When softer sounds are more distracting: Task-irrelevant whispered speech causes disruption of serial recall

Florian Kattner,[1] Julia Föcker,[2] Cleopatra Christina Moshona,[3] and John E. Marsh[4]

[1] *Institute for Mind, Brain and Behavior, Health and Medical University, Schiffbauergasse 14, 14467 Potsdam, Germany*[a]

[2] *College of Health and Science, School of Psychology, University of Lincoln, Brayford Pool, Lincoln, LN6 7TS, United Kingdom*

[3] *Engineering Acoustics, Institute of Fluid Dynamics and Technical Acoustics, Technische Universität Berlin, Einsteinufer 25, 10587 Berlin, Germany*

[4] *School of Psychology and Humanities, University of Central Lancashire, Preston, PR1 2HE, United Kingdom*

(Dated: 2 November 2024)

Abstract

Two competing accounts propose that the disruption of short-term memory by irrelevant speech arises either due to interference-by-process (e.g., changing-state effect) or attentional capture, but it is unclear how whispering affects the irrelevant speech effect. According to the interference-by-process account, whispered speech should be less disruptive due to its reduced periodic spectro-temporal fine structure and lower amplitude modulations. In contrast, the attentional account predicts more disruption by whispered speech, possibly via enhanced listening effort in the case of a comprehended language. In two experiments, voiced and whispered speech (spoken sentences or monosyllabic words) were presented while participants memorized the order of visually presented letters. In both experiments, a changing-state effect was observed regardless of the phonation (sentences produced more disruption than 'steady-state' words). Moreover, whispered speech (lower fluctuation strength) was more disruptive than voiced speech when participants understood the language (Exp. 1), but not when the language was incomprehensible (Exp. 2). The results suggest two functionally distinct mechanisms of auditory distraction: While changing-state speech causes automatic interference with seriation processes regardless of its meaning or intelligibility, whispering appears to contain cues that divert attention from the focal task primarily when presented in a comprehended language, possibly via enhanced listening effort.

[a]Correspondence to: florian.kattner@hmu-potsdam.de

I. INTRODUCTION

Most readers will have experienced disruption to their cognitive performance in the presence of task-irrelevant background sound even when the focal task information is in a different modality (e.g., visual) and therefore cannot be attributed to interference at the sensory level (e.g., perceptual masking). Instead it must emerge from an interaction between visual and auditory processing at a level beyond the sensory organs. One well-studied example of auditory distraction is the disruption of verbal-serial short-term memory produced by task-irrelevant speech (Colle and Welsh, 1976; Salamé and Baddeley, 1982). In this irrelevant sound paradigm, participants are asked to recall a series of usually visually-presented digits or words while being presented with different types of sound via headphones that they are instructed to deliberately ignore. Immediate or delayed visual-verbal serial recall accuracy is usually lower when task-irrelevant speech is presented during encoding or retention of the items, compared to silence, continuous noise, or instrumental background music (in particular when the notes are played "legato", i.e., smoothly connected without gaps of silence; Ellermeier and Zimmer, 1997; Salamé and Baddeley, 1989; Schlittmeier *et al.*, 2008), regardless of the volume of irrelevant speech (Ellermeier and Hellbrück, 1998). However, distraction of visual-verbal serial recall is not restricted to speech or "speech-like" material (e.g., music; Salamé and Baddeley, 1989), as stimuli sufficiently unlike speech such as spectro-temporally varying tones (Jones and Macken, 1993) or pitch glides randomly interrupted with quiet (Jones *et al.*, 1993) have also been found to interfere with the serial order retention of to-be-remembered items. However, the magnitude of the disruptive effect of non-speech sound (e.g., music, varying tones, or interrupted pitch glides) was found to be significantly lower than that of irrelevant speech (see effect sizes in Ellermeier and Zimmer, 2014, Table 1).

The disruptive impact of irrelevant speech is observed even if it is presented only during a retention period after encoding of the visually-presented items. Therefore, the disruption is not due to interference in the encoding of digits, but occurs at a later stage of processing within memory (Miles *et al.*, 1991). Within the context of short-term memory, two broad mechanisms have been proposed to account for auditory distraction. According to the 'interference-by-process account' (which can be considered a generalization of the 'object-oriented episodic record account', Jones *et al.*, 1996; Marsh *et al.*, 2009), processing of task-irrelevant sound produces interference with cognitive processes that are demanded by the focal task. One prototypical example of such interference is the changing-state effect, which refers to the observation that spectro-temporally varying sound (e.g., free-running speech or random sequences of syllables or tones) gives rise to the automatic formation of an ordered auditory sequence (as part of auditory scene analysis, Bregman, 1990) which then interferes with deliberate serial-order processing of to-be-remembered information. In line with this assumption, it has been found that changing-state sequences consisting of spectro-temporally varying acoustical tokens (thus conveying irrelevant order information) are more disruptive than steady-state repetitions of a single acoustical item, regardless of whether the sequences comprise speech or non-speech materials (e.g., Hadlington *et al.*, 2004; Jones and Macken, 1993; Jones *et al.*, 1993; Tremblay *et al.*, 2000, but see LeCompte *et al.*, 1997). Moreover, this well-established 'changing-state effect' (Jones *et al.*, 1992) seems to interfere primarily in tasks that require serial-order processing or with the performance of participants who report using serial rehearsal for item retention, whereas often no changing-state effect is found with non-serial memory tasks such as the missing item task or in a mental arithmetic task (Beaman and Jones, 1997; Campbell *et al.*, 2002; Hughes and Marsh, 2020; Jones and Macken, 1993; Kattner *et al.*, 2023). Importantly, according to the interference-by-process account, auditory distraction in a serial recall task should depend primarily on the acoustical profile of the irrelevant sound (e.g., the proportion of changes in rhythm, frequency, or amplitude), and it has been discussed whether psychoacoustical metrics such as the degree of amplitude or frequency modulation, 'fluctuation strength' , or spectral detail may be useful predictors of the changing-state effect (Ellermeier and Zimmer, 2014; Schlittmeier *et al.*, 2012). It is worth noting that the degree of distraction imposed by the changing-state sound may also depend on certain speech-specific properties of the sound. For instance, it has been found that artificial sinewave speech containing three formants can be as disruptive as natural speech, but a temporal reversal of the first two formants (i.e., degrading the formant transitions that are required to identify 'consonants' in sine wave speech) reduced the degree of distraction considerably (Viswanathan *et al.*, 2014). Similarly, Dorsi *et al.* 2018 found that if the spectral detail of the irrelevant speech is reduced by decreasing the number of vocoder bands, the distraction caused by task-irrelevant speech diminishes (see also Ellermeier *et al.*, 2015). Taken together, these findings suggest that speech-specific properties and signal fidelity (i.e. the internal properties of a signal, including its structural details and relationships) may play a cardinal role in modulating the effects of task-irrelevant speech.

In contrast, an alternative ‘unitary attentional account’ supposes that irrelevant sounds divert attentional or cognitive resources from the focal task (Bell *et al.*, 2008, 2012; Cowan, 1995). Specifically, certain types of irrelevant sound (e.g., speech, acoustical changes, unexpected events, or otherwise meaningful sounds) are assumed to capture attention and produce unspecific disruption to any attention-demanding task. By this approach, a sound may capture attention either because it cannot be predicted based on previous stimulation (random acoustical changes or an auditory oddball in a regular sequence, Eimer *et al.*, 1996), or because semantic or syntactic properties of the sound indicate enhanced relevance to the individual (e.g., one’s own name or an emotional word Röer *et al.*, 2013, 2017a). According to this account, the degree of disruption should not depend on the exact cognitive processes demanded by the focal task (e.g., retention of order), but it should vary as a function of the perceptual load and/or the working memory capacity available to the participant. Indeed, it has been reported that the disruptive effect of a deviant (unexpected) sound in a regular sequence is more pronounced in, or restricted to, individuals with low working-memory capacity and conditions of low task-encoding load (Hughes *et al.*, 2013; Hughes and Marsh, 2019; Marsh *et al.*, 2018; Sörqvist, 2010)(but see Körner *et al.*, 2017; Labonté *et al.*, 2022). However, in contrast to a unitary attentional account, the disruptive effect of other types of sounds, in particular changing-state sound, does not seem to depend on task load and the individuals’ working memory capacity (Hughes *et al.*, 2013) – though it might be sensitive to the listeners’ auditory processing and/or selective attention (cf. reduced distraction in blind individuals Kattner and Ellermeier, 2014; Kattner *et al.*, 2024).

To account for such findings, a duplex-mechanism account of auditory distraction has been proposed, assuming that interference-by-process and attentional capture may be two functionally distinct mechanisms that can produce task disruption (Hughes, 2014; Hughes *et al.*, 2005b, 2007). That is, irrelevant sound may either produce interference with specific cognitive processes that are demanded by the focal task (Kattner, 2024; Marsh *et al.*, 2009, e.g., changing-state sound interferes with a seriation process, and semantic properties of irrelevant sound may interfere with semantic organization; cf.) or it may capture attention due to its unpredictability or meaningfulness and cause unspecific disruption (assuming that sufficient attentional/cognitive resources are available to process the sound).

Whispered speech is an interesting stimulus to test the functional dissociation between interference-by-process (i.e., interference with seriation) and attentional capture. In contrast to voiced (modal) phonation, vocal cord vibration, periodic glottal excitation and harmonic structure are completely absent in whispered speech, due to its distinct production mechanism. The glottis is abducted, except for a small triangular opening in the cartilaginous portion (Laver, 1994). The pulmonic airstream forced through this narrow gap has a hissing, noise-like quality, produced by turbulence from the friction of the air around the larynx (Eckert and Laver, 1994). Consequently, whispered speech is dominated by strong aperiodic energy. It is further characterized by a notable decrease in vowel amplitude, typically by about 20–25 dB, flatter spectral slopes and an upwards shift of formant frequencies, affecting vowel quality and intelligibility (Ito *et al.*, 2005). These formant frequency trends have been reported across languages, with greater shifts for F1 than F2 or F3 (see e.g., Eklund and Traunmüller, 1997; Heeren, 2015; Jovičić and Šarić, 2008).

Due to the described acoustic features of whispered speech, listeners have been found to be less accurate when identifying linguistic information (Konno, 2016) and emotion (Frühholz *et al.*, 2016). Whispered speech has also been found to severely degrade speaker recognition systems, which are primarily based on neutral mode speech-processing algorithms (Zhang and Hansen, 2018). However, despite lacking a fundamental frequency (F_0), whispered speech has a clearly perceivable prosodic structure. Żygis *et al.* showed that the spectral properties of consonants change during whispering to convey intonation patterns, compensating for the absence of a fundamental frequency. Jovičić and Šarić report longer durations for consonants. Such modifications provide evidence for cue-trading relations, where one dominant cue is substituted by the integration of multiple others, which would otherwise be less prominent when considered in isolation (Żygis *et al.*, 2017).

Due to its acoustical profile with a decreased amplitude envelope and reduced periodic spectro-temporal fine structure, whispered speech as compared with voiced speech should produce either similar or less interference with serial-order processing. Previous studies have shown that modulations of the spectral detail of irrelevant speech (i.e., presenting noise vocoded speech varying in the number of independently amplitude-modulated frequency bands) influences the degree of distraction, with reduced spectral fidelity (decreasing number of frequency bands) attenuating disruption of serial recall (Dorsi *et al.*, 2018; Ellermeier *et al.*, 2015). Similarly, manipulations of speech prosody (e.g., emotional speech or urgent intonations) were found to increase disruption of serial recall performance (Kattner and Ellermeier, 2018; Ljungberg *et al.*, 2012), suggesting that enhanced amplitude (and frequency) modulations in speech intonations may increase interference with order processing (note that emotional speech prosody did not affect performance on the missing-item task, which does not required the retention of serial order). It has also been found that disruption of serial recall is determined largely by changes in vowels rather than consonants (i.e., consonant-vowel-consonant syllables are more disruptive when all components or only the vowels change, compared to when a consonant changes; Hughes *et al.*, 2005a). Hence, in particular due to the lower amplitude of whispered vowels (Ito *et al.*, 2005), it could be predicted that the changing-state effect on serial recall may be reduced with whispered compared to voiced speech (i.e., there should be an interaction between state and phonation, see Table I). More specifically, due to the lower amplitude modulations, recall accuracy should be higher in the whispered changing-state condition than in the voiced changing-state condition, whereas less phonation-related differences should be observed in the steady-state conditions. However, there are currently no studies showing that a decrease in the depth of amplitude modulations decreases distraction, and some studies found that serial recall is insensitive to the overall level and intensity changes of irrelevant speech (Ellermeier and Hellbrück, 1998; Tremblay and Jones, 1999). In contrast, more recent findings suggest that both steady-state (repeated words) and changing-state (varying words) sequences of high-intensity sound (75 dB(A)) are more disruptive than low-intensity sound sequences (45 dB(A)) in a serial recall task (Alikadic and Röer, 2022). In line with this finding, it could be argued that due to the lower amplitude modulations and overall loudness, both steady- and changing-state sequences of whispered words should be less disruptive than their voiced counterparts. In the present study this was controlled partially by normalizing the amplitudes of whispered and voiced speech recordings, but also by testing level and loudness as predictors of serial recall accuracy. More precisely, in order to test the contribution of (psycho)acoustical properties of irrelevant sound (Ellermeier and Zimmer, 2014), a regression analysis was conducted to predict serial recall accuracy based on multiple signal metrics including loudness, fluctuation strength, and tonality.

