Determinants of the irrelevant speech effect: Changes in spectrum and envelope

Josef Schlittenlacher,[a] Katharina Staab,[b] Özlem Çelebi, Alisa Samel,[c] and Wolfgang Ellermeier
Institut für Psychologie, TU Darmstadt, Alexanderstraße 10, 64283 Darmstadt, Germany

(Received 8 November 2018; revised 21 May 2019; accepted 28 May 2019; published online 20 June 2019)

The irrelevant sound effect (ISE) denotes the fact that short-term memory is disrupted while being exposed to sound. The ISE is largest for speech. The presented study investigated the underlying acoustic properties that cause the ISE. Stimuli contained changes in either the spectral content only, the envelope only, or both. For this purpose two experiments were conducted and two vocoding strategies were developed to degrade the spectral content of speech and the envelope independently. The first strategy employed a noise vocoder that was based on perceptual dimensions, analyzing the original utterance into 1, 2, 4, 8, or 24 channels (critical bands) and independently manipulating loudness. The second strategy involved a temporal segmentation of the signal, freezing either spectrum or level for durations ranging from 50 ms to 14 s. In both experiments, changes in envelope alone did not have measurable effects on performance, but the ISE was significantly increased when both the spectral content and the envelope varied. Furthermore, when the envelope changes were uncorrelated with the spectral changes, the effect size was the same as with a constant-loudness envelope. This suggests that the ISE is primarily caused by spectral changes, but concurrent changes in level tend to amplify it.
https://doi.org/10.1121/1.5111749

[JJL] Pages: 3625–3632

I. INTRODUCTION

The performance in short-term-memory tasks decreases when subjects simultaneously listen to sounds (Colle and Welsh, 1976; Salame and Baddeley, 1982). This effect is called the irrelevant sound effect or irrelevant speech effect (ISE; for reviews see Ellermeier and Zimmer, 2014; Hughes, 2014), the latter term emphasizing the fact that the effect is strongest when the presented sound is speech, and it occurs even when the subjects are told to ignore the sound that—in fact—is irrelevant to the memory task they are asked to perform. A typical task consists of recalling the order of a sequence of visually presented digits. A pervasive question of the research on irrelevant speech has been which features of speech do in fact make a sound disruptive in terms of its interference with short-term memory (e.g., Schlittmeier *et al.*, 2012; Park *et al.*, 2013). The present study attempts to answer that question by investigating the contribution of changes in spectrum versus changes in envelope by gradually degrading speech signals on either dimension. Here, "changes in envelope" shall refer to the changes in total level or loudness over time, such as would be caused by sinusoidal amplitude modulation but can also be irregular as caused by the syllables of speech that do not have a constant duration. "Changes in spectrum" shall refer to changes in relative loudness or intensity over time between frequency bands, such as auditory filters or critical bands, without affecting the total loudness or level. Thus, it addresses changes of the spectral composition.

Recent theorizing (Hughes *et al.*, 2005, 2007) has postulated two kinds of mechanisms by which irrelevant-sound interference comes about: (1) Interference-by-process resulting from the automatically processed irrelevant sound obstructing the serial rehearsal of the to-be-remembered material, and (2) attentional capture by some unexpected feature in the irrelevant sound. While the latter process is somewhat sensitive to semantic properties of the irrelevant stream (Neely and LeCompte, 1999), such as mentioning the listener's name (Röer *et al.*, 2013), or pronouncing taboo words (Röer *et al.*, 2017), the former process—observed in the classical serial-rehearsal task—may be considered largely acoustical in nature. That is underscored, for example, by the fact that the ISE is typically of equal magnitude for an unknown foreign language as for one's native language (Colle and Welsh, 1976; Ellermeier and Zimmer, 1997; Ellermeier *et al.*, 2015), thereby suggesting a focus on the acoustic properties of sound to explain the automatically occurring interference by process in the ISE.

It is well established that the overall level of a sound has no significant effect on the ISE (Colle, 1980; Ellermeier and Hellbrück, 1998), with error rates being the same for moderate as for high sound pressure levels of the distractor. Various continuous noises including stationary noise with

[a]Present address: Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, United Kingdom. Electronic mail: js2251@cam.ac.uk
[b]Present address: Department of Marketing and Human Resource Management, TU Darmstadt, Hochschulstraße 1, 64289 Darmstadt, Germany.
[c]Present address: Klinik für Kinder−und Jugendpsychatrie, Phillipps−Universität Marburg, Hans-Sachs-Straße 4, 35039 Marburg, Germany.

the average spectrum of speech produced the same error rates in the recall task as silence did (e.g., Liebl *et al.*, 2016), suggesting that particular frequencies in the absence of any temporal changes do not selectively impair short-term memory either.

Considering these findings, it seems evident that it is the time-varying nature of speech (or irrelevant sound in general) that causes the ISE. That is the central claim of the original theorizing about the "changing-state effect" (Jones *et al.*, 1992) as well as its recent reconceptualization as the "interference-by-process" mode of auditory distraction (Hughes *et al.*, 2005, 2007). Identifying a specific psychoacoustic sound feature as being at the root of what is perceived as "changing state," Schlittmeier *et al.* (2012) recently suggested the magnitude of the ISE to be proportional to psychoacoustical fluctuation strength (Fastl, 1983), i.e., the amount of fluctuation perceived in the irrelevant sound. Fluctuation strength captures both changes in envelope and spectrum, and produces a maximum for modulation frequencies around 4 Hz, which roughly corresponds to the rate of syllables per second in speech. Despite its simplicity, the model of Schlittmeier *et al.* (2012) has predicted the ISE of various sounds correctly, with at least two notable exceptions: amplitude-modulated (AM) or frequency-modulated (FM) sounds with a constant modulation frequency. A periodic pattern in a synthetic sound may lead to the perception of fluctuation, but impairs short-term memory only marginally. Ellermeier and Zimmer (2014; Fig. 2), for example, reported the error rate for an uninterrupted FM tone as a function of modulation frequency. The inverted-V pattern mimicked the rise and fall of fluctuation strength as a function of modulation frequency, but the maximum error rate was still close to that obtained for silence.

