Atten Percept Psychophys (2018) 80:222–241
DOI 10.3758/s13414-017-1425-3

Sound specificity effects in spoken word recognition: The effect of integrality between words and sounds

Dorina Strori[1] · **Johannes Zaar**[2] · **Martin Cooke**[3] · **Sven L. Mattys**[4]

Published online: 3 October 2017

Abstract Recent evidence has shown that nonlinguistic sounds co-occurring with spoken words may be retained in memory and affect later retrieval of the words. This sound-specificity effect shares many characteristics with the classic voice-specificity effect. In this study, we argue that the sound-specificity effect is conditional upon the context in which the word and sound coexist. Specifically, we argue that, besides co-occurrence, integrality between words and sounds is a crucial factor in the emergence of the effect. In two recognition-memory experiments, we compared the emergence of voice and sound specificity effects. In Experiment 1, we examined two conditions where integrality is high. Namely, the classic voice-specificity effect (Exp. 1a) was compared with a condition in which the intensity envelope of a background sound was modulated along the intensity envelope of the accompanying spoken word (Exp. 1b). Results revealed a robust voice-specificity effect and, critically, a comparable sound-specificity effect: A change in the paired sound from exposure to test led to a decrease in word-recognition performance. In the second experiment, we sought to disentangle the contribution of integrality from a mere co-occurrence context effect by removing the intensity modulation. The absence of integrality led to the disappearance of the sound-specificity effect. Taken together, the results suggest that the assimilation of background sounds into memory cannot be reduced to a simple context effect. Rather, it is conditioned by the extent to which words and sounds are perceived as integral as opposed to distinct auditory objects.

Keywords Spoken word recognition · Long-term memory · Speech perception

✉ Dorina Strori
dorina.strori@northwestern.edu

[1] Department of Communication Sciences and Disorders, Northwestern University, Evanston, IL 60208, USA

[2] Department of Electrical Engineering, Technical University of Denmark, Kongens Lyngby, Denmark

[3] Ikerbasque (Basque Science Foundation), Bilbao, Spain

[4] Department of Psychology, University of York, York, UK

Speech encompasses both a linguistic and an indexical dimension. The linguistic component conveys propositional information about objects, entities, and events in the world, whereas indexical information refers to acoustic correlates in the speech signal that provide information about the talker, including identity, age, gender, dialect, and emotional state (Pisoni, 1997; Vitevitch, 2003). These two components necessarily coexist and are integrally blended in a single auditory unit, such that is virtually impossible to perceptually segregate one from the other upon hearing an utterance. Indexical information is not the only nonpropositional dimension of a spoken word. In daily life, listeners often experience speech in the presence of environmental noise. Although there is ample evidence suggesting the integration of linguistic and indexical information in memory during speech processing, research examining whether co-occurring environmental sounds are also encoded in memory has only started to emerge. However, the available evidence indicates that, compared to indexical effects, speech-extrinsic specificity effects seem to be more fragile and their appearance, conditional on the experimental context in which they are probed. The aim of the present study was to understand the conditions in which sound-specificity effects occur by employing a close analogy to the voice-specificity effect in a context that (1) emulates the relationship between a word and a voice in its two crucial aspects: co-occurrence and integrality (Experiment 1) and

(2) only allows for co-occurrence, without integrality between the words and sounds (Experiment 2).

Indexical effects in spoken word recognition

Early models of spoken word recognition endorsed an abstractionist view of lexical representations in memory (e.g., Distributed Cohort Model: Gaskell & Marslen-Wilson, 1997, 1999, 2002; PARSYN: Luce, Goldinger, Auer, & Vitevitch, 2000; Shortlist: Norris, 1994; TRACE: McClelland & Elman, 1986; see Jusczyk & Luce, 2002, for a review), in which the underlying assumption is that the speech signal is mapped onto abstract linguistic representations. Accordingly, nonlinguistic information pertaining to the talker's voice (otherwise known as indexical information) is deemed irrelevant for spoken word recognition and is discarded early in the processing stages through a process typically referred to as normalization (Jusczyk & Luce, 2002; Lachs, McMichael, & Pisoni, 2003; Pisoni, 1997).

This approach was later challenged by an extensive body of studies that reported what are collectively referred to as *indexical effects,* emerging as a result of changing the talker's voice from exposure to test (e.g., Bradlow, Nygaard, & Pisoni, 1999; Church & Schacter, 1994; Creel, Aslin, & Tanenhaus, 2008; Goldinger, 1996, 1998; Luce & Lyons, 1998; Mattys & Liss, 2008; Mullennix, Pisoni, & Martin, 1989; Nygaard, Sommers, & Pisoni, 1994; Palmeri, Goldinger, & Pisoni, 1993; Schacter & Church, 1992; Sheffert, 1998a, b). The common finding is that words that are repeated in the same voice in both exposure/study and test phases of an experiment are recognized/identified/discriminated more accurately and/or faster than words repeated in a different voice. This indicates that listeners retain talker-specific acoustic details in memory, and that this information in turn facilitates the recognition of previously heard words as well as subsequent understanding of previously encountered speakers (e.g., Nygaard et al., 1994; see Luce & McLennan, 2005; Pisoni & Levi, 2007, for a review).

Relevant for the present study, a typical recognition memory paradigm used for probing indexical effects consists of an exposure and a test phase. The listeners are first exposed to the words during the exposure phase, where they perform a task regarding the words that is designed to promote their encoding in memory. Afterward, listeners complete a surprise recognition memory task that consists of deciding whether the word is *old* (repeated from exposure), or *new* (heard for the first time). The voice manipulation usually involves presenting half of the repeated words in the same voice as in exposure and the other half in the different voice. The voice-specificity effect is then assessed by comparing the overall recognition performance (accuracy and/or response latency) of the items repeated in the same voice to those repeated in the different voice. Better performance on the same-voice repetitions compared to the different-voice repetitions indicates the presence of a voice-specificity effect (e.g., Goldinger, 1996, 1998; Luce & Lyons, 1998; Mattys & Liss, 2008; Sheffert, 1998a).

The ample evidence supporting indexical effects in spoken word processing and encoding in memory lead to the emergence of episodic accounts of spoken word recognition. In this approach, variation in indexical dimension of the speech signal is considered crucial to explaining how listeners understand spoken words uttered at various speaking styles and rates by various speakers, each with their own vocal properties and idiolect. Accordingly, talker-related indexical information is encoded in memory and can affect subsequent word recognition (e.g., Elman, 2004, 2009; Goldinger, 1998). These models typically rely on multiple occurrences of a word (concept) that, in turn, forms clusters (networks), the size and strength of which is primarily determined by the frequency of the occurrences and their similarity to the shared word concept (e.g., Goldinger, 1998).

Speech-extrinsic specificity effects in spoken word processing

The first study to investigate the encoding of background sounds alongside spoken words in memory was carried out by Pufahl and Samuel (2014). The drive behind the study was the observation that since voices co-occur with words, the same questions that motivated indexical studies can be extended to background sounds that co-occur with spoken words. More specifically, the main question was whether changing a co-occurring sound would elicit a specificity effect in word identification similar to that elicited by changing a voice. Accordingly, words spoken by a male and a female talker were paired with one of two exemplars of environmental sounds (e.g., the word *butterfly* paired with a large barking dog (Exemplar A), or with a small barking dog (Exemplar B).[1] The nonlinguistic variation from exposure to test involved the talker's voice, the background sound, both of them, or none. Participants listened to the word–sound pairs in quiet during exposure and performed a semantic judgment task on the words, followed by a word-identification task at test, during which they heard the heavily filtered version of word–sound pairs. Results revealed the classical voice-specificity effect and, interestingly, a new specificity effect, elicited by the change of the accompanying environmental sound exemplar from exposure to test. Namely, the overall word-identification accuracy was reduced for the words repeated with the different

[1] Each word–sound association was unique and whenever a sound change from exposure to test occurred, it was within the same sound category. For example, if the word *butterfly* was paired with the large barking dog (Exemplar A) in exposure, at test it was paired with the small barking dog (Exemplar B) for the different-sound condition. The same-sound condition did not involve any change in the accompanying sound.

sound compared to those repeated with the same sound, as in exposure. This novel effect led the authors to propose that memory representations of spoken words may include both indexical (talker-related) and speech-extrinsic (sound-related) auditory information. However, inclusion of the associated sound in memory is only one possible explanation for the sound-specificity effect. Critically, it is not clear whether the sound-specificity effect is a result of the encoding of the sound in memory, or encoding of slightly different versions of the word resulting from the unique degradation generated by the sound. Thus, a drop in word-identification memory as a result of the change in the paired sound could be because the acoustic glimpse of a word formed in exposure does not match the one encountered at test. Further, the number of co-occurring sounds in the Pufahl and Samuel study was significantly greater than that of talker voices (two), because every word was paired with a unique sound exemplar. This discrepancy brings along the question as to whether the sound-specificity effect would emerge in the case of a more genuine analogy to its indexical counterpart in terms of the number of talkers and sounds, or whether it is more contingent on contextual details (e.g., the number of sounds). The sound-specificity effect was inspired in great part by its indexical counterpart, thus, it is important to understand the circumstances in which the two effects show a similar pattern of emergence, and the circumstances in which they may differ.

