Original Article

QJEP

Quarterly Journal of Experimental Psychology
2018, Vol. 71(10) 2152–2161
© Experimental Psychology Society 2017
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1747021817739257
qjep.sagepub.com
SAGE

The role of speech fidelity in the irrelevant sound effect: Insights from noise-vocoded speech backgrounds

Josh Dorsi[1], Navin Viswanathan[2], Lawrence D Rosenblum[1] and James W Dias[1]

Abstract

The Irrelevant Sound Effect (ISE) is the finding that background sound impairs accuracy for visually presented serial recall tasks. Among various auditory backgrounds, speech typically acts as the strongest distractor. Based on the changing-state hypothesis, speech is a disruptive background because it is more complex than other nonspeech backgrounds. In the current study, we evaluate an alternative explanation by examining whether the speech-likeness of the background (speech fidelity) contributes, beyond signal complexity, to the ISE. We did this by using noise-vocoded speech as a background. In Experiment 1, we varied the complexity of the background by manipulating the number of vocoding channels. Results indicate that the ISE increases with the number of channels, suggesting that more complex signals produce greater ISEs. In Experiment 2, we varied complexity and speech fidelity independently. At each channel level, we selectively reversed a subset of channels to design a low-fidelity signal that was equated in overall complexity. Experiment 2 results indicated that speech-like noise-vocoded speech produces a larger ISE than selectively reversed noise-vocoded speech. Finally, in Experiment 3, we evaluated the locus of the speech-fidelity effect by assessing the distraction produced by these stimuli in a missing-item task. In this task, even though noise-vocoded speech disrupted task performance relative to silence, neither its complexity nor speech fidelity contributed to this effect. Together, these findings indicate a clear role for speech fidelity of the background beyond its changing-state quality and its attention capture potential.

Keywords

Irrelevant sound effect; noise-vocoded speech; speech perception

Received: 7 October 2016; revised: 1 September 2017; accepted: 13 September 2017

The Irrelevant Sound Effect (ISE) is the observation that irrelevant background sounds, such as speech or tones, reduce the accuracy of serial recall, relative to background noise or silence (Colle & Welsh, 1976; Jones & Macken, 1993; Salamé & Baddeley, 1987). For example, participants in a typical ISE paradigm view a sequence of letters or numbers appearing one at a time while a background sound is presented through headphones or speakers (e.g., Colle & Welsh, 1976; Jones & Macken, 1993; Salamé & Baddeley, 1987). Even when participants are instructed to ignore these backgrounds, recall of the visually presented sequence is impaired (Colle & Welsh, 1976; Jones & Macken, 1993; Salamé & Baddeley, 1987). Interestingly, the effect of speech is found even for foreign speech or non-words (Colle & Welsh, 1976; Jones, Miles, & Page, 1990; Salamé & Baddeley, 1982).

These early findings highlighted the utility of the ISE for studying short-term memory and for understanding auditory distraction, and thus have motivated a sizable literature. For example, the ISE has been cited as strong evidence for the Working Memory Model (Baddeley & Hitch, 1974). On this model, the phonological loop of working memory maintains the serial order of targets (letters or numbers), while speech gains automatic access to this memory system (Salamé & Baddeley, 1987). The finding

[1]University of California, Riverside, Riverside, CA, USA
[2]University of Kansas, Lawrence, KS, USA

Corresponding author:
Josh Dorsi, Department of Psychology, University of California, Riverside, 900 University Avenue, Riverside, CA 92521, USA.
Email: jdors002@ucr.edu

that non-speech auditory stimuli, such as music or even tones, can also produce the ISE challenges this model's speech-specificity assumption (Jones, 1993; Jones & Macken, 1993; Salame & Baddeley, 1989).

Such findings prompted researchers to investigate the general properties of sound, not specific to speech, that affect serial recall accuracy. This investigation revealed the *changing-state effect*: serial recall accuracy changes inversely with the number of perceived auditory segments in the background, a characteristic known as the sound's changing-state quality (Jones & Macken, 1993; Macken, 2014). For example, an irrelevant background of different tones will impair serial recall accuracy more than a single repeating tone (e.g., Jones & Macken, 1993). This finding motivated the *changing-state hypothesis*, explaining that this pattern of effects occurs because, relative to the different tone condition, the changing-state nature of the single repeating tone is reduced. This highlights the essential stimulus-to-disruption relationship of the changing-state hypothesis; the ISE corresponds to the functional (acoustic/perceptual) complexity of the signal (henceforth the signal's 'changing-state complexity').

The changing-state hypothesis has the benefit of accounting for nonspeech effects in the ISE by suggesting that the ISE is the result of conflict between two serial processes; the process involved in the focal serial recall task, and the organisation of the auditory objects (e.g., Jones & Macken, 1993; Macken, 2014). Under this account, backgrounds with alternating tones are organised resulting in cues indicating the order of the changes in a sound sequence. These order cues then interfere with the maintenance of the serial order of the to-be-remembered items.

Despite this parsimony, this account has a notable limitation: a growing literature suggests that the content, in addition to the changing-state complexity, of speech can influence the amount of serial recall disruption. For example, participants have lower serial recall accuracy when the irrelevant background contains their name (e.g., Röer, Bell, & Buchner, 2013). Similarly, negative valence words such as "Apathetic" produce greater disruption than do neutral valence words such as "Curious" (Buchner, Rothermund, Wentura, & Mehl, 2004[1]). Such findings challenge the notion that the general changing-state complexity of an auditory signal is the sole cause of the ISE.