Thus, the changing state property causing the ISE is likely to consist of irregular changes in spectrum and envelope, or both. The questions remain: which of these two contributes more strongly, and are the effects of irregular AM and FM additive or subject to interactions? Slow changes at a rate below 2 Hz have not shown any effect for irregular FM (Jones *et al.*, 1993) or for short words of varying level (Tremblay and Jones, 1999). The latter study used spoken integers as words, which depending on the condition were always presented at the same level or randomly changed in level between 55 and 85 dB(A). Likewise, employing AM on a constant spectrum yielded negligible effects on the error rate, even if the modulation was taken from the envelope of speech (e.g., Salamé and Baddeley, 1989). By contrast, randomly ordered tones varying in frequency but having a constant level produced a large increase in error rate (Jones and Macken, 1993). Typically, however, irrelevant sound effects produced with varying tones or other non-speech signals are considerably smaller than those obtained with speech (see Ellermeier and Zimmer, 2014). Some studies have used the opposite strategy, of degrading free-running speech successively, eventually generating a noise-like sound that produces no more disruption (Ellermeier *et al.*, 2015; Wöstmann and Obleser, 2016; Senan *et al.*, 2018a). These studies degraded the spectral content of speech by using a vocoder (Dudley, 1939) and varying the number of channels, with noise bands as carrier signals. For 20 channels, the error rate was the same as for original speech, while for one channel, i.e., with no changes in the spectrum and thus equivalent to Salamé and Baddeley (1989), the error rate was similar to that for silence. In-between, Ellermeier *et al.* (2015), for example, demonstrated a gradual increase in error rate with an increase in the number of vocoder channels, or greater spectral variations in speech. However, by using a typical vocoder, the changes in the envelope were always the same as in the original speech signal, and the study design therefore did not show whether the effect was amplified by or independent of the simultaneous AM.

For this reason the present study used vocoding strategies that allowed for controlling the amount of change in spectrum and envelope as independently as possible. Here and hereafter, a change in envelope refers to an amplification or attenuation of all frequency bands over time by the same amount (relative to a stationary sound with a given spectrum). Likewise, a change in spectrum means that calculated loudness (experiment 1) or A-weighted level (experiment 2) remained constant over time while the spectral composition, i.e., the relative loudness or level of the frequency bands, changed. That was achieved by interpreting momentary specific loudness as a percentage of total loudness (experiment 1) or by scaling levels of frequency bands to achieve a desired constant total A-weighted level (experiment 2).

II. METHOD

Two experiments were conducted. In experiment 1, the envelope of noise-vocoded speech was either preserved as in the original speech signal, normalized to have constant loudness, or modified to have the envelope of another speech signal. The number of channels of the vocoder was varied for each of these three conditions. In experiment 2, the duration to determine the overall level, the third-octave spectrum, or both were varied. The signal was then synthesized from noise bands based on the thus extracted information.

The stimuli were derived by degrading speech. An alternative approach to obtain changes in spectrum and envelope over time could have been irregular AM or FM of pure tones or other simple synthetic stimuli. However, this would have resulted in a much simpler spectral composition than speech, and that is why "creative" ways to degrade speech that exhibit according perceptual changes over time, were tried in this study.

A. Subjects

Fifty-five subjects participated in experiment 1 (39 females and 16 males, 18 to 51 years, median 23 years). Forty subjects participated in experiment 2 (32 females and 8 males, 19 to 46 years, median 25 years). Most of them participated in both experiments, though not necessarily in the order of doing experiment 1 first and experiment 2 second. All subjects reported having normal hearing. Most subjects were students of psychology at TU Darmstadt and participated for course credit, the remaining were friends of the experimenters and participated voluntarily without compensation. No attention was paid to the age since the design was

within-subject, we assumed the effects to occur independently of age. Furthermore, the ISE had been reported to be independent of age (Rouleau and Belleville, 1996).

B. Apparatus

The experiments were run in a sound-proof booth. Stimuli were converted from digital to analogue form by a RME Hammerfall DSP Multiface II (Audio AG, Haimhausen, Germany), passed through a Behringer Powerplay Pro-8 HA8000 headphone amplifier (Willich, Germany) and presented diotically via Beyerdynamics DT-990 Pro headphones (Heilbronn, Germany).

C. Stimuli

For each experiment, a specific vocoder was implemented in Matlab. It used bandpass-filtered noises, which were generated using ArtemiS software (Head Acoustics, Herzogenrath, Germany). All bandpass filters were 6th-order Butterworth filters, and limiting frequencies denote the 3-dB cutoff points.

The stimuli were chosen from monophonic recordings of four speakers (two female and two male), who read long passages of meaningful text, both fictional, and non-fictional. These recordings were made in our lab, with native speakers of German reading German text. The primary language of the subjects was also German. 240 segments (60 per speaker) with a duration of 14 s and a separation of 4 s between segments were cut out of the recordings. The original-speech stimuli were chosen from this set, ten for experiment 1 and another ten for experiment 2. The sets of stimuli for the two experiments did not overlap. Stimuli were allowed to be from the same passage of text, but at least 18 s of reading time were in between the stimuli that were finally chosen. The vocoded stimuli were derived from these original-speech stimuli. The stimuli had energy-equivalent levels between 60 and 65 dB(A), a duration of 14 s and rise and fall times of 20 ms.