Speech-extrinsic specificity effects have also been found at a relatively early stage of processing, perceptual classification. Using a speeded classification paradigm (Garner, 1974), Cooper, Brouwer, and Bradlow (2015) investigated processing dependencies between background noise and indexical speech features (Experiment 1). Results revealed that background noise and indexical features were perceptually integrated, even when the two auditory streams were spectrally nonoverlapping. This suggests that speech and background noise are not entirely segregated at an early stage of perceptual processing. The authors also examined whether listeners encode the background noise co-occurring with spoken words in memory using a continuous recognition memory paradigm (Experiment 2). They found that recognition memory for spoken words dropped when the background noise changed between repetitions, but only when the noise and the speech signal were spectrally overlapping. Taken together, these findings favor an integrated processing of speech and background noise, modulated by the level of processing and the spectral overlap between speech and noise.

Finally, there is also evidence for speech-extrinsic auditory specificity during novel word learning. In their study, Creel, Aslin, and Tanenhaus (2012) taught English listeners to associate nonwords with unfamiliar shapes. During the learning phase, the words were heard in the clear or in white noise. Subsequent recognition was tested in either format via a forced-choice picture-selection task. Results revealed that listeners benefited from a match between learning and test contexts, such that those who were exposed to the same context at learning and test displayed the highest performance in terms of accuracy and speed. This finding was interpreted as indicating that listeners' newly formed lexical representations include auditory details pertaining to the speech-extrinsic context of the initial exposure.

In summary, the evidence on speech-extrinsic specificity effects highlights two major points: (1) Sounds/noise coexisting with spoken words may be perceptually integrated and/or retained in memory, similar to indexical features of speech, and (2) unlike indexical effects, sound-specificity effects are unstable and constrained. The second point might be related to fundamental differences between sounds and voices. Words and voices not only necessarily co-occur, they also are integral to each other. Following Vitevitch (2003) use of the term, *integrality* refers to the fact that words and the voice that utters them cannot be separated or exist without one another. In Gestalt terms, words and voices belong to a unique source and share a "common fate." In contrast, co-occurring sounds are not integral to spoken words; they exist independently and can often be segregated from them with relative ease. Therefore, the likelihood that the co-occurring element (voice or sound) is retained in memory in a format or another may be a function of the degree of perceived integrality with the spoken word. In the present study, we tested this hypothesis in two recognition memory experiments. Experiment 1 compared two conditions in which the integrality element in the stimuli was high. In a first condition (Experiment 1a), we aimed to replicate the classic voice-specificity effect, which also represents a case of "maximal integrality" between a word and a voice. In the second condition (Experiment 1b), we probed the sound-specificity effect by pairing the spoken words of Experiment 1a with either one of two environmental sounds. Crucially, the sounds were made as integral as possible to the words they were paired with through modulation along the word's intensity envelope. We predicted that if integrality between words and co-occurring sounds is an important factor in the emergence of a sound-specificity effect, then a comparable specificity effect should be expected in both conditions (voice and sound).

Experiment 2 sought to disentangle the contribution of integrality from that of mere co-occurrence in the appearance of the sound-specificity effect. Namely, while the sound-specificity effect in the high-integrality condition could be explained by the integrality element introduced in the word–sound pairs, it could also result from the mere co-occurrence of the words with two acoustically and semantically different sounds. To decouple these two possibilities, Experiment 2 was designed to be identical to Experiment 1, except that acoustic integrality was neutralized by removing any intensity modulation. If integrality plays a crucial role in the emergence of a sound-specificity effect, any sound-specificity effect emerging

in Experiment 1 should be attenuated, or even disappear in Experiment 2. Alternatively, if integrality is not a key factor and mere co-occurrence between the words and sounds is sufficient to elicit an effect, then such an effect should persist in Experiment 2.

Experiment 1

Experiment 1 examined specificity effects in recognition memory for spoken words in two contexts of high integrality regarding the component co-occurring with the linguistic dimension: voice and sound. A recognition memory paradigm similar to the ones in Luce and Lyons (1998, Experiment 2) and Mattys and Liss (2008) was used in both Experiments 1a and 1b, consisting of an exposure phase, a short delay and a memory test phase.

Experiment 1a probed the classic voice-specificity effect, which for the present purposes represents the case of "maximal integrality." Recognition memory for the words was assessed as a function of the change in the talker's voice from exposure to test. Namely, recognition accuracy and response latencies for the words repeated in the same voice were compared to those for the words repeated in the different voice. In the case of a voice-specificity effect, the recognition performance for same-voice word repetitions should be higher than the performance for different-voice word repetitions.

Experiment 1b investigated the sound-specificity effect in a high-integrality context, wherein the sounds were made to be integral to the words in a similar way that voices are integral to words. The concept of integrality endorsed in this study refers to a degree of acoustical integration between the word and sound, aimed at making their segregation challenging and promoting their perceptual blending. More specifically, we wanted the sounds to be paired with the words in such a way that every association would be acoustically and perceptually blended into one unique item, similar to a uniquely produced spoken word.

In addition, we wanted the sounds to retain their unique identity across the different pairings, like a voice preserves its identity across different utterances. With these requirements in mind, we implemented the integrality element by modulating the sounds according to the intensity envelope of each individual word. It is well established in the literature that speech intelligibility strongly depends on the intensity fluctuations over time. For instance, noise-vocoded speech is perfectly intelligible given that enough subband envelopes are used (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). Shannon et al. (1995) demonstrated that using only the speech envelopes and replacing the fine structure with noise yields perfect speech intelligibility, provided that at least three subbands are used. Further, several prominent speech-intelligibility prediction models use only modulation information (e.g., Jørgensen & Dau, 2011; Jørgensen, Ewert, & Dau, 2013). Therefore, we chose the intensity envelope of the word as the link between the words and the sounds that would create their perceptual integration. To preserve the identity of the sounds, we selected sounds whose identity is mainly conveyed by their temporal fine structure, rather than their intensity (amplitude) modulation. This quality makes them suitable candidates for amplitude modulation by another signal, in this case, the spoken word. Hence, the integral versions of the sounds were created by preserving the fine structure of the sounds and replacing their intensity envelopes with those of the words. This modulation method produces sound candidates that are uniquely tailored for each individual word by following the rhythm of the word, while also retaining their own identity as speech extrinsic sounds. As in Experiment 1a, recognition memory for the words was assessed as a function of the change in the accompanying sound from exposure to test. In case of the emergence of a sound-specificity effect, recognition memory for words repeated with the same paired sound as in exposure should be higher than that for words repeated with the different paired sound.

Experiment 1a

Method

Participants

Forty-nine students at the University of York (age range: 18–27 years) participated in exchange for either course credit or payment. All participants provided written consent prior to the experiment. They all identified themselves as native speakers of English, and none of them reported a history of hearing or speech and language related problems.

Recording

The words were recorded in a sound-attenuated booth by a male and a female talker, who spoke Standard British English. The talkers were instructed to read at a normal pace and neutral intonation in front of a microphone (SHURE SM58). The words were digitized at a 44.1-kHz sampling rate using a recording software program (Cool Edit Pro, 2000) and stored in separate audio files. All stimuli were filtered to eliminate background noise, and 100 milliseconds of silence was appended to the beginning and end of the words to avoid transition artefacts. In addition, all the sound files were normalized so that their average intensity was 68 dB using the Praat software (Boersma & Weenink, 2013).

Materials and design

The stimuli consisted of 80 disyllabic, initial-stress words, half of which represented animate (living) entities and half inanimate (nonliving) entities. All the words were of relatively high frequency, as reported in the CELEX database, with the following mean log frequency values per semantic category: $(M, SD)_{\text{animate}} = (1.22, 0.6)$; $(M, SD)_{\text{inanimate}} = (1.38, 0.45)$. These mean frequencies were not different from each other: $F(1, 72.34) = 1.67$, $p > .05$. Acoustic analyses performed on the stimuli produced by the two talkers revealed that the mean difference in fundamental frequencies (F0s) between the male and female talkers was 40.5 Hz ($M_{\text{maleF0}} = 115.55$ Hz, $M_{\text{femaleF0}} = 156.03$ Hz). The list of words is provided in Appendix A.