A recent account of the ISE that can reconcile the reliable changing-state effect, with the effects of speech content is the *duplex-mechanism account* offered by Hughes (2014; see also Hughes, Vachon, & Jones, 2007). This account assumes that irrelevant sound disrupts serial recall at two loci; the "interference-by-process" and the "attention-capture" mechanisms. Here, the interference-by-process mechanism is the same mechanism invoked by the changing-state hypotheses (discussed above). The other mechanism, attention capture, can be "specific" when the background sound is "… meaningful or of interest to the individual" (Hughes, 2014, p. 31), as when the background contains the listeners name (e.g., Röer et al., 2013). Attention capture may also be "aspecific," when the irrelevant sound alone is meaningless, but differences between it and other tokens from the irrelevant sound, causes it to exogenously capture attention (Hughes, 2014). According to Hughes, aspecific attention capture requires that an item within the auditory stimulus must violate the listeners' expectations (Hughes, 2014). For example, aspecific attention capture may occur when a single word is spoken by a male voice within a stream of words produced by a female speaker (Hughes et al., 2007; see also Hughes, 2014).

The changing-state hypothesis and the duplex-mechanism accounts share the basic approach of comparing serial recall disruption associated with different backgrounds in order to make inferences about the structure of underlying cognitive mechanisms. Other researchers have focused on the structure of the irrelevant stimulus to identify acoustic characteristics shared across backgrounds in order to offer a more precise explanation of "changing-state complexity" or to otherwise better characterise the stimulus-to-recall disruption relationship. This research revealed that even though both speech and nonspeech backgrounds produce the ISE, many of the largest disruptive effects are produced by speech. For instance, a recent study systematically examined ISEs produced by 40 different auditory backgrounds from several studies conducted within a single lab. These backgrounds included, speech, tones, music, and traffic and office noise (Schlittmeier, Weissgerber, Kerber, Fastl, & Hellbrück, 2012). Remarkably, across these diverse backgrounds, speech produced the largest ISE (Schlittmeier et al., 2012). Compatibly, a recent analysis that compared the ISE reported for several different types of background sounds also found that speech, including foreign, reversed, and laboratory transformations of speech are consistently more disruptive than non-speech backgrounds (Ellermeier & Zimmer, 2014).

The question of why speech backgrounds produce the strongest ISE is the focus of the current study. From a changing-state perspective, the potency of speech is attributed to its greater changing-state complexity relative to other nonspeech backgrounds. For instance, Tremblay, Nicholls, Alford, and Jones (2000) used sinewave-speech to investigate the role of speech perception in the ISE. Sinewave-speech is an acoustic transformation of natural speech that preserves the spectrotempral relationships of the speech signal in a series of time-varying sinusoids (see Remez, Rubin, Pisoni, & Carell, 1981). An interesting quality of sinewave-speech is that listeners may hear it as either speech or non-speech. While naive listeners may report that sinewave-speech sounds like computer beeps or bird sounds, listeners informed about the nature of sinewave-speech can perceive its linguistic content (Remez et al., 1981). Tremblay et al. (2000) investigated whether the ISE was dependent on whether perceivers identified the irrelevant sound as speech or non-speech. This permitted them to equate changing-state complexity of the signal while examining the effect of different percepts. Their results demonstrated that irrespective of training, both groups showed the same level of serial recall disruption. The authors concluded that the ISE produced by sinewave-speech was driven by its changing-state complexity and was independent of speech-likeness.

In a follow-up study, Viswanathan, Dorsi, and George (2014a), investigated this conclusion further. First, they noted that the sinewave-speech signal preserved the acoustic structure of speech irrespective of how listeners were trained to perceive it. In other words, regardless of whether the sinewave speech was identified as speech, the acoustic structure was still lawfully related to meaningful articulatory (speech) gestures. To determine if the acoustic structure produced by articulation (speech fidelity) or changing-state complexity was responsible for the ISE, Viswanathan et al. (2014a) created a special type of sinewave-speech in which they reversed two of the three sinusoids that made the sinewave speech signal (also see Viswanathan, Magnuson, & Fowler, 2014b). This manipulation disrupted the dynamic time-varying acoustic structure of the speech stimuli, reducing the lawful relationship to natural speech, while preserving acoustic complexity. In other words, sinewave speech contains both speech and changing-state complexity information, while selectively reversed sine-wave speech contains *only* changing-state complexity information. By comparing the effects of sinewave-speech, which maintained the signal's fidelity *and* its changing-state complexity, to selectively reversed sinewave-speech, which maintained its complexity but not its speech fidelity; the researchers isolated the effect of speech structure. Their results showed that higher speech-fidelity backgrounds (sinewave speech) are more disruptive than lower (no) speech-fidelity ones (selectively reversed sinewave speech) indicating that speech fidelity of the acoustic signal, beyond changing-state complexity, contributes to the disruptive properties of speech in the ISE. While this study offers preliminary evidence, because it did not independently manipulate complexity it does not conclusively indicate that speech fidelity always contributes to the ISE. It is possible that the effect of speech fidelity is only critical in reduced signals like 3-formant sinewaves.

Taken together, the studies reviewed above prompt the critical question: why is speech more disruptive than other background sounds? The goal of the current study is to evaluate the effects of speech fidelity (as was done previously by Viswanathan et al., 2014a) *and* speech signal complexity on the ISE using a different transformation of the speech signal that allows speech fidelity and changing-state complexity to be manipulated independently. To do this, we used *noise-vocoded speech* as the irrelevant background during a series of serial recall tasks. Noise-vocoded speech is a transformation of natural speech that is generated by dividing speech into frequency channels, mapping the intensity variation within each channel, and then applying these intensity variations to corresponding channels in white noise (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). Despite lacking the fine spectral detail of natural speech, noise-vocoded speech preserves its amplitude variations and can still be intelligible (Shannon et al., 1995). Prior research demonstrates that increasing the number of channels in noise-vocoded speech, from 1 to 20 channels, makes it more disruptive to serial recall (Ellermeier, Kattner, Ueda, Doumoto, & Nakajima, 2015; see also Wöstmann & Obleser, 2016).