On the other hand, it could also be argued that whispered phonation increases the perceptual demands (due to a lack of f0 cues and altered spectral fidelity) to process the speech signal and to achieve speech recognition, thus reducing the resources available to process the focal serial recall task. According to current speech processing models (Pichora-Fuller *et al.*, 2016; Rönnberg *et al.*, 2021, 2013; Wingfield, 2016, e.g., 'framework for understanding effortful listening' and 'ease of language understanding' models;) the degradation of signal clarity due to the absence of f0 cues and formant alternations in whispered speech (reducing speech quality and intelligibility) should impose additional cognitive processing load on passive listeners (e.g., speech decoding and lexical access), compared to clearly intelligible voiced speech. However, this additional load is expected only in listeners who are familiar with the language, because extra processing resources or 'listening effort' would be dedicated to irrelevant speech only when there is some degree of mismatch between the degraded (whispered) speech signal and phonological representations in the listeners' mental lexicon. Disruption in the serial recall task would thus be the consequence of the enhanced listening effort required to process acoustically degraded, whispered speech in a comprehensible language. Nevertheless, it seems more difficult to explain other effects of 'degraded' irrelevant speech in terms of enhanced listening effort, because often degraded speech and lower speech intelligibility results in less disruption of serial recall compared to more intelligible speech (e.g., noise-vocoded and locally time-reversed speech, Ellermeier and Hellbrück, 1998; Ellermeier *et al.*, 2015; Ueda *et al.*, 2019).

Similar predictions could be derived from an attentional capture account though, assuming that attention is directed to certain semantic properties or social functions associated with whispered speech. As a universal, paralinguistic phenomenon found across cultures and unique to human species, whispering has important social functions. These functions vary, depending on whether whispering is used privately or in the public domain and influence the way it is perceived (Cirillo, 2004; Cirillo and Todt, 2005). While it can be positively connotated as an expression of affection in the private domain, it may elicit negative judgements when used in the public domain. One possible explanation for this is that whispering is often used to signal secrecy and confidentiality (Laver, 1994), thereby inducing mistrust and social segregation and diverting the attention of non-addressees by increasing auditory vigilance (compare in-group and vigilance hypothesis formulated by Cirillo and Todt, 2005). This may lead to greater attentional capture, either because whispered speech is considered to be more relevant to the individual (potential self-relevance or goal-relevance of whispered content), due to the greater listening effort required to process the meaning of acoustically degraded (and potentially interesting) whispered background speech, or because of its distinctiveness in relation to the surrounding stimuli, as expressed by the salience hypothesis (Günther *et al.*, 2017).

Importantly, such an attentional capture mechanism should be independent of, and additive to, the interference with serial-order processing produced by changing-state speech (i.e., there should be no interaction between a disruptive effect of whispered speech and the changing-state effect; Hughes *et al.*, 2005b). That is, whispered speech is expected to cause more disruption than voiced speech both with steady- and changing-state sequences of irrelevant speech (see Table I). Moreover, if the disruptive effect of whispered speech was due to enhanced listening effort or the individuals' motivation to process semantic content of unattended speech, then whispered speech should be more disruptive only when whispered in a language that is comprehensible to the individual.

To the best of our knowledge, whispered speech has only been used in one previous study testing the effect of background sound on the recall of short spoken lectures, but in this study whispered speech was not contrasted with loud/voiced speech (and whispering was not a reliable predictor of lecture recall accuracy, Zeamer and Fox Tree, 2013, Exp. 3). In the present series of experiments the effect of task-irrelevant whispered speech in serial recall was contrasted both with normally-phonated modal speech and with a silent control condition. In addition, the effect of whispering was tested both with steady-state and changing-state speech. In this context, "steady state" is used as a term to describe the repetition of single auditory tokens (e.g., monosyllabic words), resulting in a "steady" stream of sounds, while "changing state" refers to altering auditory tokens as contained in spoken sentences. This allows a test of whether whispering either (a) reduces the changing-state effect because the reduced frequency and amplitude modulations interfere less with seriation, or (b) produces process-independent distraction due to a diversion of attentional (e.g., triggered by social functions of whispered speech) or additional cognitive demands (enhanced listening effort to process whispered speech). That is, according to an interference-by-process account, the changing-state effect should be reduced with whispered speech, whereas according to an

Account	**Prediction**	**Mechanism**
Interference-by-Process	*Phonation × State Interaction:* Higher recall accuracy with whispered compared to voiced changing-state speech, smaller difference between whispered and voiced steady-state speech	Reduced amplitude envelope of whispered speech should provide less order information and thus cause less interference with seriation (i.e., whispered speech becomes more like a 'steady-state' signal)
Attentional Capture	*Independent main effects of Phonation and State in comprehensible language:* Lower recall accuracy with whispered speech compared to voiced speech and lower recall accuracy with changing-state compared to steady-state speech; *Main effect of State*, but no main effect of Phonation with irrelevant speech in *foreign language*	Whispered speech should cause attentional disruption due to its potential interest or enhanced listening effort, which is independent of the automatic interference produced by changing-state speech.

TABLE I. Summary of the main theoretical accounts and their predictions tested in this study and a description of the assumed mechanisms.

attentional capture or listening effort account there should be additive disruptive effects of changing-state and whispered speech.

II. EXPERIMENT 1

The aim of Experiment 1 was to test whether whispered speech (a) reduces the changing-state effect on serial recall as predicted by an account that assumes interference-by-process (e.g., due to reduced spectro-temporal variation / fine structure caused by the absence of fundamental frequency cues and lower vowel amplitude) or (b) is more disruptive than voiced speech as predicted by an account that predicts attentional capture by specific features of whispered speech (e.g., its semantic properties). More specifically, an interference-by-process account predicts an interaction between phonation and state of irrelevant speech, with whispered changing-state speech being less disruptive in serial recall compared to voiced changing-state speech, whereas phonation should matter less for steady-state sequences (see Table I). In contrast, an attentional capture or duplex-mechanism account predicts independent main effects of the phonation and state of irrelevant speech, assuming that whispered speech should cause additional attentional capture and be more disruptive than voiced speech regardless of whether speech is presented in steady-state or changing-state sequences. The changing-state effect would thus be independent of a disruptive effect of whispered speech, in particular when the language or whispered speech is comprehensible to the listener (see also Table I).

A. Method

1. Participants

Ninety-four participants (62, female, 31 male, 1 other) were recruited at the Health and Medical University campus in Potsdam, Germany. Ages ranged between 18 and 61 years ($M = 24.0, SD = 8.6$). Participants were native speakers of German and all reported normal hearing and normal or corrected-to-normal vision. The study has been conducted strictly in accordance with the Ethical Principles of the Acoustical Society of America for Research. All participants gave written informed consent before starting the tasks, acknowledging that participation is voluntary and they were free to withdraw from the study at any time without negative consequences. Participants were also informed about the scientific purpose (mainly during debriefing), the potential discomfort during the task (e.g., due to cognitive demand), the absence of risks to mental and physical well-being, and the confidentiality of personal data. Student participants majoring in psychology ($n = 70$) were compensated with course credits. Non-student participants received no compensation.

2. Stimuli

Two native lay speakers were recruited to record twenty unique German sentences in a male and a female voice using a Behringer B1 Bundle microphone and a FMR Audio RNP 8380 preamplifier. The sentences were adapted from previous studies (Hughes and Marsh, 2020; Kattner *et al.*, 2022; Röer *et al.*, 2015) and comprised various categories such as weather forecasts, traffic reports, cooking recipes, poems, operating manuals, and scientific descriptions. Each sentence was spoken once with voiced phonation (normal speech) and once with whispered phonation by each speaker. The speakers were instructed to adjust their rate of speaking to reach about 8 s duration. The recordings were sampled at 44.1 kHz (16 bits). For each of the twenty (changing-state) sentences, a unique 8-s steady-state sequence was created (with voiced and whispered phonation) by selecting a single monosyllabic word from the sentence (e.g., 'Hand', 300-500 ms duration), and concatenating it eight times at a rate of one word per second (we note that this creates short gaps of silence between successive utterances). Thus, in total 160 sound files were created from the twenty sentences (male/female speaker × voiced/whispered phonation × changing-/steady-state).

The amplitudes of all recordings were normalised in Audacity (https://www.audacity.de/) to minimize level differences between sound conditions. Ten different sentences and steady-state sequences were selected for the voiced and whispered phonation conditions, and participants were presented either with the male or the female voice only. That is, forty unique speech recordings were presented to each participant.

Exemplary FFT spectra of whispered and voiced speech are illustrated in Fig. 1. To estimate the overall speech intensity, the A-weighted, equivalent continuous sound pressure level (LA_{eq}) in dB(A) was determined for each sound file, using ArtemiS SUITE (HEAD Acoustics GmbH, Herzogenrath, Germany). In addition, the psychoacoustic metrics 'Zwicker' loudness (cubic average) as per DIN45631/A1 (DIN Deutsches Institut für Normung e.V., 2010), fluctuation strength (Fastl, 1982; Fastl and Zwicker, 2007) and sharpness as per DIN 45692 (DIN Deutsches Institut für Normung e.V., 2009) were computed for each sound file. Roughness and tonality were calculated according to the ECMA-418-2 (2nd) standard (ECMA International, 2022). A free sound field was assumed for the calculation of all metrics. Loudness reflects how loud a sound is perceived by human listeners. In contrast, decibels (dB) quantify the physical intensity of a sound. Sharpness is another perceptual attribute, related to the spectral content of sounds and in particular the high frequency components. Using the Relative Approach Method (RAM) (Genuit, 1996), the spectro-temporal changes in the signal were quantified by extrapolating the signal history. Fluctuation strength and roughness both reflect the sensation caused by variations in the amplitude and frequency of sounds. While fluctuation strength mirrors the slow, rhythmical variations, roughness

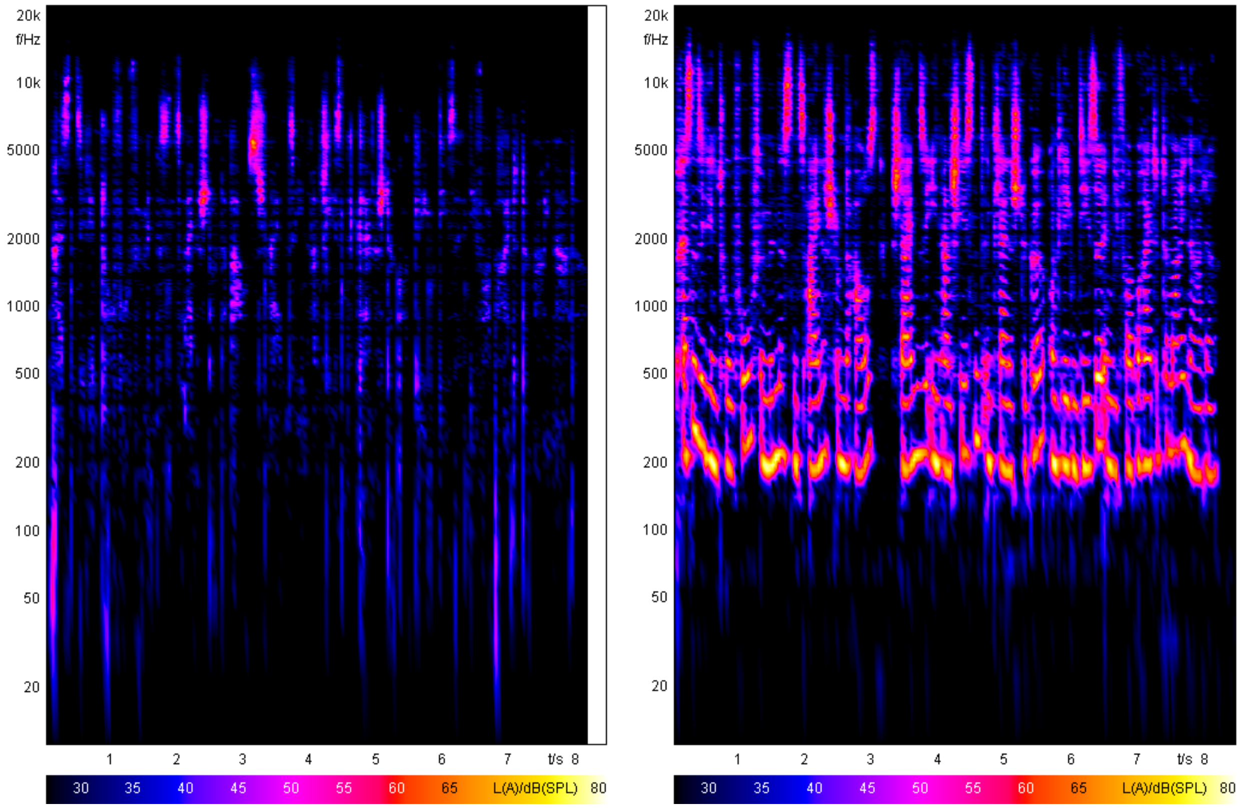

FIG. 1. FFT spectra vs. time of a changing-state sentence spoken by a female speaker with whispered (left) and voiced (right) phonation.

describes rapid, irregular variations. Finally, tonality indicates the relative prominence of the tonal elements within a specific noise spectrum.