The experiment involved two phases, exposure and test, and a short delay in between. In each phase, participants heard a block of 60 words, each spoken one at a time. None of the words were repeated within a block. Half the stimuli in each block were produced by the female voice and the other half by the male voice. The 60 words in the exposure phase (Block 1) were the same for all participants, although the voice in which they were heard was counterbalanced across participants. In the test phase (Block 2), 40 out of the 60 words that were already heard in the exposure phase were repeated (the "old" critical trials), half in the same voice as in Block 1, half in the other voice. Which words in the test phase were in the same or the different voice was counterbalanced across participants. The counterbalancing in terms of both talker (male or female) and talker sameness (same or different from exposure to test) resulted in four stimuli lists (counterbalancing groups) in total, and every participant was randomly assigned to either one of them. The remaining 20 words in Block 2 had not been heard in the exposure phase (Block 1). Hence, these were the same for all participants, with half of them spoken in the male and half in the female voice.

Procedure

Exposure phase The experiment was run on the DMDX software (Forster & Forster, 2003). Participants sat individually in a sound-attenuated booth and listened to the trials played binaurally over headphones (Sony MDR-V700) at a comfortable listening level. They were instructed to make an "animate/inanimate" decision about the word in each trial and ignore the voice change across the trials because the talker's voice was not relevant for their task. The animate and inanimate concepts were defined, and examples for each of the categories were provided (e.g., "banana is inanimate"; "professor is animate"). Participants were encouraged to be as accurate as possible and to press the response key within the allowed time frame of 10 seconds. First, the trial was played, and after 500 milliseconds, a message was displayed on the screen, prompting the participant to respond by pressing either one of the corresponding "shift" keys on the computer keyboard: the right shift key if the word was animate, and the left shift key if the word was inanimate.[2] Participants were told to wait for the message to appear on the screen before responding. The next trial followed immediately after the participant hit a response button, or after 10 seconds if no response was provided. The order of trials was randomized for each participant. No feedback was provided after each trial, and there was no mention of an upcoming recognition task.

Delay After completing the first part, participants spent 5 minutes playing an easy online game that did not involve any auditory exposure (Cube Crash 2). This was done to allow for a moderate delay before assessing their recognition memory in the test phase. All participants played the same game.

Test phase In order to assess the effect of voice change on word-recognition memory, participants completed a surprise word-recognition task. The experimenter explained that some of the words would be repeated from the first part of the experiment (i.e., old), and the other words would be presented for the first time (i.e., new). Participants were instructed to decide whether the word was old or new and, again, ignore the voice change across the trials. They were encouraged to be as accurate as possible, but to also press the response key as soon as they made their decision. Participants first saw an *x* symbol appear in the center of the screen. After 500 milliseconds, they heard the word and responded by pressing one of the shift keys on the computer keyboard (right for "old" and left for "new"). The next trial followed immediately after the participant's response, or after 10 seconds if no response was provided. The order of trials was randomized for each participant.

Results

Participants' mean accuracies in the semantic judgment task of the exposure phase were assessed to determine whether they were eligible for further analysis.[3] A correct response was coded as "1" and an incorrect one as "0." Mean accuracies were then calculated by averaging overall responses. Only the participants who displayed overall accuracies above 90% correct were included in the final analysis because it meant that they had successfully encoded the words during exposure. One participant failed to meet this criterion and was therefore excluded from further analysis. The rest of the participants displayed high mean accuracies, $(M, SD)_{\text{animate}} = (98.96, 2.19)$, $(M, SD)_{\text{inanimate}} = (99.24, 1.85)$.

[2] The message on the screen consisted of the word ANIMATE on the right side (referring them to the right "shift" key) and the word INANIMATE on the left side (referring them to the left "shift" key).

[3] The mean accuracy value represented the percentage of correct responses.

Recognition memory performance was assessed in terms of accuracy and response time. Only the critical (old) trials were included in the analysis, and response latencies were measured from the onset of the stimuli. The latencies of correct responses were submitted for analysis, and latencies longer than 2 standard deviations above the mean on a subject-by-subject basis were omitted. The data were analyzed using mixed-effects regression models (Baayen, Davidson, & Bates, 2008), with recognition accuracy (accuracy) and response time (RT) as dependent variables. The models were implemented in R (Version 3.3.1) using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015). Accuracy was coded as a binary variable, with values 1 and 0 per trial, representing a correct and an incorrect response, respectively. Linear mixed-effects regression models (LMEM) were used for the analysis of the continuous RT variable and generalized mixed-effects regression models (GLMEM) with a logistic function were used for the binary variable, accuracy.

There were three fixed factors, coded as binary variables: (1) voice sameness (1: same, −1: different voice), (2) semantics (1: animate, −1: inanimate word), and (3) exposure voice (1: female, −1: male voice). Prior to adding any fixed factors to the model, the maximal random-effects structure was tested against the basic structure for each dependent variable, to assess whether adding random slopes for the fixed factors would be necessary (see Barr, Levy, Scheepers, & Tilly, 2013). In line with Barr et al.'s (2013) argument that linear mixed-effects models generalize best when they include the maximal structure justified by the design, the maximal random structure was used whenever it converged.[4] In the instances when it did not converge, model comparisons using log-likelihood ratio tests determined whether simpler models would fit the data just as well. Henceforth, in all the present analyses, unless noted otherwise, the best fitting model with the largest random-effects structure that converged will be reported.

For every dependent variable, the fixed factors, as well as their interactions, were added incrementally to the base model, and improved fit to the model was assessed using the likelihood ratio test. The base model included only the random terms. The main effects of voice sameness, semantics, and exposure voice were obtained by testing the improvement in the model fit when each one of these factors was individually added to the base model.

Voice sameness

Assessing the main effect of voice sameness on recognition accuracy (accuracy) revealed the anticipated voice-specificity effect, $\beta = .19$, *SE* $\beta = .06$, $\chi^2(1) = 9.26$, $p = .002$. Participants were overall more accurate in recognizing previously heard (old) words that were repeated in the same-talker voice compared to the words repeated in the different voice. The voice-specificity effect did not manifest in participants' overall response time (RT), $\beta = -6.31$, *SE* $\beta = 5.48$, $\chi^2(1) = 1.32$, $p = .25$. Thus, listeners did not recognize the words repeated in the same voice faster than the words repeated in the different voice. The mean values of each dependent variable in the two voice conditions are displayed in Table 1.

Semantics

There was a main effect of semantics on recognition accuracy, $\beta = .29$, *SE* $\beta = .10$, $\chi^2(1) = 8.29$, $p = .004$, indicating that overall participants were better at recognizing animate old words compared to inanimate words. However, importantly for the present analysis, the voice-specificity effect was not affected by the semantic category of the words, as revealed by the lack of interaction between semantics and voice sameness, $\beta = .03$, *SE* $\beta = .06$, $\chi^2(1) = .18$, $p = .68$.

A main effect of semantics was also present on response time, $\beta = -23.91$, *SE* $\beta = 7.80$, $\chi^2(1) = 8.27$, $p = .004$, suggesting that participants were faster at recognizing animate old words compared to inanimate ones. However, there was no interaction between semantics and voice sameness, $\beta = 7.87$, *SE* $\beta = 5.90$, $\chi^2(1) = 1.77$, $p = .18$. Table 2 displays the mean values of each dependent variable in each semantic category.

Exposure voice

There was no main effect of the exposure voice on recognition accuracy, $\beta = -0.03$, *SE* $\beta = .06$, $\chi^2(1) = .19$, $p = .66$, meaning that the voice of the speaker in the exposure phase did not matter for listeners' accuracy performance in the test phase. Additionally, no interaction between exposure voice and voice sameness was found, $\beta = -0.12$, *SE* $\beta = .06$, $\chi^2(1) = 3.53$, $p = .06$. Similarly, there was no main effect of the exposure voice on response latency, $\beta = -8.48$, *SE* $\beta = 6.30$, $\chi^2(1) = 1.77$, $p = .18$, as well as no interaction of exposure voice and voice sameness, $\beta = .62$, *SE* $\beta = 6.24$, $\chi^2(1) = .01$, $p = .92$. Thus, the voice of the speaker in the exposure phase did not matter for participants' response speed in the test phase.

Table 1 Mean accuracy (percentage correct) and RT (ms) in each voice condition (standard deviations are shown in parentheses)

	Same voice	Different voice
Accuracy (%)	81.77 (12.01)	76.15 (12.51)
RT (ms)	1185 (148)	1201 (172)

[4] Barr et al. (2013) also noted that for categorical variables like the accuracy variable in the present analysis, it may be more difficult for the corresponding maximal generalized mixed-effects models (GLMEM) to converge, especially when mixed logit functions are involved.