While the work of Ellermeier et al. (2015) shows that increases in channel number increase the ISE, their study only used 1, 2, 4, and 20 channelled noise-vocoded speech. In Experiment 1, we investigate the effect of frequency channel number on the ISE further by examining the channels 3, 6, 9, and 12; spanning the lower range of intelligible noise-vocoded speech. While manipulating the number of vocoding channels in the background allows Experiment 1 to control the changing-state complexity of the backgrounds, this manipulation does not allow increasing changing-state complexity to be dissociated from increased speech-fidelity. Thus, in Experiment 2, we apply the selective-reversal process used in Viswanathan et al. (2014) to noise-vocoded speech to determine if the effect of channel number is independent of the effect of speech fidelity. In the context of the duplex mechanism account, it is not clear from Experiment 2 if this dissociation results from the interference-by-process or attention-capture mechanisms. To determine the locus of the effect of speech fidelity on the ISE, Experiment 3 presents the typical and selectively reversed noise-vocoded speech from Experiment 2 in the context of a missing-item task (e.g., Hughes et al., 2007).

Experiment 1

Experiment 1 investigates the effect of the number of channels in noise-vocoded speech on the ISE. Decreases in serial recall accuracy associated with increasing channel number (e.g., Ellermeier et al., 2015) may be the result of increased changing-state complexity as well as increased speech fidelity of the signal. This is because higher channel noise-vocoded speech preserves more speech information, as evident from its greater intelligibility. To better understand the influence of channel number (changing-state complexity) and speech-fidelity on the ISE, we chose 3, 6, 9, and 12 channelled noise-vocoded speech as an irrelevant background. This range of noise-vocoding channels was selected because it spans from minimally or non-intelligible to easily intelligible (e.g., Shannon et al., 1995). To provide a strong test of whether noise-vocoded speech produces the ISE, we chose to compare its disruptive effects to the effect of white noise, which has the same intensity as noise-vocoded speech but lacks the speech-like amplitude variation.

Method

Participants. Eighty-one students from the State University of New York at New Paltz received course credit for their participation. Participants were randomly assigned to one of four experimental groups: 3, 6, 9, and 12 channelled noise-vocoded speech. All subjects were native English speakers and reported normal hearing and normal or corrected to normal vision.

Materials. Noise-vocoded backgrounds were generated from the natural (non-SWS) speech tokens used by Viswanathan et al. (2014a); these tokens were as follows: *bowls, boy, day, dog, go, than*, and *view*. Noise-vocoded speech was synthesised using Praat (Davis et al., 2005). This script was modified to generate the four noise-vocoded speech conditions: 3, 6, 9, and 12 channels that were used in this study (see Appendix for additional details). White noise segments were matched in average intensity and duration to the noise-vocoded speech tokens.

The background tokens were arranged into four random ordered lists. In order to coincide with the presentation of the to-be-remembered items, each irrelevant sound list was 10 s long. For each list, the silent interval between tokens was between 150-300 ms. To reach the 10 s duration of the serial recall list, and to avoid long intervals between tokens, each list repeated each of the seven irrelevant words once (see Viswanathan et al., 2014a). Participants heard these acoustic stimuli through sound insulated headphones at 70 db. Each trial consisted of one randomly selected background list, and each experiment session repeated each list 6 times (see Viswanathan et al., 2014a). Every participant was presented with both noise-vocoded speech and white noise backgrounds. We were concerned that the potentially high intelligibility of the 12 and 9 channel conditions (e.g., Loizou, Dorman, & Tu, 1999) would bias the perception of the 3 and 6 channel conditions (e.g., Davis et al., 2005) and as such channels of noise-vocoded speech were tested between subjects.

The recall task consisted of visually presenting the targets: L R T S M K F (Tremblay et al., 2000; Viswanathan et al., 2014a). Each trial consisted of a random ordering of these target items. Participants saw these targets on a computer screen for 1000 ms each, with a 500 ms interval between items. Participant sat three feet from the computer screen, and targets appeared at the centre of the display 500 ms following a "***" fixation point. The first target appeared simultaneously with the first irrelevant sound item and the irrelevant sound persisted throughout the duration of the trial (see Viswanathan et al., 2014a for more details).

Procedure. Participants were told that they would see a series of letters on the screen and hear sounds through their headphones. Participants were instructed to report the presented letter sequence in the correct order and to ignore any sounds they heard. Participants initiated each serial recall task by pressing the spacebar on their keyboard. Participants were prompted to type in their response by a blinking cursor in the upper left corner of the computer screen 1000 ms following the presentation of the last visual item.

Results and discussion

Serial recall accuracy was measured as the number of to-be-remembered letters which were reported in their correct serial position. In total, 21 participants were placed into the 3 channel condition, 19 were placed in the 6 channel, 19 in the 9 channel, and 21 in the 12 channel conditions.

To determine whether noise-vocoded speech produced the ISE, we compared serial recall accuracy for white noise and noise-vocoded speech conditions (see Table 1) in a one-tailed paired sample t-test. This analysis found that noise-vocoded speech produced significantly lower serial recall accuracy, $t(79) = 5.259, p < .001, r = .509$, confirming that our stimuli produced the ISE. As there were not multiple levels of white noise a factorial analysis of variance (ANOVA) of these data was not appropriate. To determine the effect of the number of channels in noise-vocoded speech, we next transformed our data into difference scores by subtracting the noise-vocoded speech conditions from the white noise conditions (see Figure 1). These difference scores were submitted to a 4 level (3, 6, 9, and 12 channels) one-way ANOVA. We found that the number of noise-vocoded channels affected the degree of serial recall disruption, $F(3,\ 76) = 4.400$, $p\ <\ .007$, $\eta^2_p = .148$. In a follow-up analysis, we found a significant linear trend of channel number on serial recall accuracy, $F(3,\ 76) = 4.400$, $p = .002$, $\eta^2_p = .148$, supporting a linear relationship between channel and recall accuracy.[2]

These results indicate that increasing the number of channels in noise-vocoded speech results in increased serial recall disruption. What remains is to determine what about noise-vocoded speech channels affects serial recall disruption. Interestingly, pilot data demonstrated that the effect of vocoding channel quantity on the ISE does not correspond to the intelligibility of the noise-vocoded speech.[3] Is the effect of channel number in noise-vocoded speech due to increasing changing state-complexity, or increased speech fidelity of these backgrounds? To determine this, we conducted Experiment 2 in which we manipulated both changing-state complexity and speech fidelity independently.