1. Experiment 1

Calculations of momentary loudness and momentary specific loudness were performed for the original-speech stimuli according to DIN 45631/A1 using ArtemiS audio analysis software, including 2 s before and after the 14-s excerpt to avoid low "wrong" loudness readings during the first milliseconds of the actual stimulus to be presented. Values of momentary loudness (sometimes also called "instantaneous loudness") and momentary specific loudness were sampled every 2 ms and these values were used to manipulate the speech stimuli. There was no windowing of such a short duration. DIN 45631/A1 uses several bandpass and low-pass filters, with time constants depending on frequency. Furthermore, bandpass-filtered noise spectra with a duration of 14 s each were derived from pink noise. These included one 24-Bark-wide noise, two 12-Bark-wide noises, four 6-Bark-wide noises, eight 3-Bark-wide noises, and 24 1-Bark-wide noises. The limiting frequencies were taken from the critical-band scale (Zwicker 1961, Table I), and the bands generated for a given vocoding condition were adjacent. For example, the first of the four 6-critical-band wide noises generated for the 4-channel vocoding condition consisted of bands no. 1 to 6 (20 to 630 Hz), the second of bands no. 7 to 12 (630 to 1720 Hz), the third of bands no. 13 to 18 (1720 to 4400 Hz), and the fourth of bands no. 19 to 24 (4400 to 15 500 Hz). The bandpass-filtered noises were stored as 24-bit wav files. The vocoder (see next two paragraphs) modified the levels of these bandpass-filtered noises, varying over time, and added them to yield the vocoded stimulus.

The vocoder was intended to be a "perceptual vocoder" that used loudness to determine the envelope and specific loudness to determine the spectral content. In order to adjust the level of a bandpass-filtered pink noise to a target loudness, tables relating its loudness in sone to sound pressure level were calculated beforehand. A schematic representation of the vocoder is given in Fig. 1 for the example of a four-band vocoder.

After the preprocessing had been done in ArtemiS software, the actual vocoder was implemented in Matlab. The levels of the bandpass-filtered noises were adjusted to match their target loudness at a rate of 2 ms. The target loudness was given by the sum of specific loudness over the critical bands corresponding to the bandpass-filtered noise, and the total target loudness. Three different strategies were used to obtain the target loudness. (1) "Original loudness": The target loudness of each bandpass-filtered noise was directly given by the sum over momentary specific loudness of the respective critical bands of the original speech signal. This is very similar to a conventional noise vocoder. (2) "Constant loudness": The target loudness of the vocoded signal was given by the overall loudness of the speech signal, i.e., it was constant across the 14-s duration. The measure used for overall loudness was the LL_P (ISO, 2017; Schlittenlacher *et al.*, 2017). The contribution of each bandpass-filtered noise in sone was given by the percentage of the momentary specific loudness of the corresponding critical bands in the speech signal. For example, the band stretching from 20 to 630 Hz could have contributed 23% of the total loudness at a given moment. If the overall loudness was 10 sone, the level of the bandpass-filtered noise would have been adjusted to result in a loudness of 2.3 sone at the given moment. Thus, the total loudness remained constant, but the spectral content varied over time. (3) "Uncorrelated loudness": Determining the percentage which a band contributed to total loudness was done as previously, but the total loudness was now given by the momentary loudness of another speech signal. Thus, the momentary spectral content and momentary loudness were determined by two different speech signals. The variations over time were speech-like for both variables, but they were uncorrelated.

For example (and using fictional loudness values for this example), if a 4-band vocoded stimulus were to have a total loudness of 10 sone at time $t = 1432$ ms, and the first of these four bands spanning the first six Bark (20–630 Hz) in the original speech made up 20% of total loudness at this point in time, the bandpass-filtered noise from 20 to 630 Hz was adjusted to a level such that its specific loudness from 20 to 630 Hz summed to 2 sone (i.e., 20% of the 10 sone). The