Table 2 Mean values for accuracy (percentage correct) and RT (ms) in each semantic category (standard deviations are shown in parentheses)

	Animate	Inanimate
Accuracy (%)	83.33 (9.53)	74.58 (15.05)
RT (ms)	1,169 (164)	1,220 (162)

Discussion

Experiment 1a replicated the classical voice-specificity effect using a recognition memory paradigm that involved an explicit memory test for previously heard words (e.g., Goldinger, 1998; Luce & Lyons, 1998; Mattys & Liss, 2008). As predicted, we found that participants were more accurate in recognizing previously heard words when they were repeated in the same voice, compared to when the voice was different. The effect was not reflected in the overall response time. This pattern of results is in line with other studies that have examine voice-specificity effects with a similar recognition memory paradigm. For example, Mattys and Liss (2008) reported similar findings in their study of voice-specificity effects with normal and dysarthric speech. Namely, in the normal speech condition they found a voice effect only for recognition accuracy, not response latency.

Interestingly, we also observed an effect of the semantic category of the words, reflected in both recognition accuracy and response time. Namely, animate words were recognized more accurately and faster than inanimate words. Although this effect was not of primary interest to the present study, it is an interesting one to observe. A similar effect has also been reported by several other studies and is typically referred to as the *animacy effect* (e.g., Bonin, Gelin, & Bugaiska, 2014; Nairne, VanArsdall, Pandeirada, Cogdill, & LeBreton, 2013; VanArsdall, Nairne, Pandeirada, & Blunt, 2013). The common finding is that animate words are recalled and/or recognized better and faster than inanimate words (see Bonin et al., 2014, for a review). However, these studies involved written words, that is, words presented on a computer screen. The present study extends previous ones by finding an animacy effect in recognition memory for spoken words. Further, the lack of an interaction between this effect and the primary effect of the above experiment, the voice-specificity effect, indicates that the animacy effect does not seem to be affected by the change of the talker voice.

Experiment 1a joins several other indexical studies that showed the voice-specificity effect using a similar recognition memory paradigm (e.g., Goldinger, 1996, 1998; Luce & Lyons, 1998; Mattys & Liss, 2008; Sheffert, 1998a). Critical for the present argument, this experiment represents the condition of maximal integrality between the linguistic (word) and nonlinguistic (voice) dimensions of the speech signal. As such, it provides a solid baseline for investigating another high-integrality, speech-extrinsic dimension, namely, a co-occurring sound.

Experiment 1b

Method

Participants

Fifty-four undergraduate students at the University of York (age range: 18–27 years) participated in exchange for either course credit or payment. All participants provided written consent prior to the experiment. They all identified themselves as native speakers of British English, and none of them reported a history of hearing or speech and language related problems.

Materials and design

The stimuli consisted of 80 word–sound pairs, involving the same set of words as in Experiment 1a. In parallel to the two voices in Experiment 1a, two environmental sounds were used. The integrality between the words and sounds was implemented by modulating the sounds along the intensity envelope of each individual word. Due to the nature of the modulation, the sounds had to fit the following criteria: (1) have a continuous structure that does not fluctuate over time, and (2) their identity should be conveyed mainly by their pitch and timbre information, not by their overall intensity envelope. A cat sound and a violin sound (playing one sustained tone) were selected as the best candidates. Acoustic analyses performed on the sounds revealed that the mean difference in fundamental frequencies (F0s) between them was 203 Hz (cat_{F0} = 552 Hz, $violin_{F0}$ = 349 Hz). The temporal waveform of the sounds, their spectrograms and the pitch contours (fundamental frequency over time) are depicted in Appendix B. Prior to being paired with the words, the "integral" versions of the sounds were created by modulating their intensity envelopes according to the intensity envelope of each individual word. The intensity envelopes were extracted by filtering the words to the frequency band between 0.3 and 6 kHz, extracting their Hilbert envelopes, and low-pass filtering the envelopes with a third-order low-pass filter at a cut-off frequency of 30 Hz. To generate the cat and violin integral sounds for a given word, the sounds were limited to the same frequency band (0.3–6 kHz) and then either lengthened by adding silence at the end or shortened by cropping the end to match the duration of the speech token and its intensity envelope. The sounds were then multiplied by the intensity envelope, such that they followed the intensity envelope of the word, which defines the "rhythm" of the token. Therefore, although the same two environmental sounds were involved, the modulation process led to unique

exemplars being created for every word because the intensity envelope of the integral versions of the sounds followed the intensity envelope of the individual words they were later mixed with. However, the integral maskers did not contain any intelligible/identifiable speech information, but rather sounded like amplitude-modulated versions of the original sounds (with the type of amplitude modulation determined by the word's intensity envelope). Each word was then mixed with the corresponding integral version of the sounds at a signal-to-noise ratio (SNR) that preserved the maximal intelligibility of the word. For the majority of the words this SNR was −3 dB. However, other SNR values (−1, 0, +1 and +3 dB) were also used in some instances, to ensure the word's maximal intelligibility. The SNR values were piloted prior to the experiment, and the ones that yielded the maximum word identification accuracy (100% correct) were selected. Examples of the processing scheme for the two integral sounds and the final, mixed version of the stimuli are displayed in Appendix B (see Figs. 2 and 3). All the stimuli files were generated with a sampling rate of 44.1 kHz and a resolution of 16 bits. Every stage of the stimuli preparation process was implemented using the MATLAB software (Version R2014b, MathWorks, Natick, MA).

The experimental design was the same as in Experiment 1a. In each phase, participants heard a block of 60 trials, this time spoken only by the female talker, and each played one at a time. None of the trials were repeated within a block. Half the words in each block were paired with their corresponding integral versions of the cat sound and the other half with the integral versions of the violin sound. While the words in the exposure trials (Block 1) were the same for all participants, what sound they were paired with was counterbalanced across participants. In the test trials (Block 2), 40 of the 60 words were repeated from the exposure phase (the "old" critical trials). Half of the repeated words were paired with the same sound as in exposure, and the other half with the different sound. Which words in the test phase were paired with the same or the different sound was counterbalanced across participants. Counterbalancing sound (cat or violin) and sound sameness (same or different from exposure to test) resulted in four stimulus lists (counterbalancing groups) in total, and every participant was randomly assigned to one of them. The words in the remaining 20 trials in Block 2 had not been heard in the exposure phase (Block 1). Hence, these were the same for all participants, with half of them paired with the cat sound and the other half with the violin sound.

Procedure

The procedure was the same as in Experiment 1a, with slightly different instructions. This time, the participants were informed that they would hear words paired with background sounds and that they had to make decisions regarding the word only (i.e., animate/inanimate in exposure, and old/new in the test phase), while ignoring the sound. Prior to the experimental trials, the participants completed four practice trials that involved different words, spoken by a different (male) talker.

Results

Six participants were excluded from analysis for the following reasons: (1) technical failure of the experimental software (three); (2) judging the sounds, instead of the words, in the exposure phase (two); and (3) judging all the "inanimate" words in the exposure phase incorrectly (one). Overall, forty-eight participants were included in the analysis.

All participants displayed high mean accuracies in the semantic judgment task of the exposure phase, indicating that they had successfully encoded the words during the task, $(M, SD)_{\text{animate}} = (98.26, 2.48)$, $(M, SD)_{\text{inanimate}} = (98.75, 2.44)$.

Recognition memory performance was assessed in terms of accuracy and response time (RT), with the data analyzed in the same way as in Experiment 1a. Accordingly, there were three fixed factors, coded as binary variables: (1) sound sameness (1: same, −1: different sound), (2) semantics (1: animate, −1: inanimate word), and (3) exposure zound (1: violin, −1: cat sound). For sound sameness, random slopes for both subjects and items were included in the random structure of the maximal model, whereas for the other two factors, only random slopes for subjects were added. The main effects of sound sameness, semantics, and exposure sound were obtained by testing the improvement in the model fit when each one of these factors was individually added to the base model.

Sound sameness

As anticipated, there was a main effect of sound sameness on recognition accuracy, revealing the presence of a sound-specificity effect, $\beta = .14$, $SE\ \beta = .06$, $\chi^2(1) = 5.95$, $p = .01$. The sound-specificity effect was also present in the listeners' response time, $\beta = -19.62$, $SE\ \beta = 8.25$, $\chi^2(1) = 5.42$, $p = .02$. Thus, listeners were both more accurate and faster in recognizing previously heard words that were repeated with the same integral sound as in exposure, compared to words that were repeated with the different integral sound. Table 3 displays the mean accuracy and response time values in each condition.