Experiment 2

Experiment 2 extends Viswanathan et al. (2014a) by using selectively reversed noise-vocoded speech. In selectively reversed noise-vocoded speech, the lower two-thirds of the vocoded channels are temporally reversed relative to the

Table 1. Displays the raw serial recall accuracy scores for all conditions in Experiments 1 and 2. Bonferroni corrected paired comparisons in Experiment 1 found that all except the 3 channel condition were significantly different from white noise; across the noise-vocoded speech conditions both the 9 and 12 channel conditions were different from the 3 channel condition.

	Channels	White noise	Noise-vocoded speech	
Experiment 1	3	66%	65%	
	6	67%	64%	
	9	63%	53%	
	12	66%	58%	
	Channels	Silence	Noise-vocoded speech	Selectively reversed noise-vocoded speech
Experiment 2	6	70%	65%	67%
	12	72%	68%	71%
	18	72%	62%	65%

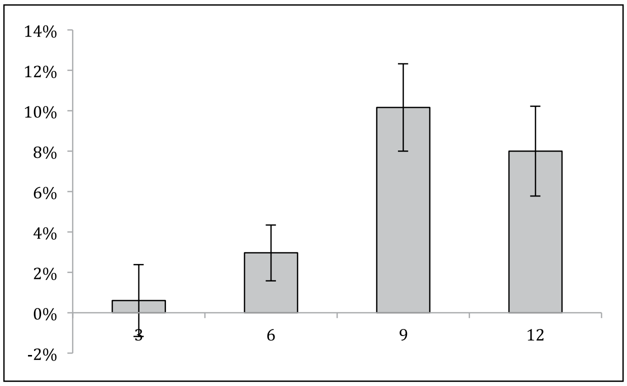

Figure 1. Difference scores (serial recall accuracy for white noise backgrounds—serial recall accuracy for noise-vocoded speech backgrounds) for four noise-vocoded speech participant groups; 3, 6, 9, and 12 channelled noise-vocoded speech used in Experiment 1.
Error bars represent standard error of the mean.

upper third. This manipulation is analogous to selectively reversed sinewave-speech and will likewise distort the speech information contained in the acoustic signal. Importantly, while selective reversal allows us to manipulate speech fidelity of an acoustic speech signal, the results of Experiment 1 suggest that the number of noise-vocoded channels allows us to manipulate changing-state complexity.

If the effect of noise-vocoded speech channels on serial recall accuracy is solely due to changing-state complexity, then within each channel condition selectively reversed and typical noise-vocoded speech should produce the same serial recall disruption; that is selective-reversal should not interact with channel number. Alternatively, if the speech fidelity of vocoded speech also matters, then, within different channel conditions selectively reversed noise-vocoded speech, with its low speech fidelity, should be less disruptive than typical noise-vocoded speech.

Experiment 1 investigated the effect of changing-state complexity in the ISE by measuring serial recall disruption associated with four different channel groups of noise-vocoded speech. These four different noise-vocoded channel conditions were presented between subjects so as to avoid cross condition learning effects. To eliminate the possibility that differences associated with noise-vocoded channels could be attributed to pre-existing group differences[4] Experiment 2 presented channel conditions within subjects, and used a blocked counterbalanced design to account for any learning effects.

To focus its investigation, Experiment 2 reduced its conditions to three levels of noise-vocoded channels. The three noise-vocoded channel groups in Experiment 2 were 6, 12, and 18, each differing by a factor of 6 channels and thus, as in Experiment 1, channel number increased linearly across conditions. The 18 channel condition was included to expand the range of channel conditions beyond what was used in the Experiment 1. Paired comparisons conducted in Experiment 1 failed to find a difference between any adjacent channel conditions; the increased channel difference between groups used in Experiment 2 may enhance any difference caused by channel number and thereby offer a better opportunity to detect channel differences. Additionally, Experiment 2 adopted larger sample sizes, which were more similar to prior work (e.g., Tremblay et al., 2000).

Method

Participants. Experiment 2 consisted of 77 participants from the University of California, Riverside. All participants were native English speakers, had normal hearing and normal or corrected to normal vision. All participants received course credit for their participation.

Materials. Experiment 2 used the same background words as Experiment 1. However, these words were recorded, synthesised into noise-vocoded speech, and arranged into random ordered lists specifically for Experiment 2. To make selectively reversed noise-vocoded speech, two-thirds of the frequency channels for each noise-vocoded speech token (approximately corresponding to 0-1700 hz range of the acoustic signal, see Appendix) were reversed relative to the remaining channels. Having established that noise-vocoded speech produces the ISE in Experiment 1, we opted to use silence (instead of white noise) as a control, in line with many studies of the ISE (e.g., Ellermeier et al., 2015; Elliott & Briganti, 2012; Elliott et al., 2016). All other aspects of the stimuli used in Experiment 2 were the same as those used in Experiment 1.

Procedure. The procedure of Experiment 1 was followed for the serial recall task of Experiment 2. A slight alteration was made to the visual presentation of the target items such that they were presented in the centre of a 1.5-inch square border located in the center of the computer monitor. Participants were prompted to type in their response by a ":" presented in the upper left corner of the display box, 1500 ms following the presentation of the last visual item. These additions affected all conditions equally. Instructions were provided orally by researchers from a prepared script. On screen instructions at the start of the experiment re-iterated the verbal instructions provided by researchers. The order of channels and the speech fidelity were manipulated within subjects with their order of presentation counterbalanced across different subjects. Prior to our main analyses we tested for and found no effect of sequence of channel presentation on the effects of channel number or selective-reversal.

