. The participants' task was to memorize the order of digits without overtly rehearsing, i.e., employing articulatory organs. They were instructed to ignore whatever sounds that were presented. After presentation of the digits, a blank screen appeared for 6 s (retention interval) before participants were asked to recall the series of digits. To that effect, a number pad was presented on the screen, and participants had to click on the digits in the respective order. The stream of irrelevant sound (14 s) was played back during the presentation of the digits and during the retention interval. Each sound condition (nine conditions in total: eight types of speech utterances as experimental conditions and pink noise as a control, baseline condition) was presented ten times, resulting in a total of 90 trials. The order of trials was randomized for each participant. The experiment was preceded by two practice trials and interrupted by three optional breaks. It took about 60 min to complete. The experiment was run with a software written in MATLAB using the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner *et al.*, 2007; Pelli, 1997). Retrospective self-reports on Japanese participants' mnemonic strategy were collected at the end of their sessions from 61 of the 83 Japanese participants in total. That is, 30 participants in the Japanese background conditions and 31 in the German background conditions provided the self-reports.

III. RESULTS

Two German participants in the Japanese background conditions chose to discontinue the experiment. The self-reports led to the omission of the results of two Japanese participants, one in the Japanese background conditions and one in the German background conditions; they were judged to have disregarded the instructions—one participant in the Japanese conditions made articulatory movements silently to memorize digits, and another participant in the German conditions tapped a desk with his finger, trying to mask the auditory stimuli. The omission did not alter statistical conclusions. About 96% of 59 Japanese participants reported that they used some kinds of mnemonics. The majority (about 63%) tried to make puns (construct meaningful words from the sequence of digits), group the digits into several sets, or combine these two methods.

Performance, i.e., correct response rates at each position in the sequence, in each of the nine background sound conditions, was averaged within and across subjects. For clarity, curves collapsed across segment durations in locally time-reversed speech conditions are depicted in Fig. 2. The following statistical analyses were performed on the arc-sine transformed (Snedecor and Cochran, 1989) rates. Performance in the pink-noise control conditions was significantly better than in any other condition for all groups of participants (Fig. 2; *post hoc* Tukey-Kramer honestly significant difference tests, $p < 0.05$). Thus, clear ISEs were observed. The curves of experimental conditions, i.e., conditions other than the pink-noise control condition, in each panel of Fig. 2, show a typical pattern of the serial-position effect (e.g., Glanzer and Cunitz, 1966; LeCompte *et al.*, 1997; LeCompte and Shaibe, 1997; Murdock, 1962). Performance of Japanese participants is generally better than that of German participants, as is confirmed with a Wilcoxon test, $z = 40.88$, $p < 0.001$.

Averaged numbers of digits correctly reported at the appropriate position in the sequence are depicted in Fig. 3. Focusing on the forward speech in the native language conditions (filled inverted triangles and filled circles in the left column of Fig. 3; F and LTR-F), unsegmented utterances and locally time-reversed speech with 20-ms segments resulted in the least numbers of digits recalled in both groups of participants, whereas performance improved as the segment durations were lengthened. Reversed speech (open inverted triangles and open squares in the left column of Fig. 3; R and LTR-R) generally produced comparable levels of performance to those of LTR-F with 70- or 120-ms segments. In the non-native language conditions (the right column of Fig. 3), no systematic difference appeared between LTR-F (filled circles) and LTR-R (open squares) in both participant groups; on the other hand, in German participants [Fig. 3(b)], no difference between F and R was observed, whereas in Japanese participants [Fig. 3(d)], R exhibited a smaller ISE than F.

These observations are supported by two groups of statistical analyses, in which the pink-noise control condition