Table 3 Mean values for accuracy (percentage correct) and RT ms) in each integral sound condition (standard deviations are shown in parentheses)

	Same integral sound	Different integral sound
Accuracy (%)	80.42 (12.54)	76.04 (9.56)
RT (ms)	1,425 (230)	1,471 (264)

Semantics

Similar to Experiment 1a, a main effect of the word's semantic category (semantics) was observed on both accuracy, β = .28, *SE* β = .1, $\chi^2(1) = 7.39$, $p = .007$; and RT, β = −39.38, *SE* β = 13.82, $\chi^2(1) = 7.45$, $p = .006$. Listeners were better and faster at recognizing animate words compared to inanimate words. However, the sound-specificity effect was not affected by the semantic category of the words, as shown by the absence of an interaction between the two factors on both accuracy, β = .09, *SE* β = .06, $\chi^2(1) = 2.2$, $p = .14$; and RT, β = 6.43, *SE* β = .8.28, $\chi^2(1) = .6$, $p = .44$. The mean values for both variables are shown in Table 4.

Exposure sound

There was no main effect of the exposure sound on either accuracy, β = −0.02, *SE* β = .06, $\chi^2(1) = .13$, $p = .71$; or RT, β = .03, *SE* β = 8.61, $\chi^2(1) = 0$, $p = 1$. Further, there was no interaction between the sound-specificity effect and the exposure sound on either accuracy, β = 0.04, *SE* β = .06, $\chi^2(1) = .57$, $p = .45$; or RT, β = 4.05, *SE* β = 8.47, $\chi^2(1) = .23$, $p = .63$. Therefore, the sound with which the words were heard during exposure did not affect either the recognition memory performance of participants at test, or the sound-specificity effect.

Comparison between the voice and the sound-specificity effects

In order to assess how similar the two specificity effects were to one another, a comparative statistical analysis was performed. The data from Experiments 1a and 1b were aggregated and analyzed using linear mixed-effects regression models. An extra fixed factor, experiment, was added to the analysis, coded as: 1 (Exp. 1a) and 2 (Exp. 1b). The main fixed factor of interest, voice/sound sameness was named *sameness* and it was coded in the same way as in the previous analyses: 1 (same), −1 (different).The crucial aspect of this comparative analysis was the interaction between sameness (specificity effect) and experiment.

Accuracy As expected, there was a robust main effect of sameness (specificity effect) on recognition accuracy, β = .16, *SE* β = .04, $\chi^2(1) = 15.51$, $p < .0001$. No main effect of experiment was found, β = −0.06, *SE* β = .14, $\chi^2(1) = .18$, $p = .67$. Importantly, there was no interaction between sameness and experiment, β = −0.04, *SE* β = .08, $\chi^2(1) = .26$, $p = .61$, indicating that the voice and sound-specificity effects were comparable.

Response time A main effect of sameness was also found on listeners' response time, β = −13.09, *SE* β = 4.97, $\chi^2(1) = 6.47$, $p = .01$. Further, there was a main effect of experiment, β = 254.67, *SE* β = 41.99, $\chi^2(1) = 31.27$, $p < .0001$. However, there was no interaction between the specificity effect and experiment, β = −12.61, *SE* β = 9.86, $\chi^2(1) = 1.64$, $p = .20$, suggesting that the specificity effect on response latency persists between experiments, but is not strong enough to elicit an interaction.

Table 4 Mean values for accuracy (percentage correct) and RT (ms) in each semantic category (standard deviations are shown in parentheses)

	Animate	Inanimate
Accuracy (%)	82.60 (8.93)	73.85 (14.85)
RT (ms)	1,415 (250)	1,484 (258)

Discussion

Experiment 1b investigated the role of a novel dimension in the coexistence of words and sounds in the emergence of the sound-specificity effect. This was motivated by the observation that words and voices not only necessarily co-occur but are also integral to one another in such a way that makes their segregation virtually impossible. We were interested to see whether inducing a similar degree of integrality between words and their accompanying sounds would create a sound-specificity effect. Therefore, in parallel with Experiment 1a, Experiment 1b represented a case of high integrality context, where the sounds were made integral to each individual word by modulation along the word's intensity envelope. That is, each sound's intensity envelope was replaced by the paired word's intensity envelope, while its fine spectral structure was kept intact.

The analysis revealed the expected sound-specificity effect in recognition accuracy and, interestingly, in their response time. Listeners were both more accurate and faster in recognizing words that were repeated with the same integral sound as in exposure, compared to words repeated with the different integral sound. Further, similar to Experiment 1a, we found an animacy effect, such that animate words were recognized better and faster than inanimate words.

The main finding in Experiment 1b highlights the role of integrality between words and sounds in the appearance of the sound-specificity effect. A question, however, is whether integrality is necessary to elicit this effect. The observed effect could be the result of the integrality element we introduced in the stimuli, but it could also have emerged from the mere co-occurrence of the words with two acoustically and semantically distinct sounds. Specifically, although the sounds were made integral to the words, they retained their identity across the different pairings, and a cat sound is clearly different from a violin sound, both acoustically and semantically. Would a sound-specificity effect emerge if the words and sounds merely co-occurred, without being integral to each other? Experiment 2 was designed to address this question.

Experiment 2

Experiment 2 was identical to Experiment 1b, except that the intensity modulation used to induce integrality between the words and sounds was removed from the stimuli. If the sound-specificity effect found in Experiment 1b represents a mere co-occurrence context effect, then it should persist in Experiment 2 as well. However, if integrality between words and sounds is the crucial factor behind the appearance of the sound-specificity effect, then removing integrality should make the sound specificity disappear.

Method

Participants

Forty-six students at the University of York (age range: 18–23 years) participated in exchange for either course credit or payment. All participants provided written consent prior to the experiment. They all identified themselves as native speakers of British English, and none of them reported a history of hearing or speech and language related problems.

Materials and design

The stimuli consisted of the same set of words and the two sounds (cat and violin) as in Experiment 1b, but without the sounds being modulated by the intensity envelope of each word. In order to ensure a fair comparison across experiments in terms of the spectral content, the sounds were filtered to the same band (0.3–6 kHz) as in their integral version. As in Experiment 1b, the words were mixed with the sounds at a signal-to-noise ratio (SNR) that preserved the maximal intelligibility of the word. For the majority of the words, this SNR was −3 dB. However, other SNR values (−1, 0, +1, and +3 dB) were used in some cases to ensure the word's maximal intelligibility. The SNR values were piloted prior to the experiment and the ones that yielded the maximum word identification accuracy (100% correct) were selected. The experimental design was identical to that of Experiment 1b.

Procedure

The procedure was identical to the one used in Experiment 1b.

Results

All participants displayed high mean accuracies in the semantic judgment task of the exposure phase, indicating that they had successfully encoded the words during the task ($M, SD)_{animate}$ = (99.20, 2.01), ($M, SD)_{inanimate}$ = (99.93, 0.49).

Recognition memory performance was assessed in terms of accuracy and response time (RT), with the data analyzed in the same way as in Experiment 1b.

Sound sameness

Unlike in Experiment 1b, no main effect of sound sameness was found on accuracy, $\beta = .007$, $SE\ \beta = .07$, $\chi^2(1) = .009$, $p = .92$; or RT, $\beta = -10.75$, $SE\ \beta = 10.84$, $\chi^2(1) = .98$, $p = .32$. The mean values of accuracy and RT in each condition are displayed in Table 5.

Semantics

Similar to the previous experiments, there was a main effect of semantic category (semantics) on recognition accuracy, $\beta = .27$, $SE\ \beta = .12$, $\chi2(1) = 4.99$, $p = .02$, but not on response times, $\beta = -24.15$, $SE\ \beta = 12.99$, $\chi^2(1) = 3.29$, $p = .07$. Thus, participants recognized animate words better, but not faster than inanimate words. No interaction between semantics and sound sameness was found on either accuracy, $\beta = -0.03$, $SE\ \beta = .06$, $\chi^2(1) = .19$, $p = .67$; or RT, $\beta = 5.09$, $SE\ \beta = 10.87$, $\chi^2(1) = .22$, $p = .64$. Table 6 shows the mean accuracy and response time in each condition.

Exposure sound

There was no main effect of the exposure sound on either accuracy, $\beta = .03$, $SE\ \beta = .06$, $\chi^2(1) = .18$, $p = .67$; or RT, $\beta = -6.75$, $SE\ \beta = 10.41$, $\chi^2(1) = 0.42$, $p = .52$. Further, there was no interaction between the sound-specificity effect and the exposure sound for either accuracy, $\beta = -0.04$, $SE\ \beta = .06$, $\chi^2(1) = .37$, $p = .54$; or RT, $\beta = -11.24$, $SE\ \beta = 10.25$, $\chi^2(1) = 1.2$, $p = .27$. Therefore, the sound with which the words were heard during exposure had no effect on listeners' recognition memory performance at test, as well as on the emergence of a sound-specificity effect. Specificity effects across experiments are graphically depicted in Fig. 1.