Results and discussion

In Experiment 2, we compared the serial recall accuracy associated with noise-vocoded speech and selectively reversed noise-vocoded speech across three levels of vocoded channel composition. As was done for Experiment 1, scores for the serial recall task were calculated as the average number of letters reported in the correct serial position.

We first conducted paired samples t-tests comparing noise-vocoded speech and selectively reversed noise-vocoded speech to silence. These tests were Bonferroni corrected for 2 comparisons (alpha = .025). These tests confirmed that both noise-vocoded speech, $t(76)= 6.408$, $p < .001$, $r = .592$, and selectively reversed noise-vocoded speech, $t(76)= 4.078$, $p < 0.001$, $r = .423$, were significantly different from silence, showing that our stimuli were successful in producing the ISE. We then calculated the amount of ISE in each condition by subtracting the accuracy for noise-vocoded trials from the accuracy in silent trials in the same block (see Figure 2). These difference scores were used for all subsequent analyses.

These difference scores were submitted to a 2 (Background: noise-vocoded vs. selectively reversed) X 3 (Channel: 6, 12, and 18) repeated measures ANOVA. This analysis revealed a main effect of background, $F(1, 76) = 9.195$, $p = .003$, $\eta^2_p = .108$, demonstrating that noise-vocoded speech was more disruptive than selectively reversed noise-vocoded speech, and consistent with our hypothesis that the ISE is sensitive to speech fidelity. This analysis also found a main effect of channel, $F(2, 152) = 5.426$, $p = .005$, $\eta^2_p = .067$, consistent with the results of Experiment 1 and prior research showing that serial recall accuracy is sensitive to the complexity of the irrelevant background signal. No interaction was found, $F(2, 152) = .575$, $p = .564$, $\eta^2_p = .008$, indicating no evidence that the speech fidelity effect depended on the number of channels.

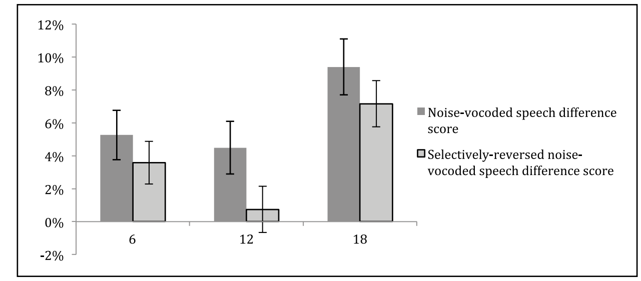

Figure 2. Difference scores (serial recall accuracy for silent backgrounds—serial recall accuracy for noise-vocoded backgrounds) for typical and selectively reversed noise-vocoded speech.
Three participant groups; 6, 12, and 18 channelled noise-vocoded speech used in Experiment 2. Error bars represent standard error of the mean.

To determine the locus of the effect of channel found in the serial recall task post hoc contrast analyses of the channel conditions for noise-vocoded speech and selectively reversed noise-vocoded speech were conducted. For noise-vocoded speech, this found a marginal linear effect of channel, $F(1, 76) = 3.766$, $p = .056$, $\eta^2_p = .047$, consistent with the result of Experiment 1. Interestingly, the selectively reversed conditions also resulted in a significant linear effect of channel, $F(1, 76) = 3.985$, $p = .049$, $\eta^2_p = .050$.[5] The results of this trend analysis indicate that the effect of channel is robust, being present even in the selectively reversed conditions. Collectively, these results suggest that the effect of changing-state complexity is present for both speech and non-speech conditions.

Based on these results it is clear that speech fidelity is important to the ISE; the selectively reversed noise-vocoded speech with less speech fidelity caused less disruption. This effect was observed despite the typical and selectively reversed noise-vocoded speech being matched in channel number. This makes it unlikely that the effect of speech fidelity could be attributed to differences in changing-state complexity.

There are two possible loci for the effect of speech fidelity on serial recall. First, it could be that the effect of speech fidelity is specific to the ISE and occurs by disrupting the serial process. Second, this effect could be reflective of signals with high speech fidelity preferentially engaging the attention-capture mechanisms that are posited by the duplex mechanism account (Hughes, 2014).

Experiment 3

We designed Experiment 3 to dissociate whether the effect of speech fidelity was due to interference-by-process or attention capture. Experiment 3 used the same stimuli as Experiment 2 but used a missing-item instead of serial recall task. The missing-item task presents participants with a sequence of items from a pre-defined limited set. The task for the participant is to indicate what item from the set was not included in the presentation. This paradigm shares many characteristics with the serial recall task (i.e., a small set of to-be-remember items presented sequentially) the critical difference being that the missing-item task does not require the participant to maintain serial order information.

The missing-item task has been used previously to determine whether effects observed in the ISE can be attributed to attention capture or interference-by-process[6] (e.g., Hughes et al., 2007). This is because the missing-item task does not require serial order information and therefore is not susceptible to interference from the serial information from the background sound (Hughes, 2014). As noted above, from the results of Experiment 2, it is unclear which process posited by the duplex mechanism account supports the observed effect of speech-fidelity in the ISE. If selectively reversed noise-vocoded speech produces disruption in the missing-item task, then the effect of selective reversal can be attributed more generally to the preferential engagement of the attention-capture mechanism (Hughes, 2014). Likewise, if the typical noise-vocoded speech produces more disruption than the selectively reversed noise-vocoded speech, then the effect of speech-fidelity is likely supported by attention capture. Alternatively, if we fail to find an effect of selective-reversal, then the effect of speech-fidelity found in Experiment 2 can be more easily attributed to interference-by-process. This effect would be consistent with our hypothesis that the ISE is sensitive to the speech fidelity of the acoustic signal.