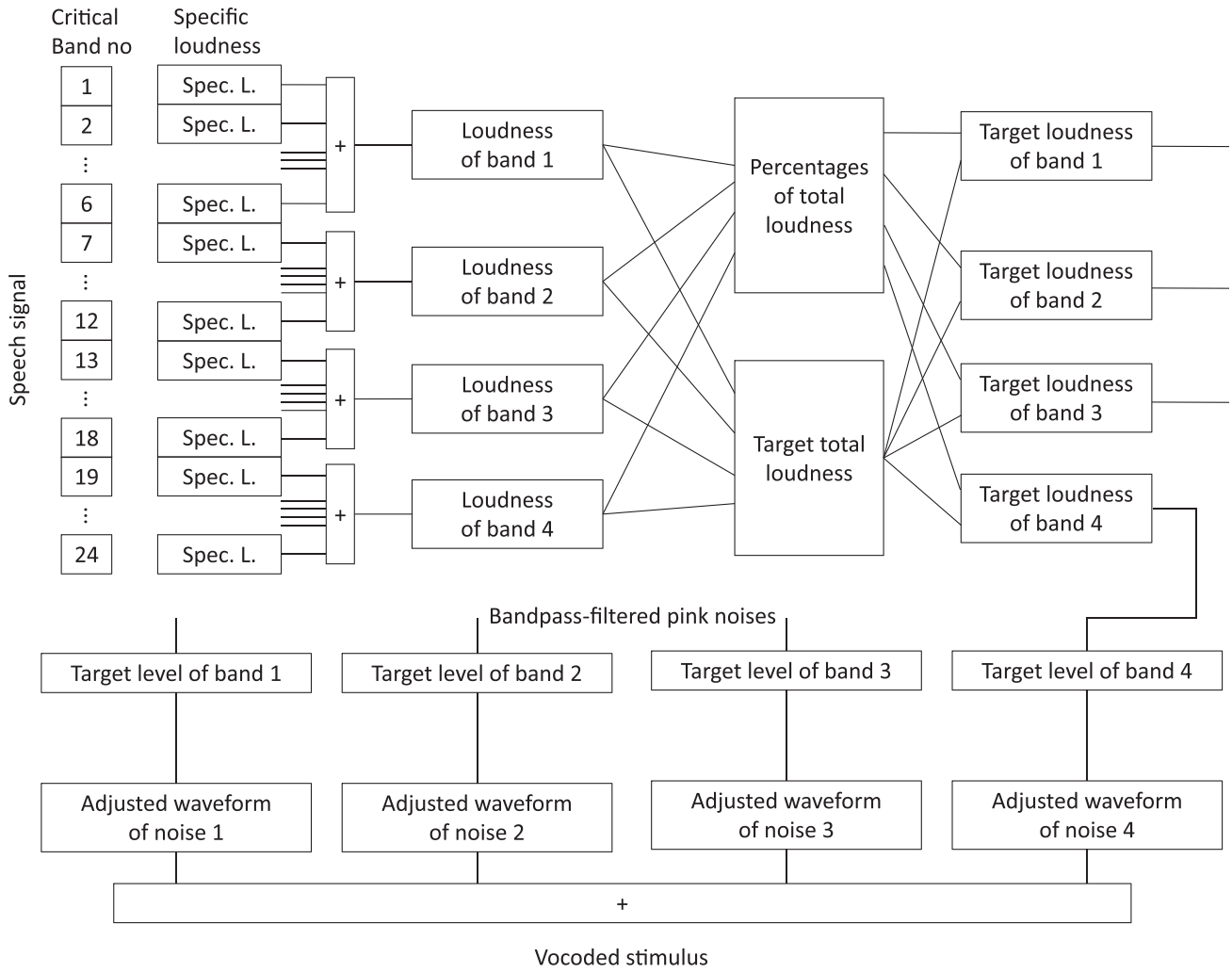

FIG. 1. "Perceptual vocoder." The schema shows the 4-band vocoder as an example. The target total loudness was determined by three different strategies. Details are given in the text. The differences to a common noise vocoder are the use of perceptual scales (critical bands, loudness), and the control of degradation in both of them, not only in the spectrum.

same procedure was repeated for the three other bands and corresponding bandpass-filtered noises (630–1720 Hz, 1720–4440 Hz, and 4440–15 500 Hz). This procedure was repeated each 2 ms, i.e., the next adjustment of level in each bandpass-filtered noise happened at $t = 1434$ ms. After the levels of all bandpass-filtered noises were adjusted at all points in time, they were added.

It should be noted that the conditions for uncorrelated loudness and constant loudness contained some clearly audible artefacts. When the original stimulus was very soft due to a pause made by the speaker, the spectrum was quasi-random and was amplified into the audible range. The actual speech content, i.e., phonemes, were not affected by this artefact. Further distortions like in a fast-acting hearing aid were not noticed, probably because the transformation to noise, i.e., the vocoder itself, dominated these distortions.

2. Experiment 2

The vocoder for this experiment was intended to control the variation of the spectral content and of the amplitude envelope by keeping them frozen for different lengths of time, thus resulting in what might be called "pixelated speech." By doing so, the spectral resolution and the changes in level were determined by segment duration, and not by the number of frequency bands, as in experiment 1. Non-overlapping discrete Fourier transformations (DFTs) with window lengths of 50, 100, 200, 500, or 14 000 ms were calculated for each speech sample to be processed. The DFT spectra were used to determine third-octave levels and the A-weighted equivalent sound pressure level (L_{Aeq}) for each segment.

Vocoded stimuli were generated by manipulating the level of third-octave wide noises, whose waveforms were subsequently added, using three different strategies. The center frequencies of the 27 third-octave wide noises ranged from 50 to 20 000 Hz. (1) Varying spectrum, varying L_{Aeq}: The levels of the third-octave noises were determined by the third-octave levels of the original speech signal for segment lengths of 50, 100, 200, 500, or 14 000 ms. The third-octave levels were frozen within a given segment, and abruptly changed between segments. The segment duration of 14000 ms resulted in a stationary noise that had the average spectrum of speech. (2) Constant spectrum, varying L_{Aeq} only: Third-octave levels were determined for a DFT length

of 14 s, i.e., the whole stimulus. The L_{Aeq} was determined for segment lengths of 50, 100, 200, or 500 ms, and the constant spectrum was amplified or attenuated over time to match each segment's target L_{Aeq}. (3) Varying spectrum, constant L_{Aeq}: The L_{Aeq} was determined for a segment duration of 14 s, i.e., based on the entire stimulus. Third-octave levels were determined for segment lengths of 50, 100, 200, or 500 ms, thus resulting in a sound of constant L_{Aeq} but subject to step-wise changes in spectral composition at constant intervals. A schematic representation of the vocoder is given in Fig. 2.

D. Procedure

A trial started with a blue square on the screen that became smaller and disappeared after 2 s. After this visual warning signal, the "irrelevant" background sound was played for 14 s. During the first 9 s of sound playback, the digits between 1 and 9 were presented in random order on the screen. Each digit was displayed for 1 s, and was not repeated. The task for the subject was to memorize the order of the digits and to ignore the sound. After the sound terminated, buttons depicting the digits 1 to 9 were displayed in a keyboard-like 3×3 matrix in a fixed layout. The subject could click on each digit only once, and had to click each digit to report the memorized order. The dependent variable was defined as how many of the nine digits of a given trial were recalled at the correct position.