Discussion

Experiment 2 examined the emergence of the sound-specificity effect in the presence of the same two background sounds used in Experiment 1b, but with the integrality component removed from the stimuli. The aim was to decouple

Table 5 Mean values for accuracy (percentage correct) and RT (ms) in each sound condition (standard deviations are shown in parentheses)

	Same sound	Different sound
Accuracy (%)	77.07 (13.65)	76.74 (10.39)
RT (ms)	1,507 (256)	1,529 (258)

Table 6 Mean values for accuracy (percentage correct) and RT (ms) in each semantic category (standard deviations are shown in parentheses)

	Animate	Inanimate
Accuracy (%)	81.09 (12.15)	72.72 (14.63)
RT (ms)	1,491 (250)	1,549 (259)

two alternative explanations for the appearance of the sound-specificity effect in Experiment 1b: integrality versus mere co-occurrence. Consistent with the integrality account, there was no sound-specificity effect on either recognition accuracy or response time in the absence of integrality between the words and the co-occurring sounds.

Additionally, in line with the previous experiments, an animacy effect was found, this time only in listeners' recognition accuracy. It is intriguing to observe this effect consistently across our experiments, which involve different contexts. This suggests that besides extending to spoken words, the animacy effect seems unaffected by changes in their context. That is, it emerges regardless of whether the words are spoken alone (Experiment 1a), with an accompanying integral sound (Experiment 1b), or with an accompanying nonintegral sound (Experiment 2).

The main finding of Experiment 2 consolidates the crucial role of integrality in the appearance of the sound-specificity effect. However, it is important to check whether this effect could be partly accounted for by masking differences, or acoustic glimpses, between the Experiments 1b and 2. Acoustic glimpse refers to the intelligible leftovers of a word after the portion affected by the masking sound has been accounted for. Two different sounds lead to two different acoustical glimpses of the same word. Therefore, it could be that the sound-specificity effect in Experiment 1b was elicited by the contrast between the different acoustic glimpses of the same word in exposure and test, rather than by the different associations in exposure and test of the same word with the two sounds. To disentangle these possibilities, the acoustic glimpses of the critical (old) words resulting from the two sounds were measured quantitatively in both experiments by means of Cooke's (2003, 2006) glimpse analysis and compared across experiments.

Comparative analysis of acoustic glimpses

The proportion of glimpsed information was calculated for each using the glimpse detection model (Cooke, 2006). The model is based on the use of glimpses of speech in spectro-

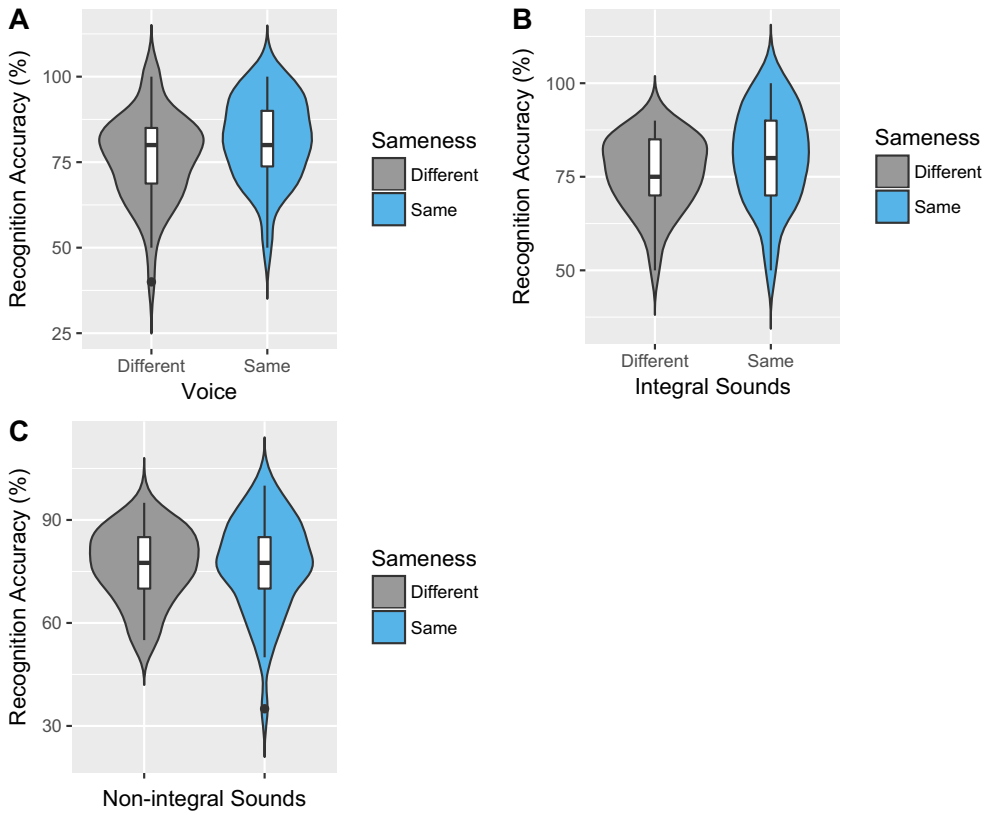

Fig. 1 Specificity effects across experiments, illustrated in terms of recognition accuracy in the two voice/sound conditions: same vs. different. The violin graphs in **a** display the data from Experiment 1a (voice), the graphs in **b** show the data from Experiment 1b (integral sounds), and the graphs in **c** represent the data from Experiment 2 (nonintegral sounds)

temporal regions where it is least affected by the background masking. It uses simulated spectro-temporal excitation patterns as input, which are smoothed and compressed representations of the envelope of the basilar membrane response to sound and are typically considered effective first-order representations of auditory stimuli at an early stage of processing (STEP; Moore, 2003). Based on the assumption that listeners may be unable to detect very brief regions of speech target dominance, or regions that occupy a very narrow portion of the spectrum, the glimpse detection model includes a *minimum glimpse area* criterion. Namely, all connected regions of spectro-temporal elements that satisfy a given local signal-to-noise (SNR) criterion also have to possess an "area" (i.e., glimpse extent) greater than a specified amount. In this context, "area" is defined as the number of time-frequency elements making up the glimpsed region.

For the present glimpse calculations, the spectro-temporal excitation pattern used as input to the model was processed by a bank of 55 gamma-tone filters (Patterson et al., 1988), between 100 and 8000 Hz. The SNR criterion was 3 dB, which meant that speech had to exceed the masker by 3 dB to be counted as a glimpse. The calculation of glimpses was based on 5-ms frames. The glimpse percentage produced by the computational analysis for a particular stimulus corresponds to the average percentage of all the individual glimpses in the input that meet the criteria mentioned above. For every word in our experiments, there were two acoustic glimpses (hence, glimpse percentages), one resulting from each of the two masking sounds.

Glimpse percentages across all the "old" words were compared for the two masking sounds in Experiments 1b and 2. First, the comparison of the glimpse differences between the experiments is provided below, followed by the analyses of the glimpse differences in each experiment. Given that these analyses involved only the stimuli, and the focus was to compare the mean glimpse values, ANOVA tests were implemented (IBM SPSS for Windows, Version 21.0, Armonk, NY), instead of linear mixed-effects regression analyses.

Comparison of the glimpse contrasts between experiments

The glimpse contrasts in both experiments were compared via a two-way repeated-measures ANOVA, with sound (two levels: cat vs. violin) and experiment (two levels) as factors. As anticipated, there was a main effect of sound, $F(1, 39) = 704.56$, $p < .0001$, $\eta^2 = .95$, as well as a main effect of experiment, $F(1, 39) = 86.20$, $p < .0001$, $\eta^2 = .69$. Crucially, there was an interaction between sound and experiment, $F(1, 39) = 71.50$, $p < 0001$, $\eta^2 = .65$, showing that the glimpse difference in Experiment 2 (diff.: 44.07% − 22.40% = 21.67%) was significantly greater than the glimpse difference in Experiment 1b (diff.: 46.10% − 34.09% = 12.01%).

Experiment 1B

A repeated-measures ANOVA, with glimpse (glimpse percentage) as the dependent variable and sound as the within-items factor revealed a significant difference between the mean acoustic glimpse of the same word(s) resulting from the two sounds, $F(1, 39) = 483.33$, $p < .0001$, $\eta^2 = .93$.

Experiment 2

The same analysis as above was performed, revealing similar results, $F(1, 39) = 405.10$, $p < .0001$, $\eta^2 = .91$. Namely, there was a significant difference between the glimpses of the same word(s) resulting from the two masking sounds. The mean glimpse percentages for each sound and experiment are displayed in Table 7.