Method

Participants. In all, 77 participants from the University of California, Riverside, participated in this experiment. All participants were native English speakers, had normal hearing and normal or corrected to normal vision and received course credit for their participation.

Materials. Experiment 3 used the same materials as Experiment 2.

Procedure. To make the results between experiments comparable, Experiment 3 used the same letter set for its missing-item task as was used in the serial recall task of Experiment 2. The missing-item task consisted of random ordered sequences of six of the letters from the seven letter set used Experiment 2 (F K L M R S T). Participants placed in this task were informed that their task would be to view sequences of six letters drawn from the seven-letter set and report the missing item. Participants were given the complete seven-letter set prior to beginning the missing-item task. As stated above, Experiment 3 used the same irrelevant backgrounds as Experiment 2.

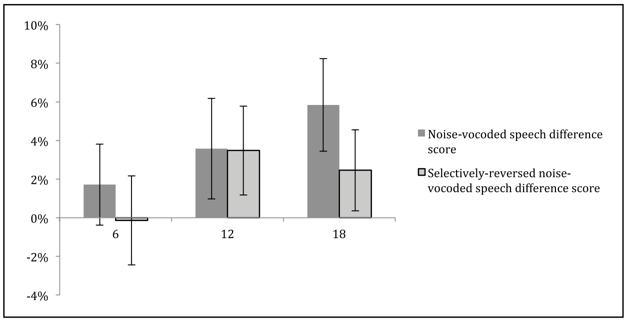

Figure 3. Difference scores (missing item accuracy for silent backgrounds—missing item accuracy for noise-vocoded backgrounds) for typical and selectively reversed noise-vocoded speech.
Three participant groups; 6, 12, and 18 channelled noise-vocoded speech used in Experiment 3. Error bars represent standard error of the mean.

Results and discussion

For the missing-item task accuracy was calculated as either correct if participants identified the missing item or incorrect if they indicated an item that was present or an item not from the set. We next calculated Bonferroni corrected paired samples *t*-tests comparing noise-vocoded speech and selectively reversed noise-vocoded speech to silence (alpha = .025). These analyses found a significant effect of noise-vocoded speech, $t(76)= 2.814$, $p = .006$, $r=.307$, showing that noise-vocoded speech caused attention capture. The effect of selectively reversed noise-vocoded speech was not significant, $t(76) = 1.439$, $p=.14$, $r=.162$. Data from this experiment were next converted into difference scores to examine the effect of changing-state complexity and speech fidelity on disruption (see Figure 3).

The ANOVA for the missing-item task failed to show an effect of channel, $F(2, 152)=.841$, $p=.433$, $\eta^2_p=.011$, demonstrating that the effect of channel found in Experiment 2 cannot be attributed to the attention-capture mechanism. This analysis is also consistent with prior studies that have also failed to find a changing-state effect in the missing-item task (e.g., Hughes et al., 2007). Critically, this analysis also failed to find an effect background, $F(1, 76)=2.232$, $p=.193$, $\eta^2_p=.029$, suggesting that performance in the missing-item task was not affected by the speech fidelity of the signal. This indicates that effect of selective-reversal reported for Experiment 2 cannot be attributed to attention capture. This analysis also failed to find an interaction, $F(2, 152)=.443$, $p=.643$, $\eta^2_p=.006$. Collectively, this analysis indicates that even though the presence of noise-vocoded backgrounds disrupts performance in this task, there is no effect of speech fidelity. Thus, the effects observed for the serial recall task from Experiment 2 cannot be attributed to the attention-capture mechanism.

General discussion

The potency of speech as a disruptive background is illustrated in recent reviews (e.g., Ellermeier & Zimmer, 2014; Schlittmeier et al., 2012) as well as recent empirical work (e.g., Viswanathan et al., 2014a). The goal of the research presented here was to understand what makes speech a more disruptive background than non-speech. One explanation is that this effect is solely the result of the greater changing-state complexity for speech relative to non-speech. This hypothesis was tested against an alternative explanation; that the speech fidelity of the background, even when controlled for its overall complexity, makes it particularly disruptive.

To test these explanations we compared the serial recall accuracy in noise-vocoded speech backgrounds. Experiment 1 confirmed that serial recall disruption linearly increased with the number of vocoding channels, a finding which according to the changing-state hypothesis confirms that changing-state complexity increases with channel number. Experiment 2 assessed the roles of changing-state complexity (operationally defined as the number of vocoding channels) and the speech fidelity of the acoustic speech signal. Speech fidelity of the noise-vocoded speech was manipulated by selectively reversing a subset of the vocoding channels. Critically, we found that across different channel conditions, selectively reversed noise-vocoded speech was less disruptive than its normal (non-reversed) noise-vocoded speech counterpart despite sharing the same changing-state complexity (number of channels). This indicates that speech fidelity has an effect on the ISE beyond the overall complexity of the signal. Experiment 3 demonstrated that overall, noise-vocoded backgrounds produce more disruption than silence in a missing-item task. Note the null effect of channel in the missing-item task of Experiment 3, in contrast to the effect of channel on the serial recall task of Experiment 2 indicates that channel number affects the ISE through the interference-by-process mechanism. Thus, these findings are consistent with our conclusion that channel number influences the changing-state complexity of the background. Critically however, the speech-fidelity also did not contribute to performance on the missing-item task. This suggests that the effect of speech fidelity on ISE is not reducible to speech's ability to preferentially capture attention. Instead, the speech structure appears to specifically interfere with the serial rehearsal process. Together the results of these experiments present interesting implications for ISE accounts. The essential prediction of the changing-state hypothesis is that the degree of serial recall disruption will correspond to the number of auditory states in the sound, and as such, speech should only be as disruptive as its changing-state complexity. We operationally defined changing-state complexity as the number of channels in the noise-vocoded speech. Consistent with the predictions of the changing-state hypothesis, our results show that serial recall disruption is related to the number of vocoding channels. However, the results of Experiment 2 show that selectively reversed noise-vocoded speech is less disruptive than typical noise-vocoded speech despite being composed of the same number of vocoding channels. Similar to Viswanathan et al. (2014a), these results highlight a role for speech fidelity and are inconsistent with the changing-state hypothesis that the only driver of the ISE is the changing-state complexity of the signal.