Descriptive statistics of the psychoacoustic metrics are shown in Table II. Due to violations of homogeneity of the covariance matrices ($\chi^2(198) = 768.94$; $p < .001$) and deviations from multivariate normality ($W = 0.87$; $p < .001$), non-parametric Kruskal-Wallis rank-sum tests were conducted to test for differences in psychoacoustical metrics between experimental conditions.

Level was significantly higher in voiced compared to whispered speech, $W(1) = 1716$, $p < .001$. Moreover, there was a significant level difference between steady- and changing-state speech, $W(1) = 53.18$, $p < .001$ (higher levels in steady-state), and also between male and female voices, $W(1) = 4.43$, $p = .035$.

'Zwicker' loudness was significantly higher for voiced than for whispered speech, $W(1) = 49.57$, $p < .001$, and steady-state sequences are louder than changing-state sentences, $W(1) = 33.05$, $p < .001$. The speaker difference in loudness was not significant, $W(1) = 1.30$, $p = .254$. We note that the loudness differences between state and voice conditions are rather small in magnitude and any detrimental effect of loudness on serial recall would work against the main hypotheses that the softer whispered speech and changing-state speech will be more disruptive than voiced speech and steady-state speech. Moreover, reducing loudness differences through normalization could have removed the characteristic attention-capturing properties of whispered speech.

Sharpness was significantly higher for whispered than for voiced speech, $W(1) = 66.72$, $p < .001$, reflecting the larger amount of high-frequency energy in whispered speech. There was also a speaker difference, $W(1) = 21.35$, $p < .001$, but no difference between steady- and changing-state sequences in sharpness, $W(1) = 0.60$, $p = .441$.

The difference in *roughness* between whispered and voiced speech was also significant, , $W(1) = 14.17$, $p < .001$, likely attributed to the absence of an F_0 in whispered speech. Roughness was also higher in male than in female speech, , $W(1) = 56.50$, $p < .001$, but there was no roughness difference between steady- and changing-state sequences, $W(1) = 0.20$, $p = .656$.

Fluctuation strength in turn was significantly higher in voiced speech compared to whispered speech, $W(1) = 47.22$, $p < .001$, indicating a reduced amplitude envelope with whispered phonation, and steady-state speech was significantly more fluctuating than changing-state speech, $W(1) = 89.55$, $p < .001$ (as to be expected due to the silent gaps between successive words in steady-state streams). There was no significant speaker difference in fluctuation strength though, $W(1) = 1.02$, $p = .31$.

There was also a significant difference in the spectro-temporal variation quantified via the *Relative Approach* metric (an extrapolation method) between steady- and changing-state sequences, $W(1) = 20.91$, $p < .001$, as well as between whispered and voiced speech, $W(1) = 30.68$, $p < .001$, reflecting more spectral and temporal variation in voiced speech. There was no speaker difference in spectro-temporal variation, $W(1) = 1.98$, $p = .159$.

Finally, *tonality* was higher in voiced than in whispered speech, $W(1) = 79.88$, $p < .001$, as well as in steady-state speech compared to changing-state sentences, $W(1) = 4.41$, $p = .036$. These distinct differences in vocal quality are also visible in Figure 1. Tonality was also higher in female speech than in male speech, $W(1) = 9.75$, $p = .002$.

3. Apparatus

The study was conducted in a single-walled sound-attenuated listening booth (Studiobox GmbH, Munich, Germany). The experiment ran on a Lenovo Thinkstation P350 desktop computer and the experimental routines were programmed in Python utilizing the PsychoPy package (Peirce *et al.*, 2019). Visual stimuli were presented on a BenQ GW2780 IPS screen (27 in).

Sounds were D/A converted by an ESI MAYA44 eX PCIe sound card (ESI Audiotechnik, Leonberg, Germany) passed through a Behringer Powerplay HA8000 amplifier (Behringer,

TABLE II. Mean psychoacoustic metrics of the 20 changing-state sentences and 20 monosyllabic steady-state word sequences, each spoken aloud (voiced) and whispered by a male and female speaker (standard deviations in parentheses).

Parameter	**Speaker**	**Changing-State**		**Steady-State**	
		whispered	**voiced**	**whispered**	**voiced**
LA_{eq}	male	71.62 (2.47)	74.79 (1.63)	78.16 (2.89)	79.60 (2.89)
(dB(A))	female	71.28 (2.51)	75.52 (1.22)	74.36 (3.32)	77.18 (2.30)
Loudness	male	17.63 (2.31)	24.75 (2.53)	28.07 (5.08)	30.52 (4.99)
(sone)	female	16.64 (2.50)	27.98 (2.65)	22.07 (4.83)	28.77 (4.50)
Sharpness	male	1.66 (0.11)	1.34 (0.09)	1.68 (0.16)	1.47 (0.22)
(acum)	female	1.89 (0.10)	1.54 (0.09)	1.78 (0.14)	1.59 (0.16)
Roughness	male	0.37 (0.08)	0.88 (0.16)	0.33 (0.08)	1.05 (0.39)
(asper)	female	0.31 (0.04)	0.26 (0.03)	0.28 (0.10)	0.35 (0.09)
Fluctuation	male	0.20 (0.04)	0.55 (0.15)	0.56 (0.16)	1.34 (0.24)
Strength (vacil)	female	0.18 (0.03)	0.36 (0.07)	0.62 (0.19)	1.32 (0.21)
Rel. Approach	male	38.72 (5.24)	54.58 (6.31)	39.24 (7.33)	43.05 (8.31)
(cPa)	female	40.77 (4.40)	48.69 (5.34)	37.02 (4.56)	38.94 (6.34)
Tonality	male	0.21 (0.05)	0.39 (0.12)	0.34 (0.09)	0.41 (0.15)
($t.u._{HMS}$)	female	0.25 (0.04)	0.91 (0.15)	0.27 (0.11)	1.12 (0.41)

Penang, Malaysia) and played diotically via Beyerdynamic DT 990 PRO headphones (Beyerdynamic, Heilbronn, Germany) at an overall average playback level of 70 dB(A), (with inter-sentence variability accounting for the differences in state and phonation, see Table II).This playback level deviated slightly from the original sound pressure level of the recordings, but was deemed comfortable for participants.

4. Experimental Design and Procedure

A 2 (State: steady, changing) × 2 (Phonation: voiced, whispered) experimental design was implemented. Silence was presented as a control condition to assess possible disruptive effects of steady-state sound (Bell *et al.*, 2019). There were ten repetitions of each auditory condition, resulting in a total of 50 trials that were presented in fully randomized order. Half of the participants ($n = 47$) were presented with irrelevant speech in the male voice, and the other half was presented with the female voice only.

Participants were instructed to memorize the order of eight consonants presented on the screen while ignoring the sound that was played via headphones. Participants started each trial at their own pace by pressing the space bar. Then an empty white square was presented in the center of the black screen for 1 s before the eight to-be-remembered consonants were presented successively within the square. The consonants were drawn randomly without replacement from 'F', 'G', 'K', 'L', 'M', 'P', 'Q', 'S', and 'T'. Each consonant was presented for 800 ms and followed by a 200-ms inter-stimulus interval showing the empty square. Irrelevant sound was presented during the visual presentation of consonants (8 s). After a silent retention interval of 6 s (showing a blank screen), a 3×3 response matrix was presented on the screen showing all nine consonants arranged alphabetically. Participants were prompted to click the consonants in the memorized order. The sequence of clicked consonants was presented on the screen (above the matrix). Participants were able to click consonants multiple times, but they could not correct their previous responses. After the last click response, the number of consonants that were recalled in the correct serial position was presented as visual feedback for 1.5 s (e.g., 'Trial 3: 6 correct'). The next trial started immediately after the feedback. After the 10^{th}, 20^{th}, 30^{th}, and 40^{th} trial, an additional text prompt was presented on the screen, indicating that participants could now take a short break before proceeding with the next trial.

B. Results

The serial recall accuracy in the five auditory conditions is illustrated in Fig. 2, both averaged and across serial positions. A 2 (state: steady, changing) × 2 (phonation: voiced, whispered) × 8 (serial position: 1-8) repeated-measures ANOVA on recall accuracy revealed a significant main effect of state, $F(1, 93) = 50.13$, $MSE = 0.05$, $p < .001$, $\hat{\eta}^2_p = .350$ (i.e., a changing-state effect: lower recall accuracy with changing-state speech compared to steady-state words), and a significant main effect of phonation, $F(1, 93) = 6.27$, $MSE = 0.05$, $p = .014$, $\hat{\eta}^2_p = .063$, with lower recall accuracy during whispered speech ($M = .53, SD = .12$) than during voiced speech ($M = .55, SD = .12$).The relative decrement in performance compared to silence ($(accuracy_{silence} - accuracy_{speech})/accuracy_{silence}$, cf. Ellermeier and Zimmer, 2014) was 12.9% for whispered speech, compared to 9.6% for voiced speech. There was no interaction between state and phonation, $F(1, 93) = 0.17$, $MSE = 0.04$, $p = .680$, $\hat{\eta}^2_p = .002$ As can be seen in Fig. 2A, whispered speech produced similar disruption of serial recall compared to voiced speech, regardless of whether speech consisted of steady-state words and changing-state sentences.

As to be expected, the ANOVA also revealed a significant serial position effect, $F(3.42, 317.70) = 317.75$, $MSE = 0.10$, $p < .001$, $\hat{\eta}^2_p = .774$, with higher accuracy for items from the beginning

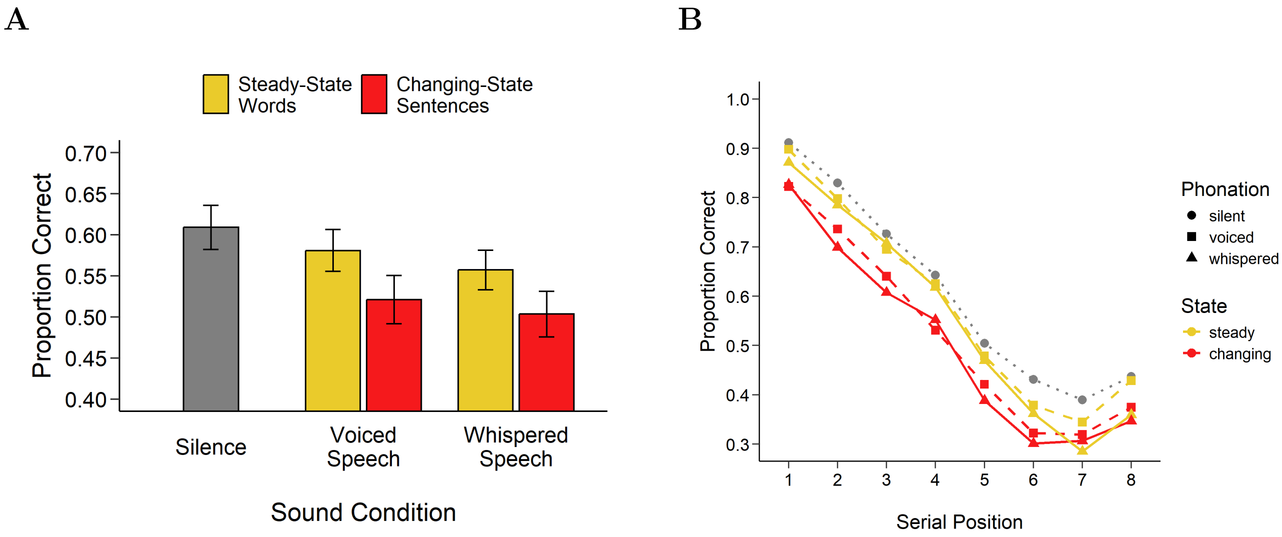

FIG. 2. **(A)** Mean serial recall accuracy in silence and when either steady-state or changing-state speech was presented with voiced or whispered phonation during item encoding in Experiment 1. Error bars represent 95% confidence intervals. **(B)** Serial recall accuracy in the five irrelevant sound conditions as a function of serial position.

of the list (position 1: $M = 0.85$ $[0.83, 0.88]$, position 2: $M = 0.75$ $[0.73, 0.78]$, position 3: $M = 0.66$ $[0.63, 0.69]$, position 4: $M = 0.58$ $[0.55, 0.62]$, position 5: $M = 0.44$ $[0.40, 0.48]$, position 6: $M = 0.34$ $[0.31, 0.38]$, position 7: $M = 0.31$ $[0.28, 0.35]$; 95% CIs in brackets) as well as a small recency effect (position 8: $M = 0.38$ $[0.34, 0.41]$). It was further tested whether the effect of whispering and changing-state sound differs for items in different serial positions (compare Fig. 2B). To that effect, the ANOVA revealed a significant interaction between state and serial position, $F(5.31, 493.77) = 4.14$, $MSE = 0.02$, $p < .001$, $\hat{\eta}^2_p = .043$, but not between phonation and serial position, $F(5.12, 476.25) = 1.70$, $MSE = 0.02$, $p = .130$, $\hat{\eta}^2_p = .018$. A planned contrasts analysis corrected for multiple comparisons (Benjamini and Hochberg, 1995) revealed that the changing-state effect was significant at serial positions 1 to 6 ($p_{\mathrm{BH}(8)} < .001$) and 8 ($p_{\mathrm{BH}(8)} = .017$), but not at serial position 7

($p_{\mathrm{BH}(8)} = .886$). There was no three-way interaction between state, phonation, and serial position, $F(6.12, 569.19) = 1.77$, $MSE = 0.02$, $p = .101$, $\hat{\eta}_p^2 = .019$.