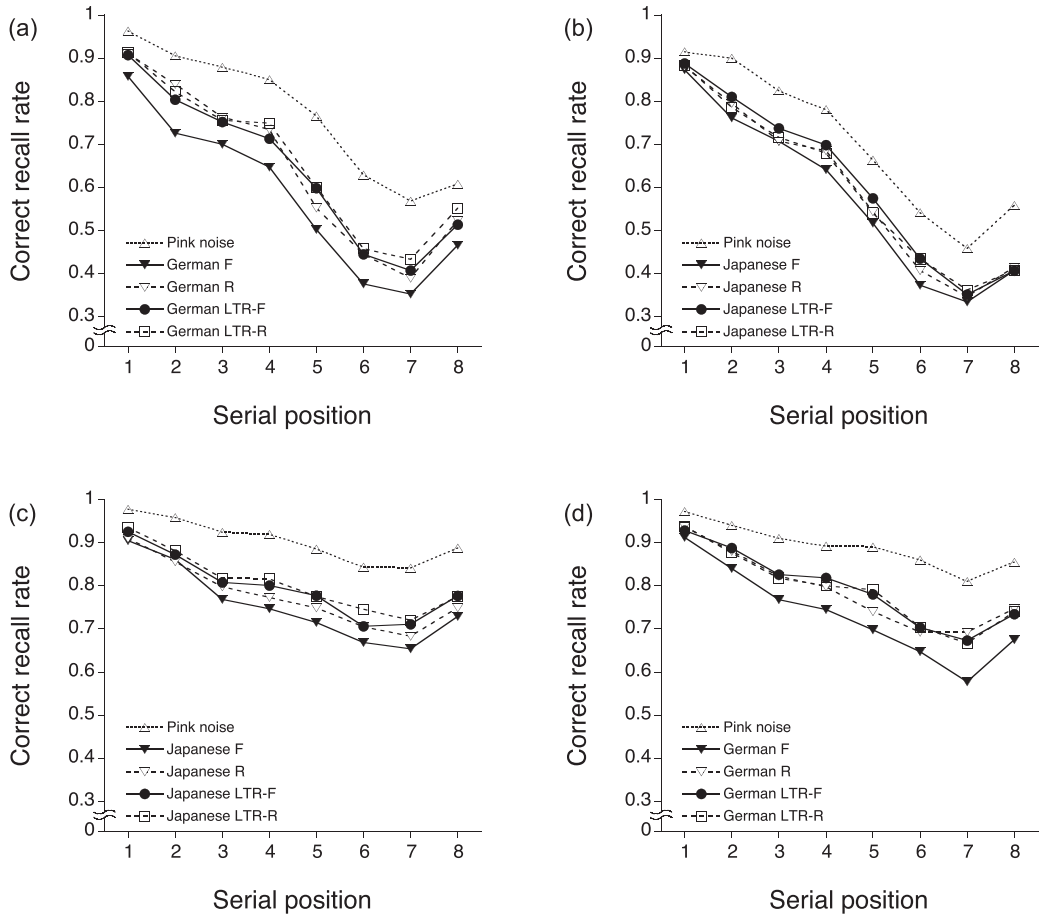

FIG. 2. Correct recall rate as a function of serial position and background condition ($n = 160$, in total). For clarity, only the curves collapsed across segment durations are shown for the locally time-reversed speech conditions. (a) German and (b) Japanese background speech was presented to German participants ($n = 38$, $n = 41$), whereas (c) Japanese and (d) German background speech was presented to Japanese participants ($n = 41$, $n = 40$). Thus, (a) and (c): Native language was presented as irrelevant speech; (b) and (d): Non-native language was presented as irrelevant speech. In the legends, F: forward; R: reversed; LTR: locally time-reversed speech.

was excluded: a $2 \times 2 \times 2 \times 4$ (Participant Group [German, Japanese] × Language [native, non-native] × Direction [forward, reversed] × Segment Duration [0, 20, 70, 120 ms]) mixed analysis of variance (ANOVA) applied to the overall results, and 2×4 (Direction [forward, reversed] × Segment Duration [0, 20, 70, 120 ms]) repeated-measures ANOVAs applied to each combination of participant groups and languages presented, which corresponds to each panel in Fig. 3.

The $2 \times 2 \times 2 \times 4$ mixed ANOVA applied to the whole set of experimental conditions revealed significant main effects of participant group, $F(1, 156) = 28.01$, $p < 0.001$, $\eta_G^2 = 0.115$, and segment duration, $F(3, 468) = 11.61$, $p < 0.001$, $\eta_G^2 = 0.009$. A significant interaction effect was found between direction and segment duration, $F(3, 468) = 3.97$, $p = 008$, $\eta_G^2 = 0.003$. No other effect was statistically significant.