The duplex-mechanism theory for the ISE proposes two mechanisms that can account for the ISE; an interference-by-process mechanism and an attention-capture mechanism. Recall that under this account one mechanism for the ISE is the interference-by-process mechanism which supports the changing-state effect. The interference-by-process mechanism, as outlined in the preceding discussion does not account for the differential disruption between noise-vocoded and selectively reversed noise-vocoded speech. This leaves the attention-capture mechanism to account for effects of speech fidelity. However, this explanation is ruled out by the results of Experiment 3.

While Viswanathan et al. (2014b) used a secondary measure of speech perception to demonstrate that the selective-reversal process disrupts speech fidelity, no such measure was used here, and it is possible that the selective reversal process affects speech fidelity differently in sinewave speech (e.g., Viswanathan et al., 2014a, 2014b) and noise-vocoded speech (i.e., the present study). More importantly, no independent measure exists to determine a signal's changing-state quality. Lacking such an independent measure makes it difficult to determine if the main effect of speech fidelity found in Experiment 2 constitutes an independent effect on the ISE, or if speech fidelity (in addition to channel composition) influences the signal's changing-state quality and that in turn influences the ISE.

Speech fidelity could directly influence the ISE by appealing to phonological interference (e.g., Larsen, Baddeley, & Andrade, 2000); however, this account, and similar accounts, alone are unable to explain the breadth of findings that the changing-state and duplex accounts successfully account for (e.g., see Jones, Macken, & Nicholls, 2004). The 'Perceptual-Motor Account' proposes that subvocal rehearsal converts to-be-remembered items into a perceptual-motor plan that maintains serial order by taking advantage of the inherent serial nature of speech, such as coarticulation (see Hughes & Marsh, 2017). On this account, speech fidelity would disrupt the maintenance of the serial order of the to-be-remember items by introducing a signal with obligatory access to the perceptual-motor plan.

While proponents of the perceptual-motor account indicate that such obligatory access is associated with auditory stimuli generally (see Hughes & Marsh, 2017, p. 3), this proposed mechanism converges with work in the speech literature. Specifically, it has been found that listening to speech produces subtle activity in the listener's articulatory tract that matches the place of articulation of the heard speech (e.g., Fadiga, Craighero, Buccino, & Rizzolatti, 2002; Sundara, Namasivayam, & Chen, 2001). The results of Sundara et al. (2001) are consistent with the perceptual-motor account's prediction that speech affects the ISE through obligatory activation of the articulatory motor system; however, it is unclear how other auditory stimuli such as tones might produce similar effects, or if this effect might be modulated by the selective-reversal process used in the present study. In short, the current results do not completely conform to established theoretical accounts.

To conclude, as reviewed earlier, speech-like signals produce the strongest ISE (Ellermeier & Zimmer, 2014; Schlittmeier et al., 2012; Viswanathan et al., 2014a). The current set of experiments suggests that the potency of speech signals is not solely due to its changing-state complexity and attention-capturing capability. Instead, there appears to be clear need for any account of the ISE to incorporate mechanisms for speech sensitivity that go beyond complexity and the content of speech and consider the dynamic stimulus level structure specific to speech.

Acknowledgements

We thank Matt Davis for his guidance and sharing the script for generating the noise-vocoded speech stimuli used in this study. We also thank the editor and two anonymous reviewers for helpful comments that substantially improved this manuscript.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by NIDCD grant R15 DC011875-01 to NV and NSF grant 1632530 to LDR.

Supplemental material

Supplementary material is available at journals.sagepub.com/doi/suppl/10.1177/1747021817739257.

Notes

1. Words used in Buchner et al. (2004) were actually presented in German; the native language of the participants in the study. In the present discussion, we have provided the English translation for ease of reading.
2. The trend analysis was conducted because the overall pattern across the channel conditions was the pattern of interest to the current study. However, it is worth noting that the Bonferroni corrected paired comparisons were consistent with this finding; the 9 channel condition was neither different from the 6 or 12 channel conditions. These comparisons did, however, reveal that despite both differing by 6 channels, the 3 to 9 channel comparison but not the 6 to 12 channel comparisons was significant.
3. The unpublished thesis by which the present work builds on reported an intelligibility measurement for Experiment 1. The responses to the measurement were too low to influence the results of the current report; indicating fewer than 1 word was accurately perceived by any participant. However, as some have pointed out (e.g. Ellermeier et al., 2015) these low intelligibility scores are interesting as the 9 and 12 channel conditions are within a range of noise-vocoded speech generally found to be easily intelligible. Indeed, these samples were easily intelligible to the authors and collaborators of the current study, and it is not readily apparent why participants had such low intelligibility scores. Ultimately, this null finding will require additional investigation that is beyond the purview of the current work.
4. We thank anonymous Reviewer 1 for bringing this limitation to our attention.
5. This experiment was originally run using a between subjects design similar to what is reported for Experiment 1. We reran this experiment using a within subjects comparison to eliminate group differences. It is noteworthy that both versions of Experiment 2 produce the same pattern of effects.
6. We thank Reviewer 2 who provided several valuable criticism that shaped Experiment 3. Most notably, this reviewer recommended the use of the missing-item task as a way of dissociating the interference-by-process and attention-capture mechanism.

References

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. *Psychology of Learning and Motivation*, *8*, 47–89.