C. Discussion

Experiment 1 demonstrated that whispered speech produced more disruption in a serial recall task compared to speech presented with (louder) voiced phonation. As expected, changing-state speech (full German sentences) was also more disruptive than steady-state speech consisting of repetitions of a single monosyllabic German word. Interestingly, these two effects seem to be independent, as indicated by the absence of an interaction. While the changing-state effect is most likely due to interference between the order information in the auditory stream and deliberate serial-order processing, the "whispering effect" may be due to attentional capture elicited either by the potential meaning of whispered information or the enhanced listening effort required to process the semantic content of whispered speech in a comprehensible language (as predicted by speech processing accounts, Pichora-Fuller *et al.*, 2016; Rönnberg *et al.*, 2021, 2013; Wingfield, 2016). To test this last assumption, a second experiment was conducted in which we tried to replicate the disruptive effect of whispered speech that is presented in a language that is foreign to the listener, making it incomprehensible. If the attentional disruption was due to enhanced listening effort in case of acoustically degraded but comprehensible whispered speech, then it should disappear when participants perceive the language as an incomprehensible, foreign language, because in this case there would be no mismatch between the task-irrelevant speech signal and phonological representations stored in the mental lexicon.

III. EXPERIMENT 2

Experiment 2 was a close replication of Experiment 1, but with a sample of participants, who did not understand the irrelevant speech language (German).

1. Participants

A power analysis based on the effect size for the whispering effect observed in Experiment 1 ($\hat{\eta}_p^2 = .063$) revealed that a sample size of $N = 51$ is required to reach a statistical power of $1 - \beta = .95$ ($\alpha = .05$) for the detection of a two-level main effect in a repeated-measures ANOVA. Fifty-one participants (42 women) who did not speak or understand German were recruited either at the University of Lincoln, UK ($n = 44$), or at Ludwig Maximilian Universität München, Germany ($n = 7$). Ages ranged between 18 and 53 years ($M = 29.3; SD = 11.8$). All participants reported normal hearing and normal or corrected-to-normal vision. Most participants of Experiment 2 were native speakers of English, but there were also a few native speakers of other languages (e.g., Chinese and Spanish). We also note that an additional data analysis including only the subsample of native speakers of English – not including speakers of a logographic language such as Chinese – produced the same overall pattern of results. All participants confirmed not speaking or understanding the German language. The study has received ethics approval by the ethics committee of the University of Lincoln (ref: 33415). All participants gave written informed consent before starting the task. Participants of Experiment 2 were compensated with course credit.

2. Stimuli and Apparatus

The set of German speech recordings from Experiment 1 was used also for Experiment 2, but 16 unique changing-state and steady-state recordings were selected. Half of the speech samples were presented with voiced phonation and half were presented with whispered phonation. Each sentence or word sequence was selected once in the male and once in the female voice, thus generating 16 unique speech recordings for each auditory condition (state × phonation).

The experiment was conducted on an HP EliteDesk computer in a testing cubical at the University of Lincoln. Visual stimuli were presented on an HP EliteDisplay E240 screen (24 in). An Intel Realtek audio controller was used and sounds were played dichotically via Sony MDR-ZX110 headphones at a level similar to Experiment 1 (approximately 70 dB(A) on average). The experiment was programmed in Python using PsychoPy (Peirce *et al.*, 2019).

3. Design and Procedure

The experimental design was the same as in Experiment 1, using five different auditory conditions (silence, voiced/whispered steady-state words, voiced/whispered changing-state speech). The procedure was identical to Experiment 1, except that the number of repetitions per experimental condition was increased to 16, resulting in a total of 80 trials. As in Experiment 1, unique sentences or unique steady-state words were presented on each trial. Moreover, half of the speech trials were presented by the male and female voice, respectively. The trial structure was also identical to Experiment 1, except that after each trial participants were asked to give a confidence judgment by clicking on a scale from 0 to 8, indicating "how many letters they thought to have recalled in the correct position". Feedback on the actual number of correct letters was presented after the confidence judgment.

A. Results

1. Whispering and changing-state effects with foreign language

In contrast to Experiment 1, an equivalent 2 (state: steady, changing) × 2 (phonation: voiced, whispered) × 8 (serial position) repeated-measures ANOVA revealed a significant main effect of state, $F(1,51) = 7.37$, $MSE = 0.03$, $p = .009$, $\hat{\eta}_p^2 = .126$, but no significant main effect of phonation, $F(1,51) = 2.85$, $MSE = 0.03$, $p = .097$, $\hat{\eta}_p^2 = .053$. As can be seen in Fig. 3A, in participants to whom the language is incomprehensible, whispered German speech tended to be less disruptive ($M = .53; SD = .12$) compared to voiced German speech ($M = .51, SD = .10$). The relative decrement in performance (compared to silence) was 12.8% for voiced speech and 10.4% for whispered speech. There was also no interaction between phonation and state in Experiment 2, $F(1,51) = 0.14$, $MSE = 0.03$, $p = .713$, $\hat{\eta}_p^2 = .003$.

The ANOVA also revealed a significant serial position effect, $F(2.54, 129.62) = 204.73$, $MSE = 0.13$, $p < .001$, $\hat{\eta}_p^2 = .801$, as well as an interaction between state and serial position, $F(5.21, 265.50) = 2.40$, $MSE = 0.01$, $p = .035$, $\hat{\eta}_p^2 = .045$. According to a planned contrasts analysis, the changing-state effect was significant only at the early serial positions 1 ($p_{\text{BH}(8)} = .046$), 2 ($p_{\text{BH}(8)} = .008$) and barely at position 3 ($p_{\text{BH}(8)} = .063$), but not at the

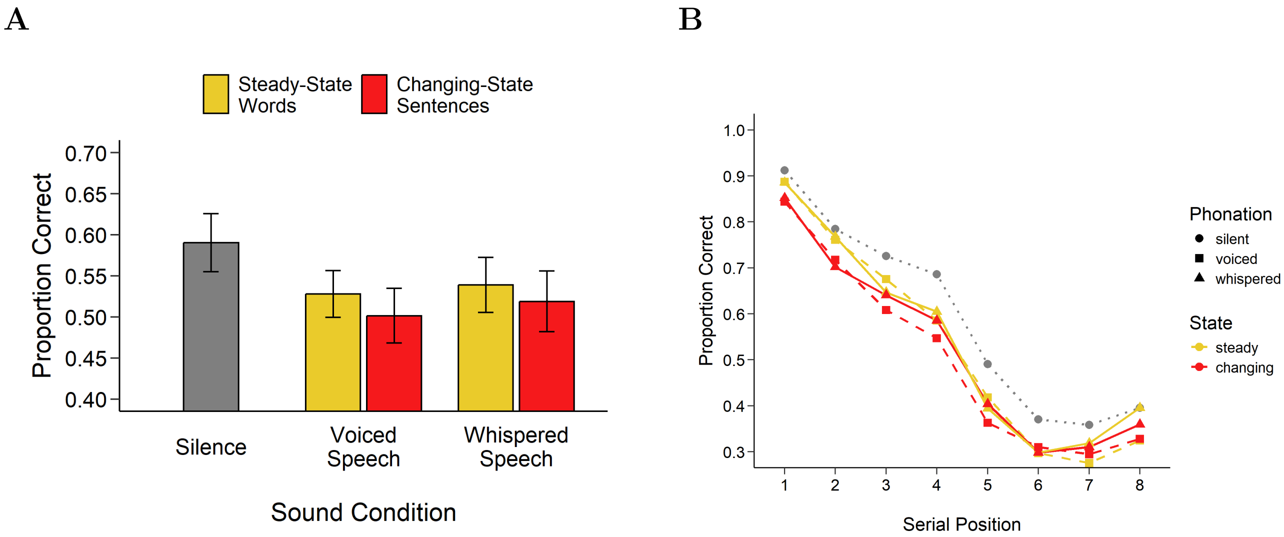

FIG. 3. **(A)** Mean serial recall accuracy in silence and when either steady-state or changing-state speech was presented with voiced or whispered phonation in a foreign language (German) during encoding in Experiment 2. Error bars represent 95% confidence intervals. **(B)** Serial recall accuracy in the five irrelevant sound conditions as a function of serial position.

later serial positions ($p_{\text{BH}(8)} \geq .179$). Consistent with Experiment 1, the interaction between phonation and serial position was not significant, $F(5.46, 278.29) = 2.04$, $MSE = 0.01$, $p = .067$, $\hat{\eta}_p^2 = .039$. However, we note that there was a non-significant trend towards whispered speech being less disruptive than voiced speech at the last serial position, $t(51) = -2.72$, $p_{\text{BH}(8)} = .071$, whereas all other contrasts were clearly non-significant ($p_{\text{BH}(8)} \geq .162$). This indicates that task-irrelevant whispered speech in a non-comprehended language is equally disruptive as voiced speech to the memorization of items from the beginning and the middle of the list, but it may restore the recency effect (compare Fig. 3B). There was also no significant three-way interaction, $F(4.84, 246.66) = 2.08$, $MSE = 0.01$, $p = .070$, $\hat{\eta}_p^2 = .039$.

2. Metacognitive confidence

To assess metacognitive awareness of the disruptive effects of changing-state and whispered speech, participants of Experiment 2 were also asked to indicate their confidence after each trial. A one-way repeated-measures ANOVA revealed a significant difference in confidence judgments between the five auditory conditions, $F(3.23, 164.95) = 19.69$, $MSE = 0.16$, $p < .001$, $\hat{\eta}_p^2 = .279$ (for descriptive statistics, see Table III). Planned contrasts revealed that confidence was higher in silence compared to both steady-state ($p < .001$) and changing-state speech ($p < .001$), but there was no significant difference in confidence between steady-state and changing-state conditions ($p = .082$; note however that it would be premature to conclude that participants did not notice the difference between the two conditions, see Bell *et al.*, 2022; Kattner and Bryce, 2022; Röer *et al.*, 2017b), nor between voiced and whispered speech conditions ($p = .217$). This indicates that participants were aware of the general disruption by task-irrelevant speech, but they did not notice the stronger impairment by specific types of speech (e.g., changing-state speech).

TABLE III. Means and standard deviations of confidence judgments of serial recall in silence and during the presentation of steady-state words or changing-state sentences with voiced or whispered phonation in Experiment 2.

	silence	**steady-state**		**changing-state**	
		voiced	**whispered**	**voiced**	**whispered**
M	4.07	3.61	3.66	3.52	3.59
SD	1.23	1.02	1.06	1.07	1.13

3. *Cross-experiment analysis*

To directly compare the effects of whispered speech in a comprehensible and incomprehensible language (i.e., Experiment 1 vs. 2), an additional 2 (experiment) × 2 (state) × 2 (phonation) mixed-factors ANOVA was conducted with experiment as a between-subjects factor and state and phonation as within-subject factors. The analysis revealed that there was no main effect of phonation, $F(1, 144) = 0.25$, $MSE = 0.01$, $p = .618$, $\hat{\eta}^2_p = .002$, but a significant interaction between phonation and experiment, $F(1, 144) = 7.50$, $MSE = 0.01$, $p = .007$, $\hat{\eta}^2_p = .049$, suggesting that whispered speech was more disruptive than voiced speech when the language is intelligible (Experiment 1), but not when it is incomprehensible to participants (Experiment 2). Planned contrasts (corrected according to Benjamini and Hochberg, 1995) revealed that there was a significant difference between whispered and voiced phonation in Experiment 1, $t(144) = 2.71$, $p_{\mathrm{BH(2)}} = .015$, but not in Experiment 2, $t(144) = -1.39$, $p_{\mathrm{BH(2)}} = .165$. Interestingly, in addition to the main effect of state, $F(1, 144) = 40.79$, $MSE = 0.01$, $p < .001$, $\hat{\eta}^2_p = .221$, the ANOVA also revealed a significant interaction between state and experiment, $F(1, 144) = 7.50$, $MSE = 0.01$, $p = .007$, $\hat{\eta}^2_p = .049$, indicating that the magnitude of the changing-state effect differed also between experiments. Planned contrasts revealed that the changing-state effect was significant in both experiments, but larger in Experiment 1, $t(144) = 7.60$, $p_{\mathrm{BH(2)}} < .001$, than in Experiment 2, $t(144) = 2.31$, $p_{\mathrm{BH(2)}} = .022$. There were no other significant effects.