The 2×4 (Direction [forward, reversed] × Segment Duration [0, 20, 70, 120 ms]) repeated-measures ANOVAs applied to each combination of participant groups and languages yielded the following. A main effect of direction was statistically significant only in German participants being presented German speech backgrounds [Fig. 3(a)], $F(1, 37) = 10.44$, $p = 0.003$, $\eta_G^2 = 0.003$. Significant main effects of segment duration were found in all groups of participants, German participants being presented German [Fig. 3(a)], $F(3, 111) = 18.30$, $p < 0.001$, $\eta_G^2 = 0.014$, German participants being presented Japanese [Fig. 3(b)], $F(3, 120) = 10.03$, $p < 0.001$, $\eta_G^2 = 0.008$, Japanese participants being presented Japanese [Fig. 3(c)], $F(3, 120) = 8.26$, $p < 0.001$, $\eta_G^2 = 0.006$, and Japanese participants being presented German [Fig. 3(d)], $F(3, 117) = 6.90$, $p < 0.001$, $\eta_G^2 = 0.005$. Significant interaction effects between direction and segment duration were found in German participants being presented German [Fig. 3(a)], $F(3, 111) = 8.75$, $p < 0.001$, $\eta_G^2 = 0.006$, Japanese participants being presented Japanese [Fig. 3(c)], $F(3, 120) = 3.10$, $p = 0.03$, $\eta_G^2 = 0.002$, and Japanese participants being presented German [Fig. 3(d)], $F(3, 117) = 4.13$, $p = 0.01$, $\eta_G^2 = 0.003$, but the interaction effect was only marginally significant in German participants being presented Japanese [Fig. 3(b)], $F(3, 120) = 2.40$, $p = 0.07$, $\eta_G^2 = 0.002$. Significant simple effects were found in German participants being presented German [Fig. 3(a)], for direction at 0 ms, $F(1, 37) = 13.80$, $p < 0.001$, $\eta_G^2 = 0.014$, and at 20 ms, $F(1, 37) = 16.79$, $p < 0.001$, $\eta_G^2 = 0.018$, and for segment duration with forward speech, $F(3, 111) = 25.46$, $p < 0.001$, $\eta_G^2 = 0.036$; in German participants being presented Japanese [Fig. 3(b)], for segment duration with forward speech, $F(3, 120) = 10.92$, $p < 0.001$, $\eta_G^2 = 0.016$; in Japanese participants being presented Japanese