This analysis undermines the possibility that a contrast in the acoustic glimpses resulting from masking, rather than the integrality between the words and sounds, could explain the presence, or lack thereof, of the sound-specificity effect. If the glimpse contrast played a role, we should have observed the opposite pattern of results between the two experiments, or at least, a sound-specificity effect in Experiment 2 as well, given that the glimpse contrast in that experiment was significantly higher than that in Experiment 1b. This leaves us with the integrality between words and sounds as the crucial factor behind the observed sound-specificity effect.

General discussion

This study investigated the corepresentation of spoken words and environmental sounds in memory in an analogous fashion to the corepresentation of spoken words and voices. To do so, we measured recognition memory for spoken words as a function of talker variability and, in parallel, variability in co-occurring sounds.

The sound-specificity effect was probed first in a context that promoted high integrality between words and sounds (Experiment 1b) and then in a context that only involved mere co-occurrence in the stimuli (Experiment 2). The novel integrality element between words and sounds was motivated by the intrinsic link between a word and a voice, which incorporates two crucial components: co-occurrence and integrality. Integrality was implemented by modulating the

Table 7 Mean percentage values for the glimpses resulting from each sound, in both experiments

	Violin sound (%)	Cat sound (%)
Experiment 1b	34.09 (4.93)	46.10 (4.61)
Experiment 2	22.40 (3.84)	44.07 (6.06)

environmental sounds according to the temporal intensity envelope of each individual word and then pairing these modulated versions with the corresponding words.

As expected, the two high-integrality conditions (Experiments 1a–b) revealed robust voice and sound-specificity effects on word recognition memory. After being exposed to words spoken in a particular voice (Experiment 1a) or paired with a particular sound (Experiment 1b), listeners were later less accurate in recognizing the words that were repeated in the different voice, or with the different paired sound, compared to the same-voice/same-sound word repetitions.

Experiment 2 aimed at decoupling the contributions of integrality and mere co-occurrence in the appearance of the effect by eliminating the intensity modulation from the word–sound pairs and keeping everything else identical to Experiment 1b. The absence of an effect in this condition strengthened the argument that integrality between words and sounds elicited the sound-specificity effect. This interpretation was further consolidated by the results of the comparative glimpse analysis performed on the stimuli of Experiment 1b and Experiment 2. This analysis revealed that the contrast between the acoustic glimpses of the same words could not explain the pattern of results in the two experiments. Taken together, the results suggest that co-occurrence per se is not sufficient for the appearance of the sound-specificity effect. However, this conclusion should be interpreted with caution, since Pufahl and Samuel (2014) found a sound-specificity effect in a context that only involved co-occurrence between the words and sounds, without the integrality element implement in the present study. The distinction between the two studies supports the general observation that sound-specificity effects are fragile and conditional on the context in which they are probed.

In summary, the present results make a compelling case for the role of integrality between words and sounds in the appearance of the sound-specificity effect on recognition memory for spoken words. Similar to the intrinsic link between a word and a voice, wherein it is impossible to perceptually segregate one from the other, inducing a similar degree of integration between a word and a co-occurring sound, leads to a similar perceived functional/causal link between the two. Namely, the harder it is to segregate a background sound from a word, the easier it is to perceive the pair as a blended, integrated auditory item.

Integrality as an instance of the "common fate" Gestalt principle

The integrality effect is reminiscent of the "common fate" Gestalt principle of grouping (Wertheimer, 1923). This relates to work by Bregman and colleagues, who adapted the principle to the auditory domain to provide a plausible account for how the auditory system analyses auditory scenes consisting of multiple elements, or "streams" of information (Bregman, 1990). What is particularly relevant here is the fact that the adaptation of the *common fate* principle concerns changes/manipulations in the sound over time, with the heuristics being that if different parts of the spectrum change in a correlated way, they are bound together into a common perceptual unit (Bregman, 1990). In the case of integral words and sounds, the common fate heuristic is a domain-general principle that could easily explain the effect we observed. Co-occurring words and sounds constitute two different auditory "objects" that, in normal circumstances, can be segregated with relative ease, as demonstrated by the results of Experiment 2. However, when modulated to undergo the same changes over time, apparently these two objects blend perceptually to form a unified object, which in turn may promote a similarly unified encoding in memory.

It is worth pointing out that the view of integrality adopted in this study does not dissociate between the integral processing of two co-occurring sound sources (a word and a sound) and failure to segregate them. The question of whether the integral processing of two co-occurring sound sources and failure of segregation are the same or separable phenomena has not been addressed in existing studies of sound-specificity effects (Cooper et al., 2015; Pufahl & Samuel, 2014). In our view, the induced integrality between words and sounds makes their perceptual segregation challenging and, as such, promotes their integration. In this respect, we consider integration and failure of segregation to be two sides of the same phenomenon. There may well be cases where integrality does not fully prevent segregation, due to top-down knowledge, for example, but this possibility would require further testing.

Sound-specificity effects: context-general or an extension of indexical effects?

The emergence of sound-specificity effects has raised the question of whether an external, irrelevant auditory stimulus co-occurring with a spoken word is also included in the memory episode of the word, similar to indexical features. Sound-specificity effects were motivated by indexical effects, which share characteristics with them, and have been probed by means of the typical indexical paradigms. In this respect, it seems appealing to posit a common processing mechanism, and/or place in the memory episode of the word for the voice and the co-occurring sound. However, our results indicate that sound-specificity effects do not readily qualify as an extension of indexical effects. Indeed, we observed that the sound-specificity effect can behave similarly to the voice-specificity effect, but this similarity is constrained by the context in which the word and sound coexist. Namely, we found a sound effect that was comparable to the voice effect by using the same indexical paradigm, but only when the stimuli were manipulated in a way that made the word–sound link highly similar to the word–voice link.

Further, there is evidence indicating that nonauditory changes in the physical context in which spoken words are first encountered also impairs subsequent word recall performance.

One classical example of this phenomenon comes from the study by Godden and Baddeley (1975), in which participants were trained divers who listened to a list of words either on land or 20 feet under water. Every participant was then tested in each of four different exposure/test combinations: (1) land/land; (2) land/water; (3) water/water; (4) water/land. Divers recalled significantly fewer words when the context of test was different from that of exposure, compared to when the context was the same in both phases. This finding was interpreted as supporting a context-dependent memory model, which views memory as sensitive to changes in the environmental context in which words are encountered. This type of evidence weakens a view that treats sound-specificity effects as another type of indexical effects, since memory for spoken words appears to be sensitive to a range of contextual changes associated with a spoken word that are not necessarily confined to the auditory domain.

In addition, evidence from studies that examined sound-specificity effects suggests differences in the processing and encoding in memory of the voice and sound information. Notably, although in their first experiment Pufahl and Samuel (2014) found that sounds co-occurring with words behaved similarly to indexical properties of speech, the results of their second experiment pointed to an asymmetry between the two. More specifically, in their Experiment 2, they had participants hear the same word–sound pairs as in the first experiment, but, this time, the tasks involved judging the animacy of the sound (exposure) and identifying the sound (test) instead of the word. The results did not reveal the anticipated specificity effect in the sound identification performance. This discrepancy indicates that indexical properties of speech and environmental sounds may be processed and retained differently in memory. Similarly, Cooper et al. (2015) reported an asymmetry in the perceptual interference observed in their first experiment: irrelevant indexical feature variation in the speech signal slowed noise classification to a greater extent than irrelevant noise variation slowed speech classification.

Therefore, from a broad perspective, sound-specificity effects may be seen as a type of general context effect. However, as our results show, they cannot be reduced to a *simple* context effect, in the sense that their emergence seems bound to specific contexts. Namely, the results of Experiment 1b and Experiment 2 highlight the conditional nature of the sound-specificity effect and suggest that the encoding of background sounds in memory is contingent upon the extent to which words and sounds are perceived as integral compared to distinct auditory objects. The conditional nature of speech-extrinsic specificity effects has also been pointed out in Cooper et al. (2015), who found that the encoding in memory of the background noise co-occurring with a spoken word was constrained by whether the two were spectrally overlapping or not. Specifically, in their continuous memory experiment, they found that recognition memory for spoken words was impaired as a result of the variation in the background noise across repetitions, but only when the word and noise were spectrally overlapping.

Theoretical implications

The present findings could be accommodated by models of spoken word recognition and the mental lexicon that allow for episodic occurrences of the word and the inclusion of rich auditory details in its memory episode (e.g., Goldinger, 1998; Hawkins & Smith, 2001). For example, Hawkins and Smith (2001) proposed a framework of speech understanding (Polysp) that combines a richly structured, polysystemic linguistic model with psychological and neuropsychological approaches to organization of sensory experience into knowledge. In this view, episodic multimodal sensory experience of speech can be simultaneously processed into different types of linguistic and nonlinguistic knowledge at various stages of abstraction. Accordingly, listeners retain the rich acoustic details of the incoming speech input, at least until the meaning has been extracted. The authors argue that the speech signal could be considered an integral aspect of meaning, rather than only a simple carrier of meaning, and that phonetic categories, like other linguistic categories (e.g., words), are emergent, dynamic, plastic throughout life, and importantly, context-sensitive.