4. Psychoacoustical predictors

To test whether the disruption of serial recall can be predicted by the psychoacoustic properties of irrelevant speech, a backward stepwise multiple linear regression analysis was conducted to predict the average serial recall accuracy associated with each sound file that was presented in the two experiments. In addition to the three dummy-coded categorical predictors phonation (0 = voiced, 1 = whispered), speaker gender (0 = male, 1 = female) and experiment (0 = Exp. 1, 1 = Exp. 2), the z-transformed psychoacoustic metrics 'Zwicker' loudness, sharpness, roughness, fluctuation strength, relative approach (i.e., spectro-temporal variation determined with the 'Relative Approach' method) and tonality were entered as continuous predictor variables. The starting model also contained interaction terms for each psychoacoustic metric with experiment (except loudness due to a high variable inflation factor), but other interaction terms were not included due to multicollinearity (as indicated by variable inflation factors). The regression analysis revealed that the best-fitting model includes only the intercept ($b = 0.55$, 95% CI $[0.50, 0.60]$) and experiment ($b = 0.01$, 95% CI $[-0.02, 0.04]$), fluctuation strength ($\beta = 0.05$, 95% CI $[0.03, 0.08]$) and relative approach ($\beta = -.001$, $t(139) = -1.97$, $p = .050$) as well as the interaction term between fluctuation strength and experiment ($\beta = -0.05$, 95% CI $[-0.08, -0.01]$) as predictors of serial recall accuracy, $R^2 = .18$, $F(4, 139) = 7.43$, $p < .001$.

To further investigate the predictive power of individual psychoacoustic metrics while avoiding multicollinearity, additional backward step-wise regression analyses were conducted for each psychoacoustic predictor variable including the respective two-way interaction terms with experiment, phonation and speaker gender.

'Zwicker' loudness was found to be a small but significant predictor, $\beta = 0.00$, 95% CI $[0.00, 0.00]$, $t(141) = 2.76$, $p = .007$, and together with an interaction term with experiment, $\beta = 0.00$, 95% CI $[0.00, 0.00]$, $t(141) = -2.86$, $p = .005$, it accounted for about 8% of the variance in serial recall accuracy, $R^2 = .08$, $F(2, 141) = 6.36$, $p = .002$. As can be seen in Fig. 4, increasing loudness was associated with better performance in the serial recall task, and this relationship was stronger with comprehensible speech in Experiment 1 than in Experiment 2. The same predictive relationships were found also for a model including A-weighted *sound pressure level* and its interaction with experiment, which accounted for even 10% of the variance, $R^2 = .10$, $F(2, 141) = 7.81$, $p < .001$ (see Fig. 4).

Fluctuation strength was also found to be an important predictor, with increasing fluctuation strength predicting higher recall accuracy (see Fig. 5; in contrast to the assumption of stronger disruption of serial recall by sounds of higher fluctuation strength, Schlittmeier *et al.*, 2012), $\beta = 0.05$, 95% CI $[0.03, 0.07]$, $t(140) = 5.08$, $p < .001$. In addition, there was a significant interaction term between fluctuation strength and experiment, $\beta = -0.04$, 95% CI $[-0.06, -0.02]$, $t(140) = -3.42$, $p < .001$, indicating that the predictive power of fluctuation strength was stronger in Experiment 1 (see Fig. 5). This suggests that the positive relationship between fluctuation strength and recall accuracy is stronger in participants who are able to understand the distractor language, and it also reflects the fact that there was only a changing-state effect but no whispering effect in the absence of speech comprehension (Experiment 2). The model also contained a small, but non-significant interaction term between fluctuation strength and phonation, $\beta = 0.03$, 95% CI $[0.00, 0.06]$, $t(140) = 1.84$, $p = .068$, but this may be biased by the lower and smaller range of fluctuation strength in whispered speech sounds (see Fig. 5). Together, the fluctuation strength model accounted for 17% of the variance in serial recall, $R^2 = .17$, $F(3, 140) = 9.55$, $p < .001$.

The regression model with *relative approach* as an indicator of spectro-temporal variation in the signal accounted for 8% of the variance in serial recall, $R^2 = .08$, $F(3, 140) = 3.85$, $p = .011$, with higher relative approach predicting lower serial recall accuracy (see Fig. 5). However, both the main effect of relative approach, $\beta = 0.00$, 95% CI $[0.00, 0.00]$, $t(140) = -2.03$, $p = .045$, and its interaction with experiment, $\beta = 0.00$, 95% CI $[0.00, 0.00]$, $t(140) = -2.33$, $p = .021$, were both rather small though significant predictors.

Psychoacoustic *roughness* was only a small and non-significant predictor of serial recall accuracy, $\beta = 0.02$, 95% CI $[-0.01, 0.05]$, $t(141) = 1.45$, $p = .148$, but together with the interaction with experiment, $\beta = -0.04$, 95% CI $[-0.07, -0.01]$, $t(141) = -2.94$, $p = .004$, it also accounted for some variance in serial recall accuracy, $R^2 = .06$, $F(2, 141) = 4.32$, $p = .015$. This may indicate that the rougher voiced speech sounds were less disruptive with comprehensible speech in Experiment 1, whereas they tend to be more disruptive with incomprehensible speech in Experiment 2 (see also Fig. 5).

The model with *sharpness*, $\beta = 0.00$, 95% CI $[-0.04, 0.05]$, $t(141) = 0.17$, $p = .867$, and its interaction term with experiment, $\beta = -0.01$, 95% CI $[-0.02, 0.00]$, $t(141) = -2.00$, $p = .047$, did not achieve a significant fit and accounted only for a small portion of the variance in serial recall, $R^2 = .03$, $F(2, 141) = 2.01$, $p = .138$.

The same is true for *tonality*, $R^2 = .03$, $F(2, 141) = 2.53$, $p = .084$, which tends to be positively related with serial recall accuracy, $\beta = 0.03$, 95% CI $[0.00, 0.06]$, $t(141) = 1.90$, $p = .060$. Moreover, the relationship between tonality and recall also differed between experiments, $\beta = -0.03$, 95% CI $[-0.06, 0.00]$, $t(141) = -1.98$, $p = .049$.

B. Discussion

Experiment 2 revealed that whispered German speech was equally disruptive as voiced speech to listeners who did not understand the language. This suggests that listening to whispered phonation may not demand additional listening effort compared to voiced phonation if participants cannot understand the language of the whispered speech. If anything, it tends to be even less disruptive, presumably due to the reduced spectro-temporal variation and lower amplitude modulations, indicating that with an incomprehensible language, disruption may be driven primarily by psychoacoustic properties of irrelevant sound.

While whispered speech was equally disruptive as voiced speech, indicating no (additional) attentional capture with incomprehensible speech, Experiment 2 still revealed a clear changing-state effect, indicating interference-by-process (Jones and Tremblay, 2000). However, although almost the same speech recordings were used, the size of the changing-state effect was larger in Experiment 1 than in Experiment 2, indicating that the interference with order processing may be more pronounced in a language that is comprehensible to participants. This could be explained with more efficient auditory grouping of speech tokens in a familiar language, thus forming a more stable auditory stream and in case of a changing-state stream more disruption of serial-order processing.

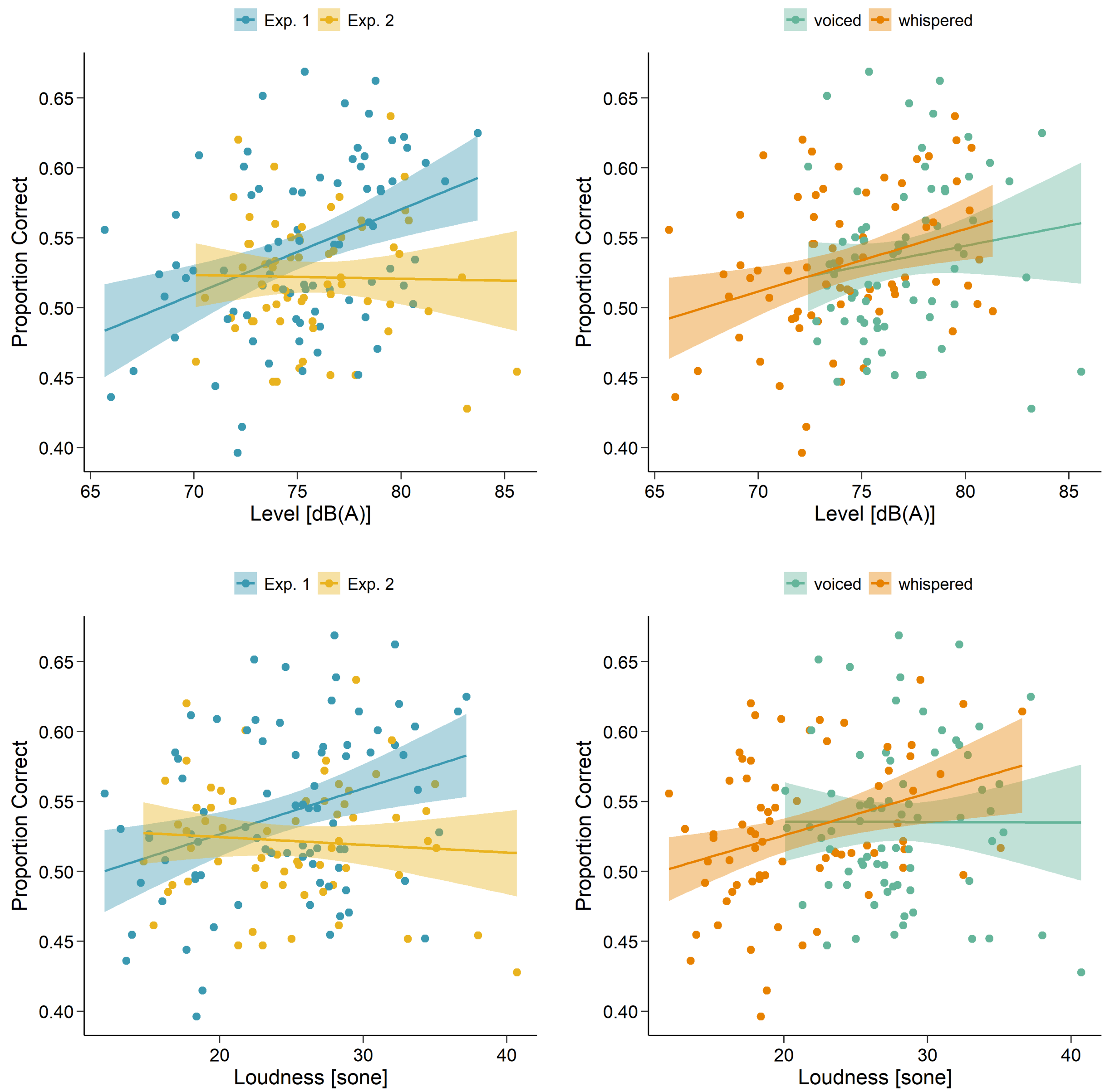

FIG. 4. Serial recall accuracy predicted by the sound pressure level [dB(A)] and ‘Zwicker’ loudness [sone] of whispered and voiced task-irrelevant German speech presented in Experiment 1 (German listeners) and 2 (foreign listeners).