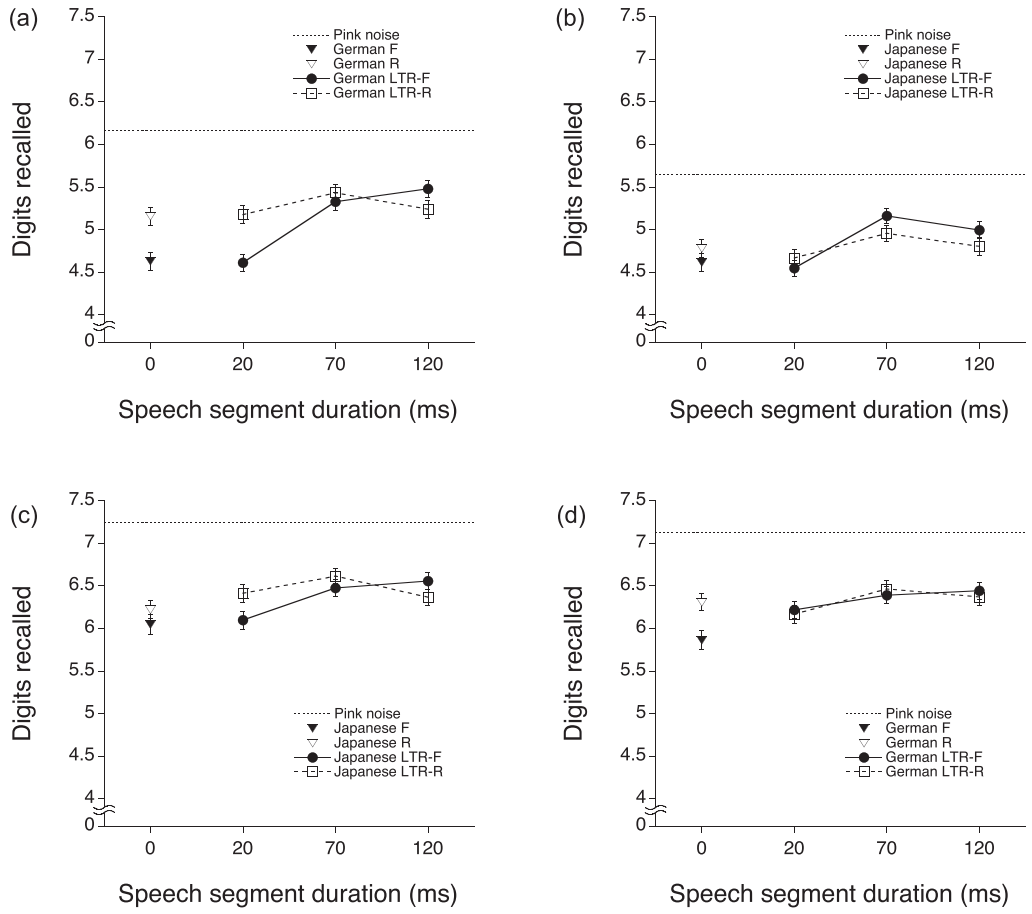

FIG. 3. Averaged performance ($n = 160$, in total) under irrelevant sound conditions. (a) German and (b) Japanese background speech was presented to German participants ($n = 38$, $n = 41$), whereas (c) Japanese and (d) German background speech was presented to Japanese participants ($n = 41$, $n = 40$). (a) and (c): Native language was presented as irrelevant speech; (b) and (d): Non-native language was presented as irrelevant speech. In the legends, F: forward; R: reversed; LTR: locally time-reversed speech. Error bars represent SEM.

[Fig. 3(c)], for direction at 20 ms, $F(1, 40) = 6.39$, $p = 0.02$, $\eta_G^2 = 0.006$, for segment duration with forward speech, $F(3, 120) = 8.37$, $p < 0.001$, $\eta_G^2 = 0.012$, and for segment duration with reversed speech, $F(3, 120) = 3.14$, $p = 0.028$, $\eta_G^2 = 0.005$; in Japanese participants being presented German [Fig. 3(d)], for direction at 0 ms, $F(3, 39) = 13.51$, $p < 0.001$, $\eta_G^2 = 0.011$, and for segment duration with forward speech, $F(3, 117) = 9.05$, $p < 0.001$, $\eta_G^2 = 0.012$. No other effect was statistically significant.

IV. DISCUSSION

Substantial ISEs were obtained in both German-speaking and Japanese-speaking participants for both background languages: Performance was considerably worse in all conditions involving speech (processed or not) than in the pink-noise control conditions (Figs. 2 and 3). This was true regardless of whether the participants' native language (left column in Fig. 3) or the foreign (incomprehensible) language was presented (right column of Fig. 3). Both sets of curves show ISEs of approximately equal magnitude: Average performance decrements of German participants being exposed to German background speech amounted to 17%; when exposed to Japanese, it amounted to 15%. Likewise, the performance decrement in Japanese participants was 12% when exposed to Japanese and 12% when exposed to German in the background. This suggests that neither the semantics of the irrelevant speech, nor the particular language-specific phonetics account for the overall disruption. When contrasting the results for free-running forward speech and the same signal played backwards (leftmost, filled and open inverted triangles in each panel of Fig. 3), it becomes obvious that reversed speech produces significant disruption, consistent with the bulk of the evidence in the literature (Jones *et al.*, 1990; LeCompte *et al.*, 1997; Röer *et al.*, 2014; Röer *et al.*, 2017; Surprenant *et al.*, 2007) and suggesting that the semantics of irrelevant speech are not crucial in producing auditory distraction.

However, the present results also suggest that global integrity of speech cues, linguistic comprehensibility, and intelligibility affected the performance of native participants. The following three points are the main findings as to locally time-reversed speech and its global reversal, i.e., LTR-F and LTR-R. First, in the native language background conditions (Fig. 3, left column), LTR-R did not produce greater disruption than LTR-F, and segment duration had no effect in LTR-R, supporting hypothesis H_G and rejecting hypothesis H_L. Thus, the global integrity dominated the size of ISE in the native participants. Second, in the non-native language background conditions (Fig. 3, right column), no statistically significant difference between LTR-F and LTR-R was

found, supporting neither H_G nor H_L. This suggests that the effects found in participants being exposed to their own native language were most plausibly caused by linguistic comprehensibility, i.e., their capability of processing the language or its elements to a greater extent. Third, in the native language background conditions (Fig. 3, left column), LTR-F with short segment duration produced as much disruption of serial recall as forward speech (F), suggesting that intelligibility of a comprehensible language has an effect on the magnitude of the ISE.