Buchner, A., Rothermund, K., Wentura, D., & Mehl, B. (2004). Artificially induced valence of distractor words increases the effects of irrelevant speech on serial recall. *Memory & Cognition*, *34*, 1055–1062. doi:10.3758/BF03195862

Colle, H. A., & Welsh, A. (1976). Acoustic masking in primary memory. *Journal of Verbal Learning and Verbal Behavior*, *15*, 17–31. doi:10.1016/S0022–5371(76)90003–7

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, *134*, 222–241. doi:10.1037/0096–3445.134.2.222

Ellermeier, W., Kattner, F., Ueda, K., Doumoto, K., & Nakajima, Y. (2015). Memory disruption by irrelevant noise-vocoded speech: Effects of native language and the number of frequency bands. *The Journal of the Acoustical Society of America*, *138*, 1561–1569. doi:10.1121/1.4928954

Ellermeier, W., & Zimmer, K. (2014). The psychoacoustics of the irrelevant sound effect. *Acoustical Science and Technology*, *35*, 10–16. doi:10.1250/ast.35.10

Elliott, E. M., & Briganti, A. M. (2012). Investigating the role of attentional resources in the irrelevant speech effect. *Acta Psychologica*, *140*, 64–74.

Elliott, E. M., Hughes, R. W., Briganti, A., Joseph, T. N., Marsh, J. E., & Macken, B. (2016). Distraction in verbal short-term memory: Insights from developmental differences. *Journal of Memory and Language*, 88(2016), 39–50.

Fadiga, L., Craighero, L., Buccino, G., & Rizzolatti, G. (2002). Speech listening specially modulates the excitability of tongue muscles: A TMS study. *European Journal of Neuroscience*, *15*, 399–402.

Hughes, R. W. (2014). Auditory distraction: A duplex-mechanism account. *Psych Journal*, *3*, 30–41. doi:10.1002/pchj.44

Hughes, R. W., & Marsh, J. E. (2017). The functional determinants of short-term memory: Evidence from perceptual-motor interference in verbal serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 537–551.

Hughes, R. W., Vachon, F., & Jones, D. M. (2007). Disruption of short-term memory by changing and deviant sounds: support for a duplex-mechanism account of auditory distraction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 1050–1061. doi:10.1037/0278–7393.33.6.1050

Jones, D. M. (1993). Objects, streams and threads of auditory attention. In A. D. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness and control: A tribute to Donald Broadbent* (pp. 87–104). Oxford, UK: Oxford University Press.

Jones, D. M., & Macken, W. J. (1993). Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 369–381. doi:10.1037/0278–7393.19.2.369

Jones, D. M., Macken, W. J., & Nicholls, A. P. (2004). The phonological store of working memory: is it phonological and is it a store? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*, 656–674. doi:10.1037/0278–7393.30.3.656

Jones, D. M., Miles, C., & Page, J. (1990). Disruption of proofreading by irrelevant speech: Effects of attention, arousal or memory? *Applied Cognitive Psychology*, *4*, 89–108.

Larsen, J. D., Baddeley, A., & Andrade, J. (2000). Phonological similarity and the irrelevant speech effect: implications for models of short-term verbal memory. *Memory*, *8*, 145–157. doi:10.1080/096582100387579

Loizou, P. C., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *The Journal of the Acoustical Society of America*, *106*, 2097–2103.

Macken, B. (2014). Auditory distraction and perceptual organization: Streams of unconscious processing: Distraction and perceptual organization. *PsyCh Journal*, *3*, 4–16. doi:10.1002/pchj.46

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carell, T. D. (1981). Speech perception without traditional speech cues. *Science*, *212*, 947–950. doi:10.1126/science.7233191

Röer, J. P., Bell, R., & Buchner, A. (2013). Self-relevance increases the irrelevant sound effect: Attentional disruption by one's own name. *Journal of Cognitive Psychology*, *25*, 925–931. doi:10.1080/20445911.2013.828063

Salamé, P., & Baddeley, A. (1982). Disruption of short-term memory by unattended speech: Implications for the structure of working memory. *Journal of Verbal Learning and Verbal Behavior*, *21*, 150–164.

Salamé, P., & Baddeley, A. (1987). Noise, unattended speech and short-term memory. *Ergonomics*, 30(8), 1185–1194. doi:10.1080/00140138708966007

Salamé, P., & Baddeley, A. (1989). Effects of background music on phonological short-term memory. *The Quarterly Journal of Experimental Psychology*, *41*, 107–122. doi:10.1080/14640748908402355

Schlittmeier, S. J., Weissgerber, T., Kerber, S., Fastl, H., & Hellbrück, J. (2012). Algorithmic modeling of the irrelevant sound effect (ISE) by the hearing sensation fluctuation strength. *Attention, Perception & Psychophysics*, *74*, 194–203. doi:10.3758/s13414-011-0230-7

Shannon, R. V., Zeng, F., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303–304.

Sundara, M., Namasivayam, A. K., & Chen, R. (2001). Observation-execution matching system for speech: A magnetic stimulation study. *NeuroReport*, *12*, 1341–1344. doi:10.1097/00001756

Tremblay, S., Nicholls, a P., Alford, D., & Jones, D. M. (2000). The irrelevant sound effect: does speech play a special role? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *26*, 1750–1754. doi:10.1037/0278-7393.26.6.1750

Viswanathan, N., Dorsi, J., & George, S. (2014a). The role of speech-specific properties of the background in the irrelevant sound effect. *Quarterly Journal of Experimental Psychology*, *67*, 581–589. doi:10.1080/17470218.2013.821708

Viswanathan, N., Magnuson, J. S., & Fowler, C. A. (2014b). Information for coarticulation: Static signal properties or formant dynamics? *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 1228–1236. doi:10.1037/a0036214

Wöstmann, M., & Obleser, J. (2016). Acoustic detail but not predictability of task-irrelevant speech disrupts working memory. *Frontiers in Human Neuroscience*, *10*, Article 538. doi:10.3389/fnhum.2016.00538