IV. GENERAL DISCUSSION

The present study investigated whether task-irrelevant whispered speech produces more or less disruption of serial recall from visual-verbal short-term memory compared to voiced

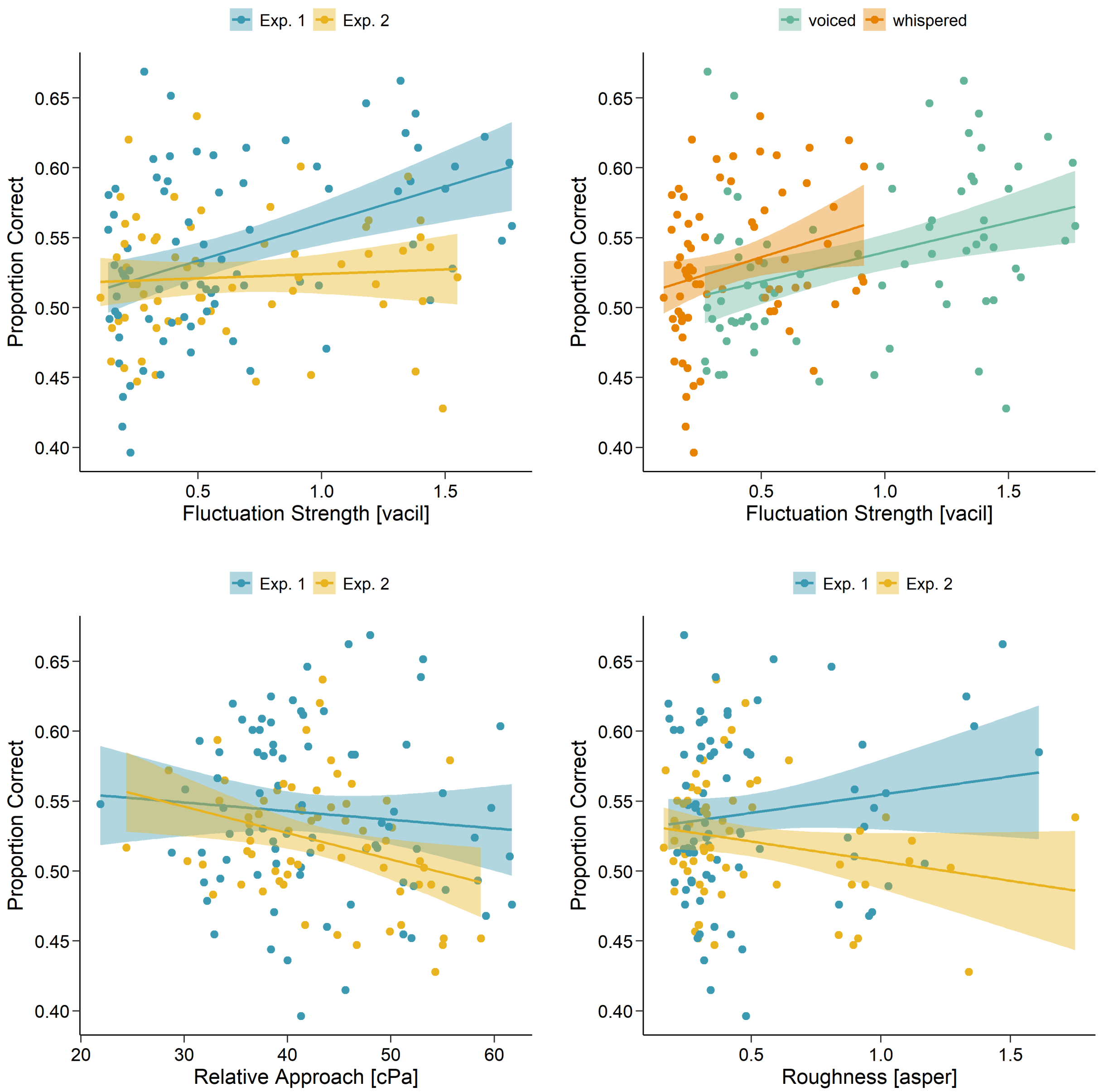

FIG. 5. Serial recall accuracy predicted by fluctuation strength [vacil], relative approach [cPa], and roughness [asper] of whispered and voiced German speech presented in Experiment 1 (German listeners) and 2 (foreign listeners).

speech. Specifically, according to an attentional account of auditory distraction, whispered speech could be expected to capture more attentional-cognitive resources than voiced speech, either due to its potential self-relevance (e.g., Röer *et al.*, 2013, 2017a) or because it requires additional listening effort to process 'degraded' whispered speech in a comprehensible language (e.g., due to the missing harmonic structure). In contrast, according to an interference-by-process account, less disruption by whispered speech would be expected, given that whispered speech sounds with a reduced level, amplitude envelope and fluctuation strength may cause less interference with serial-order processing (compare Alikadic and Röer, 2022; Jones *et al.*, 2000). In line with the attentional account, it was found in Experiment 1 that whispered speech in a comprehensible language causes about 35% more disruption of serial recall than voiced speech (i.e., the relative accuracy decrements were 9.6% with voiced speech and 12.9% with whispered speech). Interestingly, the disruptive effect of whispered speech was found to be independent of the changing-state effect. That is, whispered speech was more disruptive both in steady-state sequences (repetitions of a single monosyllabic word) and changing-state spoken sentences, and changing-state speech was more disruptive than steady-state speech regardless of whether the phonation type was whispered or voiced. This suggests that distraction by changing-state speech and distraction by whispered speech is the result of two distinct mechanisms: While changing-state speech is more disruptive due to interference with deliberate serial-order processing (i.e., interference-by-process Jones *et al.*, 1996; Marsh *et al.*, 2009), whispered phonation may cause additional disruption due to attentional capture or by demanding cognitive resources to process whispered speech in a comrehensible language (i.e., speech decoding and lexical access, which may be a more automatic / less conscious process compared to attentional capture). Interestingly, the magnitude of the changing-state effect was larger than the magnitude of the whispering effect ($\hat{\eta}_p^2$ is about 5.5 times larger for the changing-state effect in

Experiment 1), suggesting that interference-by-process causes considerably more disruption than attentional capture by specific features of a whispered voice. It is also possible that attentional capture by whispered speech has been partially reduced through the listeners' cognitive control (depending on their available working memory capacity during the task; e.g., Kattner, 2021; Sörqvist, 2010) (but see Körner *et al.*, 2017), whereas the more automatic interference-by-process probably cannot be reduced at a cognitive level (Hughes *et al.*, 2013). There are several possibilities concerning the cues in whispered speech that may capture attention. Whispered speech is often used to convey secret or personal information and it may thus be considered as potentially more important or self-relevant to a listener. Similarly, whispering may indicate social exclusion from a group and thus trigger an emotional response that directs attention to whispered speech. However, it is also possible that it just requires more listening effort and thus attentional control to process and understand whispered speech due to the absence of certain phonetic cues (e.g., the periodic excitation pattern and harmonic structure of modal speech).

Experiment 2 was conducted to test whether the impairment of serial short-term memory with whispered speech may be related to attentional capture. Therefore, the same German irrelevant speech materials were presented to participants who did not understand the language. In a foreign and therefore incomprehensible language, participants are not expected to engage in additional listening effort when processing whispered speech than when processing voiced speech (e.g., in line with the 'framework for understanding effortful listening' or the 'ease of language understanding' account Pichora-Fuller *et al.*, 2016; Rönnberg *et al.*, 2013). Moreover, whispering in an incomprehensible language may not be a useful cue of enhanced importance or self-relevance of the 'task-irrelevant' information. Thus, whispered speech would not be expected to capture more attention than voiced speech when presented in an incomprehensible language. In contrast, interference with serial order processing (due to the changing-state nature of speech) should be unaffected by a change of the language (e.g., Jones *et al.*, 1990). It was found in Experiment 2 that serial recall was disrupted by the presence of changing-state speech (compared to steady-state speech) – though less than in Experiment 1 – but it was not affected by the phonation type of irrelevant speech. Whispered speech even tended to be less disruptive than voiced speech, presumably due to the lower level or spectro-temporal variation reducing interference with serial-order processing (compare Alikadic and Röer, 2022). The results of Experiment 2 appear to rule out the notion that whispered speech produces greater disruption than normally phonated speech due to the triggering of affective responses - since these should arguably transcend language.

The greater disruption produced by whispered against normally phonated speech for native language listeners coheres with previous findings, demonstrating that meaningful sentences produce greater disruption of serial recall than incomprehensible degraded speech or random sequences of spoken syllables, presumably due to higher familiarity or interest (e.g., Hughes and Marsh, 2020; Kattner *et al.*, 2022). Moreover, the results also gel with the finding that ignoring a telephone conversation whereby only one of the two speakers was audible produced more disruption to a visually-based task than the same conversation wherein both speakers could be heard (Emberson *et al.*, 2010; Marsh *et al.*, 2018). Similarly, it has been reported that disruption decreases with an increasing number of voices in multi-speaker speech babble background situations (Jones and Macken, 1995; Zaglauer *et al.*, 2017). All three instances (whispered speech, single-sided telephone conversations and multi-speaker background babble) contain intelligible/semi-intelligible speech that involuntarily engages the listener's attention due to the semantic content (or the potential for meaning). All three types of speech engage cognitive processes related to understanding language, even when the listener is trying to focus on an unrelated, visual task. The same pattern did not emerge when the speech was incomprehensible (Marsh *et al.*, 2018), suggesting that the semantic properties of the half conversation generated a "need to listen" or "involuntary eavesdropping". It is possible that whispered speech, meaningful to the listener, provokes a similar mechanism of attentional diversion. In contrast to previous findings, however, whispered stimuli do not have to be semantically rich to attract attention. In Experiment 1 the whispering effect was similar in magnitude for sequences comprising a repeated single word (e.g., "hand") as it was for multi-word semantically rich sentences (e.g., prose). Thus, it would appear that lexical identification of a single-item is sufficient to drive the additional disruption produced by whispering as compared to normally phonated speech (possibly due to increased cognitive demand for successful decoding and lexical access, see Pichora-Fuller *et al.*, 2016; Rönnberg *et al.*, 2013).

The notion that the whispering effect emerges due to the recruitment of more listening effort for lexical-semantic identification of whispered speech, implies that similar disruptive effects should be observed for speech that is rendered slightly less intelligible via other means of acoustical manipulation. However, such a pattern appears to be absent from previous studies in which the degree of disruption in serial recall typically declines with continuous degradation of the speech signal (lower numbers of frequency bands in noise-vocoded speech or longer segment durations in locally time-reversed speech; see Ellermeier *et al.*, 2015; Ueda *et al.*, 2019). This may be because there is an optimal level of intelligibility required for the recruitment of listening effort, or because some other factor (e.g., socio-affective; Cirillo and Todt, 2005; Laver, 1994) provokes greater listening effort.

Disruption of serial recall in the present experiments was also related to certain psychoacoustic properties of the irrelevant speech sounds, particularly sound pressure level, loudness, fluctuation strength, and tonality. However, in contrast to previous reports of louder sounds being equally or more disruptive (Alikadic and Röer, 2022; Ellermeier and Hellbrück, 1998), recall accuracy (not distraction) increased with both the sound pressure level and the loudness of the irrelevant speech samples – in particular in Experiment 1. This finding most likely reflects the fact that whispered speech, when presented in a comprehensible language, was more disruptive despite its lower intensity and reduced vowel amplitudes compared to voiced speech (but see Hughes *et al.*, 2005a). In future work, it may be worth validating whether the disruptive effects of whispered speech in a comprehensible language hold true when whispering is presented at more realistic playback levels. Similarly, irrelevant speech with higher fluctuation strength also led to higher recall accuracy in Experiment 1, but not in Experiment 2. Hence, in line with other previous findings (Ellermeier *et al.*, 2015, also observing higher recall accuracy in the conditions with maximum fluctuation strength), fluctuation strength alone does not seem to be an appropriate predictor of auditory distraction (i.e., in the present experiments, it was associated with less distraction; in contrast to Schlittmeier *et al.*, 2012). A similar relationship was observed also between roughness and memory performance, with voiced speech being characterized by higher roughness – in particular for the male voice – which was associated with higher recall accuracy when the language was comprehensible (Experiment 1), but with lower accuracy when the language was incomprehensible (Experiment 2). Finally, higher tonality was also associated with higher accuracy in the serial recall task, indicating that the tonality of voiced speech (characterized by more pronounced harmonics) does not necessarily produce more disruption in a serial recall task. Specifically, it appears that certain cues in comprehensible whispered speech (e.g., semantics or social-cognitive aspects) capture attention and produce even more disruption compared to the acoustically driven interference due to changes in vowel amplitudes in voiced speech.

While the results of Experiments 1 and 2 support the notion that the changing-state effect and the whispering effect are underpinned by distinct cognitive mechanisms, further studies could add weight to the proposed dichotomy of distraction effects. Previous research suggests that the changing-state effect occurs most prominently in tasks drawing on serial rehearsal (e.g., Beaman and Jones, 1997), whereas attentional capture effects should arise for any cognitively demanding task (e.g., Vachon *et al.*, 2017). If the whispering effect for native language listeners is indeed attributable to attentional diversion then it should also be observed on focal tasks that do not draw upon serial processing, such as the missing-item task (Hughes *et al.*, 2007; Jones and Macken, 1993). Further, the whispering effect unlike the changing-state effect should be influenced by extrinsic or intrinsic cognitive control. For example, the magnitude of the whispering effect for native listeners should be reduced under high task-encoding or cognitive load (see Hughes *et al.*, 2013; Marsh *et al.*, 2020, 2018) and for individuals with higher working memory capacity (which reflects a trait capacity for cognitive/attentional control; Hughes *et al.*, 2013; Marsh *et al.*, 2017; Sörqvist *et al.*, 2012). Furthermore the whispering effect should be reduced by previous exposure to distractors (i.e., foreknowledge) which has been shown to reduce the additional disruption produced by comprehensible over incomprehensible spoken sentences (Kattner *et al.*, 2022) and emotional (e.g., taboo) over neutral words (Rettie *et al.*, 2024), through reducing the personal relevance, interest (Hughes and Marsh, 2020; Kattner *et al.*, 2022), or affective responses (Rettie *et al.*, 2024) produced by the stimuli.