In addition, focusing on unsegmented forward vs reversed speech (filled and open inverted triangles in each panel of Fig. 3), one may notice that German background speech (in panels a and d of Fig. 3) exhibited a clear effect of reversal, whereas Japanese background speech [in panels (b) and (c) of Fig. 3] did not. The results provide further evidence showing that reversed speech does not always produce ISEs of comparable magnitude as does forward speech. Although it is difficult to pin down the reason, a noticeable difference in phonemic structures of the two languages possibly caused the contrasting results: German exhibits a much larger proportion of consonant durations in a sentence, 61%, than Japanese, 49% (Ueda *et al.*, 2017; cf. Ramus *et al.*, 1999), reflecting much more frequent occurrences of consonants in German than in Japanese—Japanese in most cases allows only one consonant in a syllable, whereas many of German syllables contain more than one consonant. A language with more rapidly changing spectra, German, may produce a larger contrast between forward and reversed speech than a language with less rapidly changing spectra, Japanese, because the more frequent anomalous changes in reversed German speech may interfere with the automatic speech processing to a greater extent, hence a smaller ISE.

Generally, Japanese participants performed better than German participants. This may partially explain why the performance differences caused by direction of playback tended to be smaller in Japanese participants. Retrospective self-reports in Japanese participants revealed that most of them used mnemonics, especially puns and rhythmic grouping. Since the short form of a traditional way of counting numbers in Japanese uses only one mora (short syllable) for each digit, like "hi, fu, mi, yo, i, mu, na, ya, ko," corresponding from one to nine, and each mora can be a building block of Japanese words (that consist of one or more morae), it might be easy to make puns with just a few morae for some Japanese participants. Moreover, the short form with regular mora timing should facilitate grouping, promote chunking, and save time for rehearsal. These factors probably contributed to the better performance of Japanese participants. The same explanation may hold for earlier findings of Japanese participants performing better in the serial recall of digits than native speakers of other languages (Ellermeier *et al.*, 2015; Hellbrück *et al.*, 1996).

In sum, the largest ISEs were obtained for forward speech and locally time-reversed speech with short segment duration, but only when the participants' native language was presented. Any speech stimuli in either language, however, produced an ISE to some extent. For non-native speakers who did not master the language, the direction of playing and segment duration had no systematic effect on the disruption caused by locally time-reversed speech, suggesting that the effects found in the native speakers most plausibly depended on linguistic comprehensibility. This implies that the global integrity takes precedence over local integrity of speech cues in governing the disruptive effects of irrelevant speech, at least for participants mastering the irrelevant language.

ACKNOWLEDGMENTS

The authors would like to thank Karla Salazar Espino and Maria Hernando for running a part of the experiment with German participants, Katharina Rost and Akie Shibata for running a part of the experiment with Japanese participants, and Ger Remijn for valuable discussion. This work was partly sponsored by Grants-in-Aid for Scientific Research Nos. 14101001, 25242002, 17K18705, and 19H00630 from the Japan Society for the Promotion of Science (JSPS).