1. Experiment 1

Experiment 1 used 15 irrelevant-sound conditions. Conditions 1 to 5 ("original envelope") were vocoded with resolutions of 1, 2, 4, 8, and 24 bands, respectively. The momentary loudness was given by the original momentary loudness of the speech signal. Conditions 6 to 10 ("constant loudness") were vocoded using the same numbers of bands as conditions 1 to 5, but the momentary loudness was kept constant at the overall loudness (LL_P) of the original speech signal. Conditions 11 to 13 ("uncorrelated envelope") were vocoded with resolutions of 2 bands, 4 bands, and 24 bands. Their loudness was adjusted to the momentary loudness of another speech signal while the percentages of the spectral contributions of each band were determined by the original speech signal, i.e., spectral content and loudness were uncorrelated. 1 band was not used for the uncorrelated-envelope because that would have resulted in a speech-like modulated constant spectrum, perceptually the same as the 1-band condition with the "original envelope." Eight bands were not used for the "uncorrelated envelope" condition either. Condition 14 was stationary pink noise, and condition 15 was the original speech recording.

The 150 trials resulting from 15 conditions and ten different underlying original-speech samples per condition were divided in two sessions, each having 75 trials. The order of trials was random and different for each participant. Each session began with an additional three practice trials. Participants received feedback about their percentage correct after every ten trials, and could use this opportunity to take a rest.

2. Experiment 2

Experiment 2 used 14 irrelevant sound conditions. For conditions 1 to 4 (varying spectrum, constant L_{Aeq}), the spectral content changed every 50, 100, 200, or 500 ms, respectively, while the A-weighted level was kept constant. For conditions 5 to 8 (varying L_{Aeq}), the A-weighted level changed each 50, 100, 200, or 500 ms while the spectrum was determined by the average spectrum of the entire 14-s utterance, i.e., it was fixed and basically a speech-shaped broadband noise. For conditions 9 to 12 (varying spectrum, varying L_{Aeq}), both spectrum and level changed each 50, 100, 200, or 500 ms. Condition 13 was stationary noise, with both spectrum and level being determined by the whole 14 s. Condition 14 was the original speech recording.

The procedure was equivalent to experiment 1, with ten different underlying original-speech samples per condition, two sessions, the same amount of practice, and same procedure for feedback and opportunities for rest.

III. RESULTS

Numbers of digits correctly recalled were averaged across trials and subjects using the arithmetic mean. Figure 3 shows the results of experiment 1. It can be seen that no matter what the envelope manipulation was, the number of digits correctly recalled decreased with the number of bands of the vocoder (i.e., the better the spectral detail was preserved). Fewer digits were recalled when the envelope of the original sound was preserved (solid line) than when the envelope was set to a constant loudness (dashed line). Performance was even better when the envelope and spectral content were determined by different speech signals (dashed-dotted line).

The main part of the experiment consisted of conditions 1 to 10, i.e., the variation of the number of vocoder bands, either with the original envelope in a given frequency band, or with constant loudness. A 5×2, bands (1,2,4,8,24) $\times$ envelope (original, constant), within-subjects analysis of variance yielded a statistically significant main effect of the number of vocoder bands, $F(4,216) = 67.9$, $p < 0.001$ and a

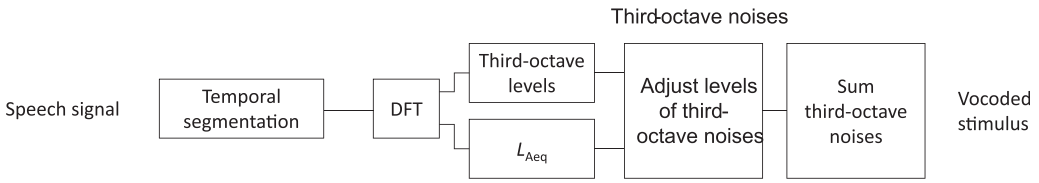

FIG. 2. "Pixelated speech" vocoder. A speech signal is split into several temporal segments, during which spectrum and/or level are frozen. The resolution of this vocoder is determined by the segment duration, not by the number of bands, and it can degrade both spectrum and envelope in the temporal domain. Details are given in the text.

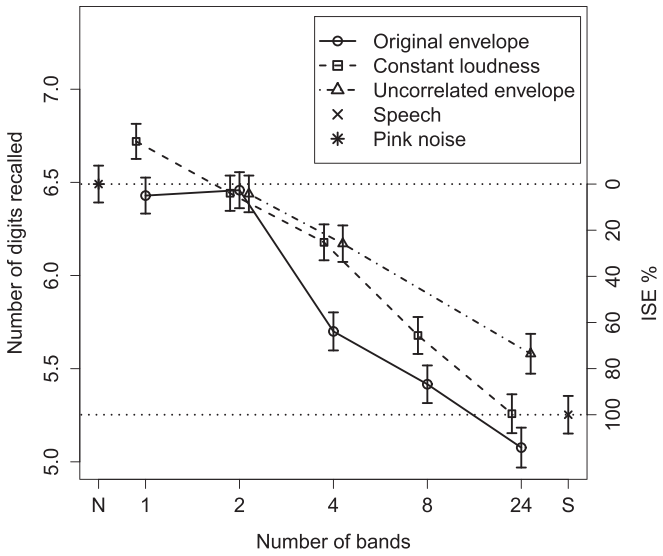

FIG. 3. Results of experiment 1. The average number of digits recalled (maximum 9) is shown as a function of the number of bands used in the vocoder. Circles connected by a solid line show conditions with the original envelope, squares connected by a dashed line conditions having a constant loudness, and triangles connected by a dashed-dotted line conditions where the envelope was taken from another speech signal but uncorrelated to the spectral content. The cross shows the original-speech condition, the asterisk pink noise. Error bars denote standard errors of the mean. The scale on the right hand side shows the ISE as the performance decrement in percent, referenced to the stationary pink noise condition (0% ISE) and playback of the original speech (100% ISE).