V. CONCLUSION

Taken together the present study shows that task-irrelevant whispered speech can be more – not less – disruptive to cognitive performance when the language is comprehensible to the listener, but not when the listeners did not understand the language. This suggests that certain semantic and/or social-cognitive features conveyed by whispered voices may capture attention or encourage enhanced listening effort, leading to a lack of cognitive resources being available for the focal short-term memory task. In line with an attentional interpretation of the whispering effects, distraction did not increase with psychoacoustic loudness or fluctuation strength – as would have been predicted by a unitary interference-by-process account of auditory distraction (i.e., whispered speech is softer, less tonal, and less fluctuating and should therefore cause less interference with serial-order processing). At the same time, it was observed that changing-state speech is more disruptive than steady-state speech regardless of whether the phonation was voiced or whispered. This indicates two functionally distinct and additive mechanisms of distraction, with one being based on interference between auditory grouping (of changing-state sounds) and deliberate seriation processes, and the other being based on attentional capture by whispered voices.

The findings of the current study may have significant practical implications across various real-world contexts, particularly when the whispered speech is intelligible to listeners. While whispering is often employed in open office settings to lower conversational volume for politeness and to avoid disturbing others, as well as to communicate sensitive information (Cirillo, 2004; Cirillo and Todt, 2005), our study reveals that intelligible whispered speech may attract more attention and lead to increased errors and reduced efficiency compared to voiced speech. Over time, these declines in cognitive performance could result in substantial costs for businesses, highlighting the need for sound management strategies or workspace redesigns, such as designated quiet zones, to minimize disruptive intelligible whispers. The results of the present study suggest that whispered conversations can be more distracting than previously thought, calling into question the effectiveness of such policies in maintaining a focused environment. Similarly, in educational environments, intelligible whispers might hinder classroom learning and academic achievement, necessitating strict noise management during study and exam sessions. Whispering during lectures, a common and allegedly unproblematic behavior, may negatively influence academic performance of other students and/or disrupt the lecturer even more than voiced conversations. Also the concept of 'whisper zones' in libraries, which are intended to minimize disruption with reading, may be based on a misguided assumption. In high-stakes cognitive settings, such as hospital operating rooms or control rooms, whispering may disrupt critical tasks, underscoring the importance of stringent sound management policies in these areas. Moreover, while whispered speech is commonly used to maintain privacy in public or semi-public spaces like libraries, our research suggests that intelligible whispers may undermine this intent and disrupt others more than anticipated. Organizations should encourage staff to reconsider the use of whispered conversations and explore physical barriers or soundproofing options to enhance privacy and minimize disruption. Finally, the tendency for intelligible whispering to be more disruptive than voiced speech has implications for individuals with breathy or soft voices, whether due to natural variations or clinical conditions characterized by breathiness (e.g., vocal fold paralysis; Macdonell and Holmes, 2007). Such individuals may inadvertently create more disruptions than those with clearer vocal tones, especially if their speech is easily understood. This highlights the potential need for voice training or sound management strategies in shared environments.

VI. AUTHOR DECLARATIONS

A. Conflict of Interest

The authors have no conflicts of interest to declare.

B. Ethics Approval

The two experiments in this study were conducted strictly in accordance with the Ethical Principles of the Acoustical Society of America and the Declaration of Helsinki. All participants were informed about the duration and procedure, potential risks, data protection regulations, and their right to withdraw from participating at any time, without consequence. Written informed consent was obtained prior to the start of the experiment. The experimental protocols were approved by the ethics committee of the University of Lincoln (ref: 33415).

VII. DATA AVAILABILITY

The data and analysis scripts of the experiments in this study are available upon request from the corresponding author.

ACKNOWLEDGMENTS

We are thankful to our student research assistants Leonardo Stuff, Ramona Spitz, Patrizia Scholz, Cosima Stokar von Neuforn, Marko Sanden, and Maike Gerlach for their help with the production of speech recordings, the recruitment of participants and the data collection of Experiment 1. We also thank Lia Downing, Veronika Schaer, Mia Dennis, and Joshua Dudley for their help with the recruitment of participants and the data collection for Experiment 2.

VIII. REFERENCES

Alikadic, L., and Röer, J. P. (**2022**). "Loud Auditory Distractors Are More Difficult to Ignore after All: A Preregistered Replication Study with Unexpected Results," Experimental Psychology **69**(3), 163–171, doi: 10.1027/1618-3169/a000554.

Beaman, C. P., and Jones, D. M. (**1997**). "Role of serial order in the irrelevant speech effect: Tests of the changing-state hypothesis," Journal of Experimental Psychology: Learning Memory and Cognition **23**(2), 459–471, doi: 10.1037/0278-7393.23.2.459.

Bell, R., Buchner, A., and Mund, I. (**2008**). "Age-Related Differences in Irrelevant-Speech Effects," Psychology and Aging **23**(2), 377–391, doi: 10.1037/0882-7974.23.2.377.

Bell, R., Mieth, L., Röer, J. P., and Buchner, A. (**2022**). "The metacognition of auditory distraction: Judgments about the effects of deviating and changing auditory distractors on cognitive performance," Memory & Cognition 2021 **50**(1), 1–14, doi: 10.3758/s13421-021-01200-2.

Bell, R., Röer, J. P., Dentale, S., and Buchner, A. (**2012**). "Habituation of the irrelevant sound effect: Evidence for an attentional theory of short-term memory disruption," Journal of Experimental Psychology: Learning Memory and Cognition **38**(6), 1542–1557, doi: 10.1037/a0028459.

Bell, R., Röer, J. P., Lang, A. G., and Buchner, A. (**2019**). "Distraction by steady-state sounds: Evidence for a graded attentional model of auditory distraction," Journal of Experimental Psychology: Human Perception and Performance **45**(4), 500–512, doi:

10.1037/xhp0000623.

Benjamini, Y., and Hochberg, Y. (**1995**). "Controlling the false discovery rate: A practical and powerful approach to multiple testing," Journal of the Royal Statistical Society, Series B **57**(1), 289 – 300, doi: 10.2307/2346101.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The perceptual organization of sound* (MIT Press, Cambridge, MA).

Campbell, T., Beaman, C. P., and Berry, D. C. (**2002**). "Auditory memory and the irrelevant sound effect: Further evidence for changing-state disruption," Memory **10**(3), 199–214, doi: 10.1080/09658210143000335.

Cirillo, J. (**2004**). "Communication by unvoiced speech: the role of whispering," Annals of the Brazilian Academy of Sciences **76**, 413–423, doi: 10.1590/S0001-37652004000200034.

Cirillo, J., and Todt, D. (**2005**). "Perception and judgement of whispered vocalisations," Behaviour **142**(1), 113–129, doi: 10.1163/1568539053627758.

Colle, H. A., and Welsh, A. (**1976**). "Acoustic masking in primary memory," Journal of Verbal Learning and Verbal Behavior **15**(1), 17–31, doi: 10.1016/S0022-5371(76)90003-7.

Cowan, N. (**1995**). *Attention and Memory: An Integrated Framework* (Oxford University Press).

DIN Deutsches Institut für Normung e.V. (**2009**). "Measurement technique for the simulation of the auditory sensation of sharpness," Standard DIN 45692:2009-08.

DIN Deutsches Institut für Normung e.V. (**2010**). "Calculation of loudness level and loudness from the sound spectrum - zwicker method - amendment 1: Calculation of the loudness of time-variant sound; with cd-rom," Standard DIN 45631/A1:2010-03.

Dorsi, J., Viswanathan, N., Rosenblum, L. D., and Dias, J. W. (**2018**). “The role of speech fidelity in the irrelevant sound effect: Insights from noise-vocoded speech backgrounds,” Quarterly Journal of Experimental Psychology **71**(10), 2152–2161, doi: 10.1177/1747021817739257.

Eckert, H., and Laver, J. (**1994**). *Menschen und ihre Stimmen: Aspekte der vokalen Kommunikation* (Beltz/Psychologie Verlags Union).

ECMA International (**2022**). “Psychoacoustic metrics for itt equipment - part 2 (models based on human perception),” Standard ECMA 418-2 (2nd edition).

Eimer, M., Nattkemper, D., Schröger, E., and Prinz, W. (**1996**). “Involuntary attention,” in *Handbook of Perception and Action*, edited by O. Neumann and F. Sanders (Academic Press, London, UK), Chap. 5, pp. 389–446, doi: 10.1016/S1874-5822(96)80022-3.

Eklund, I., and Traunmüller, H. (**1997**). “Comparative Study of Male and Female Whispered and Phonated Versions of the Long Vowels of Swedish,” Phonetica **54**(1), 1–21, doi: 10.1159/000262207 publisher: De Gruyter Mouton.

Ellermeier, W., and Hellbrück, J. (**1998**). “Is Level Irrelevant in ”Irrelevant Speech”? Effects of Loudness, Signal-to-Noise Ratio, and Binaural Unmasking,” Journal of Experimental Psychology: Human Perception and Performance **24**(5), 1406–1414, doi: 10.1037/0096-1523.24.5.1406.

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,” The Journal of the Acoustical Society of America **138**(3), 1561–1569, doi: 10.1121/1.4928954.

Ellermeier, W., and Zimmer, K. (**1997**). “Individual differences in susceptibility to the ’irrelevant speech’ effect,” Journal of the Acoustical Society of America **102**, 2191–2199, doi: 10.1121/1.419596.

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

Emberson, L. L., Lupyan, G., Goldstein, M. H., and Spivey, M. J. (**2010**). “Overheard Cell-phone Conversations: When Less Speech is More Distracting,” Psychological Science **21**(10), 1383–1388, doi: 10.1177/0956797610382126.

Fastl, H. (**1982**). “Fluctuation strength and temporal masking patterns of amplitude-modulated broadband noise,” Hearing Research **8**(1), 59–69, doi: 10.1016/0378-5955(82)90034-X.

Fastl, H., and Zwicker, E. (**2007**). *Psychoacoustics: Facts and models*, 3rd ed. ed. (Springer, Heidelberg, Germany).

Frühholz, S., Trost, W., and Grandjean, D. (**2016**). “Whispering - The hidden side of auditory communication,” NeuroImage **142**, 602–612, doi: 10.1016/j.neuroimage.2016.08.023.

Genuit, K. (**1996**). “Objective Evaluation of Acoustic Quality Based on a Realtive Approach,” in *Inter-Noise’96, 25th Anniversary Congress Liverpool*, pp. 1061 p1 – 1061 p6.

Günther, F., Müller, H. J., and Geyer, T. (**2017**). “Salience, attention, and perception,” in *Entrenchment and the psychology of language learning: How we reorganize and adapt linguistic knowledge*, Language and the human lifespan series (De Gruyter Mouton, Boston, MA, US), pp. 289–312, doi: 10.1037/15969-014.

Hadlington, L., Bridges, A. M., and Darby, R. J. (**2004**). "Auditory location in the irrelevant sound effect: The effects of presenting auditory stimuli to either the left ear, right ear or both ears," Brain and Cognition **55**, 545–557, doi: 10.1016/j.bandc.2004.04.001.

Heeren, W. F. L. (**2015**). "Vocalic correlates of pitch in whispered versus normal speech," The Journal of the Acoustical Society of America **138**(6), 3800–3810, doi: 10.1121/1.4937762.

Hughes, R. W. (**2014**). "Auditory distraction: A duplex-mechanism account," PsyCh Journal **3**(1), 30–41, doi: 10.1002/pchj.44.

Hughes, R. W., Hurlstone, M. J., Marsh, J. E., Vachon, F., and Jones, D. M. (**2013**). "Cognitive control of auditory distraction: impact of task difficulty, foreknowledge, and working memory capacity supports duplex-mechanism account.," Journal of experimental psychology. Human perception and performance **39**(2), 539–553, doi: 10.1037/a0029064.

Hughes, R. W., and Marsh, J. E. (**2019**). "Dissociating two forms of auditory distraction in a novel Stroop serial recall experiment," Auditory Perception and Cognition **2**(3), 129–142.

Hughes, R. W., and Marsh, J. E. (**2020**). "When is forewarned forearmed? Predicting auditory distraction in short-term memory," Journal of Experimental Psychology: Learning Memory and Cognition **46**(3), 427–442, doi: 10.1037/xlm0000736.

Hughes, R. W., Tremblay, S., and Jones, D. M. (**2005**a). "Disruption by speech of serial short-term memory: The role of changing-state vowels," Psychonomic Bulletin and Review **12**, 886–890, doi: 10.3758/BF03196781.

Hughes, R. W., Vachon, F., and Jones, D. M. (**2005**b). "Auditory attentional capture during serial recall: Violations at encoding of an algorithm-based neural model?," Journal of Experimental Psychology: Learning, Memory, and Cognition **31**(4), 736–749, doi: 10.1037/0278-7393.31.4.736.

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.," Journal of Experimental Psychology: Learning, Memory, and Cognition **33**(6), 1050–1061, doi: 10.1037/0278-7393.33.6.1050.

Ito, T., Takeda, K., and Itakura, F. (**2005**). "Analysis and recognition of whispered speech," Speech Communication **45**(2), 139–152, doi: 10.1016/j.specom.2003.10.005.

Jones, D. M., Alford, D., Macken, W. J., Banbury, S. P., and Tremblay, S. (**2000**). "Interference from degraded auditory stimuli: Linear effects of changing-state in the irrelevant sequence," The Journal of the Acoustical Society of America **108**(3), 1082–1088, doi: 10.1121/1.1288412.