Baddeley, A. (**1986**). *Working Memory* (Clarendon Press, Oxford).
Banbury, S., Macken, W., Tremblay, S., and Jones, D. (**2001**). "Auditory distraction and short-term memory: Phenomena and practical implications," Human Factors **43**(1), 12–29.
Bell, R., Röer, J. P., Lang, A.-G., and Buchner, A. (**2019**). "Reassessing the token set size effect on serial recall: Implications for theories of auditory distraction," J. Exp. Psychol.: Learn. Mem. Cogn. (published online).
Boersma, P., and Weenink, D. (**2016**). "Praat: Doing phonetics by computer [computer program]," Version 6.0.21, http://www.praat.org/ (Last viewed 9 November 2016).
Brainard, D. H. (**1997**). "The psychophysics toolbox," Spatial Vision **10**, 433–436.
Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, UK), p. 773.
Caplan, D., Waters, G., and Howard, D. (**2012**). "Slave systems in verbal short-term memory," Aphasiology **26**(3–4), 279–316.
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.
Colle, H. A., and Welsh, A. (**1976**). "Acoustic masking in primary memory," J. Verb. Learn. Verb. Behav. **15**, 17–31.
Dorsi, J. (**2013**). "Recall disruption produced by noise-vocoded speech: A study of the irrelevant sound effect," M.S. thesis, State University of New York, New Paltz, NY.
Ellermeier, W., Kattner, F., Ueda, K., Doumoto, K., and Nakajima, Y. (**2015**). "Memory disruption by irrelevant noise-vocoded speech: Effects of native language and the number of frequency bands," J. Acoust. Soc. Am. **138**(3), 1561–1569.
Ellermeier, W., and Zimmer, K. (**2014**). "The psychoacoustics of the irrelevant sound effect," Acoust. Sci. Technol. **35**, 10–16.
Giraud, A.-L., and Poeppel, D. (**2012**). "Cortical oscillations and speech processing: Emerging computational principles and operations," Nat. Neurosc. **15**(4), 511–517.
Glanzer, M., and Cunitz, A. R. (**1966**). "Two storage mechanisms in free recall," J. Verb. Learn. Verb. Behav. **5**, 351–360.
Greenberg, S., and Arai, T. (**2001**). "The relation between speech intelligibility and the complex modulation spectrum," in *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech-2001)*, September 3–7, Aalborg, Denmark, pp. 473–476.
Hellbrück, J., Kuwano, S., and Namba, S. (**1996**). "Irrelevant background speech and human performance: Is there long-term habituation?," J. Acoust. Soc. Jpn. (E) **17**, 239–247.
Hickok, G., and Poeppel, D. (**2007**). "The cortical organization of speech processing," Nat. Rev. Neurosci. **8**, 393–402.
Hughes, R. W., and Marsh, J. E. (**2017**). "The functional determinants of short-term memory: Evidence from perceptual-motor interference in verbal serial recall," J. Exp. Psychol.: Learn. Mem. Cogn. **43**(4), 537–551.