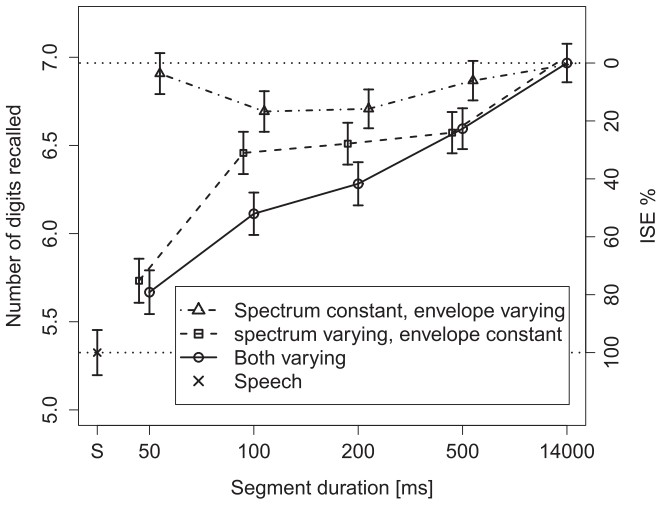

FIG. 4. Results of experiment 2. Average number of digits recalled as a function of segment duration in "pixelated speech." The circles on the solid line show conditions where both spectral content and envelope varied between segments, the squares on the dashed line those, where only the spectral content varied, and the triangles on the dashed-dotted line show conditions in which only the envelop varied. The original speech condition is shown by a cross. Error bars denote the standard error of the mean. The scale on the right hand side shows the ISE as a percentage between a stationary noise resulting in a 0% performance decrement, and original speech producing a 100% ISE.

statistically significant main effect of envelope, $F(1,54) = 24.7$, $p < 0.001$. The interaction between the two was not statistically significant, $F(4,216) = 2.32$, $p = 0.058$.

The three conditions with the envelope being uncorrelated to the spectral changes (dashed-dotted line in Fig. 3) had been added in the experimental design to check whether they produced a smaller ISE than the original envelope. Surprisingly, they even produced a slightly smaller error rate than the constant-loudness envelope. However, this difference just missed statistical significance according to a 3×2, bands (2,4,24) $\times$ envelope (constant, uncorrelated), analysis of variance, $F(1,54) = 3.50$, $p = 0.067$, and neither was the interaction between bands and envelope, $F(2,108) = 2.36$, $p = 0.099$.

On average, 0.2 digits more were recalled in the one-band constant-loudness condition (speech-shaped noise) than with pink noise, despite both of them being stationary noises. However, this difference was not statistically significant in a two-tailed paired t-test, $t(54) = 1.66$, $p = 0.103$. The difference between the one-band original-envelope condition and pink noise also was not statistically significant, $t(54) = 0.46$, $p = 0.65$, suggesting that changes in envelope alone did not increase the error rate.

The results of experiment 2 are shown in Fig. 4. In general the number of digits recalled increased as the "frozen" segments got longer, i.e., as speech became more "pixelated" or more and more degraded. An exception is given by the condition in which only the amplitude envelope varied from segment to segment, while the spectral content remained constant: Here, performance is hardly affected at all (triangles, dashed-dotted line).

Overall, performance disruption ranged from about 25% of the maximal ISE with a segment duration of 500 ms to about 80% for a segment duration of 50 ms, that is—for the conditions in which the spectrum, or both spectrum and level varied—the segment lengths chosen produced effects that were monotonic with segment length and covered almost the entire range of potential effects. Up to 0.5 digits more were recalled on average when only the spectral content varied (squares, dashed line) than when both spectral content and level varied between segments (circles, solid line).

The outcomes thus characterized were confirmed by a within-subject 4×3, segment duration (50, 100, 200, 500 ms) $\times$ variation (constant level, constant spectral content, both varying), within-subjects analysis of variance. The main effect of segment duration was statistically significant, $F(3,117) = 15.4$, $p < 0.001$, as were the main effect of variation, $F(2,78) = 27.6$, $p < 0.001$, and the interaction between the two, $F(6,234) = 8.10$, $p < 0.001$.

The conditions involving changes in level only (triangles, dashed-dotted line) were not statistically different as a function of segment length, as is shown by a one-way, within-subject (segment duration; 50, 100, 200, 500 ms) analysis of variance, $F(3,117) = 1.49$, $p = 0.22$. By contrast, the difference between the conditions with varying spectral content only (dashed line) and additionally varying level (solid line) was statistically significant as confirmed by a within-subject 4×2, segment duration $\times$ variation, analysis of variance which produced a significant effect of variation, $F(1,39) = 4.38$, $p = 0.043$, and no significant interaction.

IV. DISCUSSION

Two different vocoder strategies were used to degrade the spectrum and envelope of free-running speech independently: One strategy operating in the spectral domain (experiment 1), the other one in the temporal domain (experiment 2). Their

effects shall be discussed with respect to: (A) spectral noise vocoding, (B) freezing speech segments in spectrum or level, i.e., "pixelated speech," (C) the role of spectral and level changes, and (D) potential predictors for the magnitude of irrelevant speech effects.

A. Spectral noise vocoding

Experiment 1 shows a substantial effect of the number of frequency channels used in degrading the original speech recordings (Fig. 3), much like in similar, earlier work employing conventional vocoding by preserving the amplitude envelope in each channel (e.g., Dorsi, 2013; Ellermeier *et al.*, 2015; Wöstmann and Obleser, 2016; Senan *et al.*, 2018a): Notably, the effect of the number of channels continues beyond what is required for decent speech intelligibility (4 noise-vocoder channels already yield some 80% correct, see Ellermeier *et al.*, 2015). That suggests that spectral-energy changes between critical bands (corresponding to our 24-band condition) are most important for explaining the effect of frequency changes on the impairment of short-term memory performance. Any further spectral resolution—as given in the present "original-recordings" condition—might just make additional short-range frequency modulations or subtle pitch changes available, contributing to enhancing the quality of the sound, but not to its processing as (interfering) speech. Consequently, no statistically significant performance difference was observed between the 24-channel condition and free-running (original) speech.