Jones, D. M., Beaman, C. P., and Macken, W. J. (**1996**). "The object-oriented episodic record model," in *Models of short-term memory*, edited by S. E. Gathercole (Psychology Press, Hove), pp. 209–238.

Jones, D. M., and 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., and Macken, W. J. (**1995**). "Auditory Babble and Cognitive Efficiency: Role of Number of Voices and Their Location," Journal of Experimental Psychology: Applied **1**(3), 216–226, doi: 10.1037/1076-898X.1.3.216.

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," Memory & Cognition **21**, 318–328, doi: 10.3758/BF03208264.

Jones, D. M., Madden, C., and Miles, C. (**1992**). "Privileged access by irrelevant speech to short-term memory: The role of changing state," The Quarterly Journal of Experimental Psychology **44A**, 645–669.

Jones, D. M., Miles, C., and Page, J. (**1990**). "Disruption of proofreading by irrelevant speech: Effects of attention, arousal or memory?," Applied Cognitive Psychology **4**, 89–108, doi: 10.1002/acp.2350040203.

Jones, D. M., and Tremblay, S. (**2000**). "Interference in memory by process or content? A reply to Neath (2000)," Psychonomic Bulletin and Review **7**(3), 550–558, doi: 10.3758/BF03214370.

Jovičić, S. T., and Šarić, Z. (**2008**). "Acoustic Analysis of Consonants in Whispered Speech," Journal of Voice **22**(3), 263–274, doi: 10.1016/j.jvoice.2006.08.012.

Kattner, F. (**2021**). "Transfer of working memory training to the inhibitory control of auditory distraction," Psychological Research **85**, 1–15, doi: 10.1007/s00426-020-01468-0.

Kattner, F. (**2024**). "False memories through auditory distraction: When irrelevant speech produces memory intrusions in the absence of semantic interference," Quarterly Journal of Experimental Psychology doi: 10.1177/17470218241235654.

Kattner, F., and Bryce, D. (**2022**). "Attentional control and metacognitive monitoring of the effects of different types of task-irrelevant sound on serial recall," Journal of Experimental Psychology: Human Perception and Performance **48**(2), 139–158, doi:

10.1037/xhp0000982.

Kattner, F., and Ellermeier, W. (**2014**). "Irrelevant speech does not interfere with serial recall in early blind listeners," Quarterly Journal of Experimental Psychology **67**(11), 2207–2217, doi: 10.1080/17470218.2014.910537.

Kattner, F., and Ellermeier, W. (**2018**). "Emotional prosody of task-irrelevant speech interferes with the retention of serial order," Journal of Experimental Psychology: Human Perception and Performance **44**(8), 1303–1312, doi: 10.1037/xhp0000537.

Kattner, F., Fischer, M., Caling, A. L., Cremona, S., Ihle, A., Hodgson, T., and Föcker, J. (**2024**). "The disruptive effects of changing-state sound and emotional prosody on verbal short-term memory in blind, visually impaired, and sighted listeners," Journal of Cognitive Psychology **36**, 28–41, doi: 10.1080/20445911.2023.2186771.

Kattner, F., Hanl, S., Paul, L., and Ellermeier, W. (**2023**). "Task-specific auditory distraction in serial recall and mental arithmetic," Memory and Cognition **51**(4), 930–951.

Kattner, F., Richardson, B. H., and Marsh, J. E. (**2022**). "The Benefit of Foreknowledge in Auditory Distraction Depends on the Intelligibility of pre-exposed Speech," Auditory Perception & Cognition **5**(3-4), 151–168, doi: 10.1080/25742442.2022.2089525.

Konno, H. (**2016**). "Analysis on Acoustical and Perceptual Characteristics of Whispered Speech and Whisper-to-Normal Speech Conversion" doi: 10.14943/doctoral.k12482.

Körner, U., Röer, J. P., Buchner, A., and Bell, R. (**2017**). "Working memory capacity is equally unrelated to auditory distraction by changing-state and deviant sounds," Journal of Memory and Language **96**, 122–137, doi: 10.1016/j.jml.2017.05.005.

Labonté, K., Marsh, J. E., and Vachon, F. (**2022**). “Distraction by auditory semantic deviations is unrelated to working memory capacity: Further evidence of a distinction between acoustic and categorical deviation effects,” Auditory Perception & Cognition .

Laver, J. (**1994**). Cambridge Textbooks in Linguistics *Principles of Phonetics* (Cambridge University Press), pp. 190–192.

LeCompte, D. C., Neely, C. B., and Wilson, J. R. (**1997**). “Irrelevant speech and irrelevant tones: The relative importance of speech to the irrelevant speech effect,” Journal of Experimental Psychology: Learning Memory and Cognition **23**(2), 472–483, doi: 10.1037/0278-7393.23.2.472.

Ljungberg, J. K., Parmentier, F. B., Hughes, R. W., Macken, W. J., and Jones, D. M. (**2012**). “Listen Out! Behavioural and Subjective Responses to Verbal Warnings,” Applied Cognitive Psychology **26**(3), 451–461, doi: 10.1002/acp.2818.

Macdonell, R. A., and Holmes, R. (**2007**). *Motor Speech and Swallowing Disorders*, 155–170 (Elsevier), doi: 10.1016/B978-0-323-03354-1.50016-X.

Marsh, J. E., Campbell, T., Vachon, F., Taylor, P., and Hughes, R. W. (**2020**). “How the deployment of visual attention modulates auditory distraction,” Attention, Perception, and Psychophysics **82**, 350–362, doi: 10.3758/s13414-019-01800-w.

Marsh, J. E., Hughes, R. W., and Jones, D. M. (**2009**). “Interference by process, not content, determines semantic auditory distraction,” Cognition **110**(1), 23–38, doi: 10.1016/j.cognition.2008.08.003.

Marsh, J. E., Vachon, F., and Sörqvist, P. (**2017**). “Increased distractibility in schizotypy: Independent of individual differences in working memory capacity?,” Quarterly Journal of

Experimental Psychology **70**(3), 565–578, doi: 10.1080/17470218.2016.1172094.

Marsh, J. E., Yang, J., Qualter, P., Richardson, C., Perham, N., Vachon, F., and Hughes, R. W. (**2018**). "Postcategorical auditory distraction in short-term memory: Insights from increased task load and task type.," Journal of Experimental Psychology: Learning, Memory, and Cognition **44**(6), 882–897, doi: 10.1037/xlm0000492.

Miles, C., Jones, D. M., and Madden, C. A. (**1991**). "Locus of the Irrelevant Speech Effect in Short-Term Memory," Journal of Experimental Psychology: Learning, Memory, and Cognition **17**(3), 578–584, doi: 10.1037/0278-7393.17.3.578.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., and Lindeløv, J. K. (**2019**). "PsychoPy2: Experiments in behavior made easy," Behavior Research Methods **51**, 195–203, doi: 10.3758/s13428-018-01193-y.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., and Wingfield, A. (**2016**). "Hearing impairment and cognitive energy: The framework for understanding effortful listening (fuel)," Ear and Hearing **37**, 5S–27S, doi: 10.1097/AUD.0000000000000312.

Rettie, L., Potter, R. F., Brewer, G., Degno, F., Vachon, F., Hughes, R. W., and Marsh, J. E. (**2024**). "Warning—taboo words ahead! Avoiding attentional capture by spoken taboo distractors," Journal of Cognitive Psychology **36**(1), 61–77, doi: 10.1080/20445911.2023.2285860.

Röer, J. P., Bell, R., and Buchner, A. (**2013**). "Self-relevance increases the irrelevant sound effect: Attentional disruption by one's own name," Journal of Cognitive Psychology **25**(8), 925–931, doi: 10.1080/20445911.2013.828063.

Röer, J. P., Bell, R., and Buchner, A. (**2015**). “Specific foreknowledge reduces auditory distraction by irrelevant speech,” Journal of Experimental Psychology: Human Perception and Performance **41**(3), 692–702, doi: 10.1037/xhp0000028.

Röer, J. P., Körner, U., Buchner, A., and Bell, R. (**2017**a). “Attentional capture by taboo words: A functional view of auditory distraction,” Emotion **17**(4), 740–750, doi: 10.1037/emo0000274.

Röer, J. P., Rummel, J., Bell, R., and Buchner, A. (**2017**b). “Metacognition in Auditory Distraction: How Expectations about Distractibility Influence the Irrelevant Sound Effect,” Journal of Cognition **1**(1), 2, doi: 10.5334/joc.3.

Rönnberg, J., Holmer, E., and Rudner, M. (**2021**). “Cognitive hearing science: three memory systems, two approaches, and the ease of language understanding model,” Journal of Speech, Language, and Hearing Research **64**, 359–370, doi: 10.1044/2020_JSLHR-20-00007.

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Örjan Dahlström, Signoret, C., Stenfelt, S., Pichora-Fuller, M. K., and Rudner, M. (**2013**). “The ease of language understanding (elu) model: Theory, data, and clinical implications,” Frontiers in Systems Neuroscience **7**, 48891, doi: 10.3389/fnsys.2013.00031.

Salamé, P., and Baddeley, A. D. (**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, doi: 10.1016/S0022-5371(82)90521-7.

Salamé, P., and Baddeley, A. D. (**1989**). "Effects of background music on phonological short-term memory," The Quarterly Journal of Experimental Psychology Section A **41A**(1), 107–122, doi: 10.1080/14640748908402355.

Schlittmeier, S. J., Hellbrück, J., and Klatte, M. (**2008**). "Does irrelevant music cause an irrelevant sound effect for auditory items?," European Journal of Cognitive Psychology **20**, 252–271, doi: 10.1080/09541440701427838.

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," Attention, Perception, and Psychophysics **74**(1), 194–203, doi: 10.3758/s13414-011-0230-7.

Sörqvist, P. (**2010**). "High working memory capacity attenuates the deviation effect but not the changing-state effect: Further support for the duplex-mechanism account of auditory distraction," Memory and Cognition **38**(5), 651–658, doi: 10.3758/MC.38.5.651.

Sörqvist, P., Nöstl, A., and Halin, N. (**2012**). "Working memory capacity modulates habituation rate: Evidence from a cross-modal auditory distraction paradigm," Psychonomic Bulletin and Review **19**, 245–250, doi: 10.3758/s13423-011-0203-9.

Tremblay, S., and Jones, D. M. (**1999**). "Change of intensity fails to produce an irrelevant sound effect: Implications for the representation of unattended sound," Journal of Experimental Psychology: Human Perception and Performance **25**(4), 1005–1015, doi: 10.1037/0096-1523.25.4.1005.

Tremblay, S., Nicholls, A. P., Alford, D., and Jones, D. M. (**2000**). "The Irrelevant Sound Effect: Does Speech Play a Special Role?," Journal of Experimental Psychology: Learning

Memory and Cognition **26**(6), 1750–1754, doi: 10.1037/0278-7393.26.6.1750.

Ueda, K., Nakajima, Y., Kattner, F., and Ellermeier, W. (**2019**). "Irrelevant speech effects with locally time-reversed speech: Native vs non-native language," The Journal of the Acoustical Society of America **145**(6), 3686, doi: 10.1121/1.5112774.

Vachon, F., Labonté, K., and Marsh, J. E. (**2017**). "Attentional capture by deviant sounds: A noncontingent form of auditory distraction?," Journal of Experimental Psychology: Learning Memory and Cognition **43**(4), 622–634, doi: 10.1037/xlm0000330.

Viswanathan, N., Dorsi, J., and George, S. (**2014**). "The role of speech-specific properties of the background in the irrelevant sound effect," Quarterly Journal of Experimental Psychology **67**(3), 581–589, doi: 10.1080/17470218.2013.821708.

Wingfield, A. (**2016**). "Evolution of models of working memory and cognitive resources," Ear and Hearing **37**, 35S–43S, doi: 10.1097/AUD.0000000000000310.

Zaglauer, M., Drotleff, H., and Liebl, A. (**2017**). "Background babble in open-plan offices: A natural masker of disruptive speech?," Applied Acoustics **118**, 1–7, doi: 10.1016/J.APACOUST.2016.11.004.

Zeamer, C., and Fox Tree, J. E. (**2013**). "The process of auditory distraction: Disrupted attention and impaired recall in a simulated lecture environment," Journal of Experimental Psychology: Learning Memory and Cognition **39**(5), 1463–1472, doi: 10.1037/a0032190.

Zhang, C., and Hansen, J. H. L. (**2018**). "Advancements in whispered speech detection for interactive/speech systems," in *Signal and Acoustic Modeling for Speech and Communication Disorders* (De Gruyter), pp. 9–32, doi: 10.1515/9781501502415-002.

Żygis, M., Pape, D., Koenig, L. L., Jaskuła, M., and Jesus, L. M. (**2017**). "Segmental cues to intonation of statements and polar questions in whispered, semi-whispered and normal speech modes," Journal of Phonetics **63**, 53–74, doi: 10.1016/j.wocn.2017.04.001.