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**, 1749.
J Software (**2016**). "The J programming language," http://www.jsoftware.com (Last viewed 12 June 2019).
Jones, D. (**1993**). "Objects, streams, and threads of auditory attention," in *Attention: Selection, Awareness, and Control: A Tribute to Donald Broadbent*, edited by A. Baddeley and L. Weiskrantz (Clarendon Press, Oxford), pp. 87–104.
Jones, D. M., and Macken, W. J. (**1995**). "Phonological similarity in the irrelevant speech effect: Within- or between-stream similarity?," J. Exp. Psychol.: Learn. Mem. Cogn. **21**, 103–115.
Jones, D. M., Miles, C., and Page, J. (**1990**). "Disruption of proofreading by irrelevant speech: Effects of attention, arousal or memory?," App. Cog. Psych. **4**(2), 89–108.
Kellogg, E. W. (**1939**). "Reversed speech," J. Acoust. Soc. Am. **10**, 324–326.
Kiss, M., Cristescu, T., Fink, M., and Wittmann, M. (**2008**). "Auditory language comprehension of temporally reversed speech signals in native and non-native speakers," Acta Neurobiol. Exp. **68**, 204–213.
Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., and Broussard, C. (**2007**). "What's new in psychtoolbox-3?," Perception **36**(14), 1–16.
Lange, E. B. (**2005**). "Disruption of attention by irrelevant stimuli in serial recall," J. Mem. Lang. **53**, 513–531.
Larsen, L. D., Baddeley, A., and Andrade, J. (**2000**). "Phonological similarity and the irrelevant speech effect: Implications for models of short-term verbal memory," Memory **8**(3), 145–157.
LeCompte, D., Neely, C., and Wilson, J. (**1997**). "Irrelevant speech and irrelevant tones: The relative importance of speech to the irrelevant speech effect," J. Exp. Psychol.: Learn. Mem. Cogn. **23**, 472–483.
LeCompte, D. C., and Shaibe, D. M. (**1997**). "On the irrelevance of phonological similarity to the irrelevant speech effect," Quart. J. Exp. Psychol. **50A**(1), 100–118.
Licklider, J. C. R., and Miller, G. A. (**1951**). "The perception of speech," in *Handbook of Experimental Psychology*, edited by S. S. Stevens (John Wiley, New York), pp. 1040–1074.
Marsh, J. E., and Jones, D. M. (**2010**). "Cross-modal distraction by background speech: What role for meaning?," Noise Health **12**(49), 210–216.
Marsh, J. E., Yang, J., Qualter, P., Richardson, C., Perham, N., Vachon, F., and Hughes, R. W. (**2018**). "Postcategorical auditory distraction in short-term memory: Insights from increased task load and task type," J. Exp. Psychol.: Learn. Mem. Cogn. **44**(6), 882–897.
Meunier, F., Cenier, T., Barkat, M., and Magrin-Chagnolleau, I. (**2002**). "Mesure d'intelligibilité de segments de parole à l'envers en français" ("Measuring intelligibility of reversed speech segments in French"), in *XXIVèmes Journées d'Étude sur la Parole, Nancy*, June 24–27, Nancy, France, pp. 117–120.
Meyer-Eppler, W. (**1950**). "Reversed speech and repetition systems as means of phonetic research," J. Acoust. Soc. Am. **22**(6), 804–806.
Murdock, B. B. J. (**1962**). "The serial position effect of free recall," J. Exp. Psychol. **64**(5), 482–488.
Nakajima, Y., Matsuda, M., Ueda, K., and Remijn, G. B. (**2018**). "Temporal resolution needed for auditory communication: Measurement with mosaic speech," Front. Human Neurosci. **12**, 149.
Neath, I. (**2000**). "Modeling the effects of irrelevant speech on memory," Psychonom. Bull. Rev. **7**(3), 403–423.
Pelli, D. G. (**1997**). "The videotoolbox software for visual psychophysics: Transforming numbers into movies," Spatial Vision **10**, 437–442.
Poeppel, D. (**2003**). "The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time,'" Speech Commun. **41**, 245–255.
Ramus, F., Nespor, M., and Mehler, J. (**1999**). "Correlates of linguistic rhythm in the speech signal," Cognition **73**(3), 265–292.
Remez, R. E., Thomas, E. F., Dubowski, K. R., Koinis, S. M., Porter, N. A. C., Paddu, N. U., Moskalenko, M., and Grossman, Y. S. (**2013**). "Modulation sensitivity in the perceptual organization of speech," Atten. Percept. Psychophys. **75**(7), 1353–1358.
Röer, J. P., Bell, R., and Buchner, A. (**2014**). "Evidence for habituation of the irrelevant sound effect on serial recall," Mem. Cogn. **42**, 609–621.
Röer, J. P., Körner, U., Buchner, A., and Bell, R. (**2017**). "Semantic priming by irrelevant speech," Psychonom. Bull Rev. **24**, 1205–1210.
Saberi, K., and Perrott, D. R. (**1999**). "Cognitive restoration of reversed speech," Nature **398**, 760.
Salamé, P., and Baddeley, A. (**1982**). "Disruption of short-term memory by unattended speech: Implications for the structure for working memory," J. Verb. Learn. Verb. Behav. **21**, 150–164.
Sanders, L. D., and Poeppel, D. (**2007**). "Local and global auditory processing: Behavioral and ERP evidence," Neuropsychologia **45**(6), 1172–1186.
Schlittmeier, S. J., Weißgerber, T., Kerber, S., Fastl, H., and Hellbrück, J. (**2012**). "Algorithmic modeling of the irrelevant sound effect (ISE) by the hearing sensation fluctuation strength," Atten. Percept. Psychophys. **74**(1), 194–203.
Senan, T. U., Jelfs, S., and Kohlrausch, A. (**2018a**). "Cognitive disruption by noise-vocoded speech stimuli: Effects of spectral variation," J. Acoust. Soc. Am. **143**(3), 1407–1416.
Senan, T. U., Jelfs, S., and Kohlrausch, A. (**2018b**). "Erratum: Cognitive disruption by noise-vocoded speech stimuli: Effects of spectral variation [J. Acoust. Soc. Am. **143**(3), 1407–1416 (2018)]," J. Acoust. Soc. Am. **144**(3), 1330.
Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.
Snedecor, G. W., and Cochran, W. G. (**1989**). *Statistical Methods*, 8th ed. (Iowa State University Press, Ames, IA), pp. 289–290.
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, Germany), 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.
Surprenant, A. M., Neath, I., and Bireta, T. J. (**2007**). "Changing state and the irrelevant sound effect," Can. Acoust. **35**, 86–87.
Teng, X., Tian, X., and Poeppel, D. (**2016**). "Testing multi-scale processing in the auditory system," Sci. Rep. **6**, 34390.
Ueda, K., Nakajima, Y., Ellermeier, W., and Kattner, F. (**2017**). "Intelligibility of locally time-reversed speech: A multilingual comparison," Sci. Rep. **7**, 1782.
Viswanathan, N., Dorsi, J., and George, S. (**2014**). "The role of speechspecific properties of the background in the irrelevant sound effect," Quart. J. Exp. Psychol. **67**(3), 581–589.