As regards intermediate conditions, the four-band condition with the original envelope of experiment 1 exhibited an ISE of 64%, with the error rate expressed as a percentage of the maximal performance decrement observed when comparing stationary noise with speech. This is comparable to other vocoder studies. With four bands, both Ellermeier *et al.* (2015) and Senan *et al.* (2018a) found ISEs of approximately 60% of the range between silence and speech.

B. Freezing speech segments: "pixelated speech"

In experiment 2, freezing the spectrum, or spectrum and level/loudness for speech segments increasing in length from 50 to 500 ms, monotonically improved serial recall (Fig. 4), i.e., the more "pixelated" the speech signal became, the less it affected cognitive performance. That may also be interpreted as a "dosage" effect, reflecting the number of changes per time. Bridges and Jones (1996) showed that the error rate increases with the number of words presented in the irrelevant sound stream per time unit. Similarly, the number of changes was varied in the present study, and error rate increased. However, it should be kept in mind that when averaging spectrum or level across a longer duration, the magnitude of the changes also decreases because the average content of a long segment gets closer to the long-term average of speech. The results might also be explained by the modulation frequencies that are present in speech. Typically, the most important modulations for speech perception have rates between 4 and 16 Hz (Drullman *et al.*, 1994). The shortest segment duration of 50 ms meant up to 20 changes per second, and thus may have just missed the fastest of the relevant modulations, but still produced an ISE of 80% of that for speech. Longer segment durations removed the higher modulation frequencies, similar to a low-pass filter, and thus reduced the amount of the ISE.

It is interesting to compare the results of experiment 2 with recent work on the intelligibility of pixelated speech. In studying what they call "mosaic speech," Nakajima *et al.* (2018) found speech with temporally frozen (i.e., our "both varying" condition) critical-band segments under 40 ms length to be near perfectly intelligible, with performance dropping to 50% for 80-ms segments, and below 10% for segments of 160 ms or longer. Again, as with noise-vocoding, in the present cognitive-disruption experiment, the beneficial effects of degrading speech by freezing temporal segments appear to extend beyond what might be expected from the intelligibility of the same material.

C. Spectral versus level changes

The point of the present study was to investigate changes in loudness (or envelope) in addition to the spectral changes manipulated by vocoding, since a typical vocoder would simply preserve all changes in loudness or level. In both experiments, the ISE turned out to be bigger when both spectral content and loudness varied compared to when only the spectral content varied but loudness remained constant. This seems surprising since changes in the envelope alone did not have any effect (see also Tremblay and Jones, 1999). That may suggest that changes in envelope have an amplifying effect in producing an ISE: Spectral changes might be perceptually enhanced when there are simultaneous loudness/envelope changes. By contrast, when the spectral content does not change, there is nothing to enhance.

This line of reasoning is also consistent with the results of the three conditions of experiment 1 in which an uncorrelated envelope was applied to the respective vocoder channels (see Fig. 3). In these conditions, the temporal variation of loudness did not coincide with the temporal variation of spectral content. Therefore, one might argue, it did not amplify the ISE but showed the same task disruption as the constant-loudness conditions, with the ISE being entirely determined by the changes of the spectral content.

D. Indices predicting the ISE

Schlittmeier *et al.* (2012) showed that psychoacoustical fluctuation strength correlates with the magnitude of the ISE for many real-world sounds (speech, traffic noise, animal sounds). For these stimuli changes in envelope and spectral content are not artificially separated like in the present study. Schlittmeier *et al.* (2012) also pointed out—as have others confirmed, e.g., Ellermeier *et al.* (2015)—that a subset of the stimuli used in the present study, broadband noise that only varies in amplitude over time but has fixed spectral content, fluctuation strength is not a good predictor of performance disruption. Therefore, no attempt was made to predict the current results by determining the fluctuation strength of the signals employed. However, it is noteworthy that the ISE tended to be larger when the original envelope was used compared to when the dynamic range of possible critical-

band levels was compressed by enforcing a constant loudness or constant L_{Aeq}.

Altogether, both experiments showed that the main determinant of the ISE are changes in spectral content, thereby lending support to the idea that an estimator of the amount of spectral change between speech tokens might constitute a better predictor of the ISE. Park *et al.* (2013) and subsequently Senan *et al.* (2018a,b) proposed the frequency domain correlation coefficient [FDCC; eq. (1) in the papers] to capture the amount of spectral change between successive speech elements, though with limited success in predicting their own ISE data. It is conceivable that an improved version of this measure, supplemented by a method to factor in coherent amplitude changes might be a promising predictor of the ISE.

V. CONCLUSIONS

(1) In both experiments, the main factor that determined the magnitude of the ISE was spectral degradation. Using 24 vocoder channels or a temporal resolution of 50 ms for rendering spectral change had almost the same effect as presenting original speech.
(2) Changes in envelope alone produced nearly the same error rate as stationary noise, i.e., had no effect on serial recall performance.
(3) However, changes in envelope and spectrum had a greater effect than changes in spectrum alone. This suggests that changes in envelope amplify the irrelevant speech effect, although they do not cause an ISE themselves.
(4) When the changes in envelope are determined by a different speech signal than are the spectral changes, they do not amplify the ISE and have the same effect as a constant envelope.

ACKNOWLEDGMENTS

J.S. was supported by the Engineering and Physical Sciences Research Council (UK, Grant No. RG78536) at the time of writing. Portions of the results of experiment 2 were presented at the Annual Meeting of Experimental Psychologists (TeaP) in Heidelberg, Germany, March 21–23, 2016, and at the Inter-noise in Hamburg, Germany, August 21–24. We thank Armin Kohlrausch for valuable comments.

Bridges, A. M., and Jones, D. M. (**1996**). "Word dose in the disruption of serial recall by irrelevant speech: Phonological confusions or changing state?," Q. J. Exp. Psychol. Sec. A **49**, 919–939.
Colle, H. A. (**1980**). "Auditory encoding in visual short-term recall: Effects of noise intensity and spatial location," J. Verbal Learn. Verbal Behav. **19**, 722–735.
Colle, H. A., and Welsh, A. (**1976**). "Acoustic masking in primary memory," J. Verbal Learn. Verbal Behav. **15**, 17–31.
Dorsi, J. (**2013**). "Recall disruption produced by noise-vocoded speech: A study of the irrelevant sound effect," State University of New York, New Paltz, NY.
Drullman, R., Festen, J. M., and Plomp, R. (**1994**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**, 1053–1064.
Dudley, H. (**1939**). "Remaking speech," J. Acoust. Soc. Am. **11**, 169–177.
Ellermeier, W., and Hellbrück, J. (**1998**). "Is level irrelevant in 'irrelevant speech'? Effects of loudness, signal-to-noise ratio, and binaural unmasking," J. Exp. Psychol. Hum. Percept. Perform. **24**, 1406–1414.
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**, 1561–1569.
Ellermeier, W., and Zimmer, K. (**1997**). "Individual differences in susceptibility to the 'irrelevant speech effect'," J. Acoust. Soc. Am. **102**, 2191–2199.
Ellermeier, W., and Zimmer, K. (**2014**). "The psychoacoustics of the irrelevant sound effect," Acoust. Sci. Technol. **35**, 10–16.
Fastl, H. (**1983**). "Fluctuation strength of modulated tones and broadband noise," in *Hearing—Physiological Bases and Psychophysics*, edited by R. Klinke and R. Hartmann (Springer, Stuttgart, Germany), pp. 282–288.
Hughes, R. W. (**2014**). "Auditory distraction: A duplex-mechanism account," Psych. J. **3**, 30–41.
Hughes, R. W., Vachon, F., and Jones, D. M. (**2005**). "Auditory attentional capture during serial recall: Violations at encoding of an algorithm-based neural model?," J. Exp. Psychol.: Learn. Mem. Cognit. **31**, 736–749.
Hughes, R. W., Vachon, F., and Jones, D. M. (**2007**). "Disruption of short-term memory by changing and deviant sounds: Support for a duplex-mechanism account of auditory distraction," J. Exp. Psychol.: Learn. Mem. Cognit. **33**, 1050–1061.
ISO (**2017**). ISO 532-1, "Acoustics—Methods for calculating loudness—Part 1: Zwicker method," (International Organization for Standardization, Geneva, Switzerland).
Jones, D., Madden, C., and Miles, C. (**1992**). "Privileged access by irrelevant speech to short-term memory: The role of changing state," Q. J. Exp. Psychol. **44**, 645–669.
Jones, D. M., and Macken, W. J. (**1993**). "Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in working memory," J. Exp. Psychol.: Learn. Mem. Cognit. **19**, 369–381.
Jones, D. M., Macken, W. J., and Murray, A. C. (**1993**). "Disruption of visual short-term memory by changing-state auditory stimuli: The role of segmentation," Mem. Cognit. **21**, 318–328.
Liebl, A., Assfalg, A., and Schlittmeier, S. J. (**2016**). "The effects of speech intelligibility and temporal–spectral variability on performance and annoyance ratings," Appl. Acoust. **110**, 170–175.
Nakajima, Y., Matsuda, M., Ueda, K., and Remijn, G. B. (**2018**). "Temporal resolution needed for auditory communication: Measurement with mosaic speech," Front. Hum. Neurosci. **12**, 149.
Neely, C. B., and LeCompte, D. C. (**1999**). "The importance of semantic similarity to the irrelevant speech effect," Mem. Cognit. **27**, 37–44.
Park, M., Kohlrausch, A., and van Leest, A. (**2013**). "Irrelevant speech effect under stationary and adaptive masking conditions," J. Acoust. Soc. Am. **134**, 1970–1981.
Röer, J. P., Bell, R., and Buchner, A. (**2013**). "Self-relevance increases the irrelevant sound effect: Attentional disruption by one's own name," J. Cognit. Psychol. **25**, 925–931.
Röer, J. P., Körner, U., Buchner, A., and Bell, R. (**2017**). "Attentional capture by taboo words: A functional view of auditory distraction," Emotion **17**, 740–750.
Rouleau, N., and Belleville, S. (**1996**). "Irrelevant speech effect in aging: An assessment of inhibitory processes in working memory," J. Gerontol. Ser. B **51**, 356–363.
Salame, P., and Baddeley, A. (**1982**). "Disruption of short-term memory by unattended speech: Implications for the structure of working memory," J. Verbal Learn. Verbal Behav. **21**, 150–164.
Salamé, P., and Baddeley, A. (**1989**). "Effects of background music on phonological short-term memory," Q. J. Exp. Psychol. Sec. A **41**, 107–122.
Schlittenlacher, J., Hashimoto, T., Kuwano, S., and Namba, S. (**2017**). "Overall judgment of loudness of time-varying sounds," J. Acoust. Soc. Am. **142**, 1841–1847.
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**, 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**, 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**, 1330–1330.
Tremblay, S., and Jones, D. M. (**1999**). "Change of intensity fails to produce an irrelevant sound effect: Implications for the representation of unattended sound," J. Exp. Psychol. Hum. Percept. Perform. **25**, 1005–1015.
Wöstmann, M., and Obleser, J. (**2016**). "Acoustic detail but not predictability of task-irrelevant speech disrupts working memory," Front. Hum. Neurosci. **10**, 538.
Zwicker, E. (**1961**). "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," J. Acoust. Soc. Am. **33**, 248.