Our results also seem consistent with a distributed view of the mental lexicon, that allows for the co-activation of the co-occurring variation available in the auditory episode of the word (e.g., Elman, 2004, 2009; Gaskell & Marslen-Wilson, 1997, 1999, 2002; Hinton, McClelland, & Rumelhart, 1986; Hintzman, 1986). In Elman's simple recurrent network (SRN) and Gaskell and Marslen-Wilson's distributed cohort model (DCM), lexical representations are defined in a way that may allow the incorporation of the episodic information incidental to spoken words entailed by our results. For instance, in Gaskell and Marslen-Wilson's DCM, as well as in Goldinger's echo model, the mapping of a spoken word to the lexicon is defined as a vector in a high-dimensional space. If this vector was extended to include entries that are not limited to speech-intrinsic dimensions (e.g., speech sounds and voices) but also reflect broader aspects of acoustic variation (e.g., co-occurring sounds), the specificity effects observed here could be accommodated. In a similar fashion, Elman's model posits a distributed representation of word knowledge in which categories emerge over time and are determined by the distributional properties of the input that enters the system. This approach considers words to be cues that activate the co-occurring information with which they have appeared, based on the frequency of the co-occurrence.

It is worth noting that, from a lexical memory perspective, it may seem counterintuitive to posit a memory system that retains redundant and irrelevant information regarding spoken words. Generally speaking, the recognition of a word is not typically aided by the inclusion of details about a certain environmental sound that happens to be present at the time the word is heard. The alternative to having this type of memory system would be to heavily rely on the online processing of

the input, such that the listener continuously evaluates the input and decides what information to include and exclude from the word's auditory episode. Performing such evaluations and decisions under real-time constraints may present serious challenges to the processing capacity, arguably more so than having a memory system with a high-storage capacity. These alternatives reflect what has been broadly termed as the "storage versus computation" challenge, which remains largely unresolved and is beyond the scope of the present discussion (see Baayen, 2007, for a review).

Conclusion

Our results are in line with previous studies that found sound-specificity effects in spoken word processing (Cooper et al., 2015; Pufahl & Samuel, 2014). They suggest that similar to indexical features, background sounds accompanying spoken words may also be assimilated into memory. However, this assimilation seems contingent upon the extent to which words and sounds are perceived as integral compared to distinct auditory entities.

Author note This research was funded by the European Union's Seventh Framework Programme for research, technological development and demonstration under Grant Agreement No. FP7-PEOPLE-2011-290000.

The authors would like to thank Ann R. Bradlow for her comments on an earlier draft of this article.

Appendix A

List of the word stimuli

Word	Semantic category
Dolphin	Animate
Eagle	Animate
Squirrel	Animate
Rabbit	Animate
Baby	Animate
Doctor	Animate
Teacher	Animate
Student	Animate
Actor	Animate
Singer	Animate
Tiger	Animate
Monkey	Animate
Writer	Animate
Donkey	Animate
Zebra	Animate
Hamster	Animate
Panther	Animate
Parrot	Animate
Penguin	Animate
Pigeon	Animate
Scorpion	Animate
Spider	Animate
Turtle	Animate
Lizard	Animate
Dentist	Animate
Waiter	Animate
Dancer	Animate
Artist	Animate
Painter	Animate
Plumber	Animate
Lawyer	Animate
Driver	Animate
Worker	Animate
Banker	Animate
Sculptor	Animate
Soldier	Animate
Athlete	Animate
Chemist	Animate
Scholar	Animate
Leopard	Animate
Basket	Inanimate
Biscuit	Inanimate
Sofa	Inanimate
Table	Inanimate
Bottle	Inanimate
Apple	Inanimate
Orange	Inanimate
Olive	Inanimate
Lemon	Inanimate
Chapel	Inanimate
Cabin	Inanimate
Oven	Inanimate
Pencil	Inanimate
Pillow	Inanimate
Candle	Inanimate
Onion	Inanimate
Taxi	Inanimate
Coffee	Inanimate
Window	Inanimate
Jacket	Inanimate
Bucket	Inanimate
Sugar	Inanimate
Berry	Inanimate
Paper	Inanimate
Mirror	Inanimate
Butter	Inanimate
Carriage	Inanimate
Peanut	Inanimate
Panel	Inanimate
Pepper	Inanimate
Sausage	Inanimate
Ribbon	Inanimate
Building	Inanimate
Bracelet	Inanimate
Necklace	Inanimate
Collar	Inanimate
Blanket	Inanimate
Freezer	Inanimate
Heater	Inanimate
Carpet	Inanimate

Appendix B

Examples of integral and nonintegral stimuli

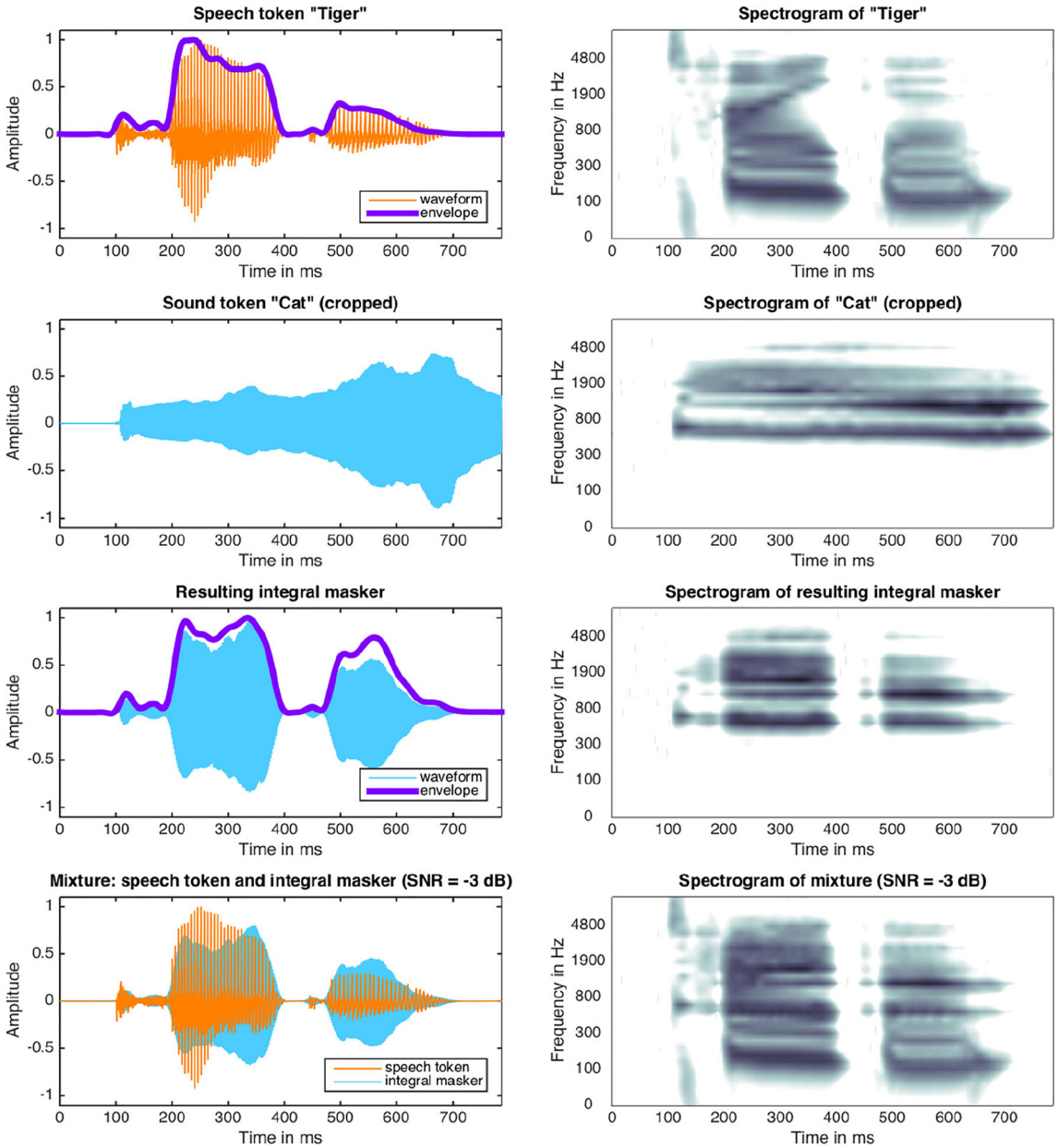

Fig. 2 Processing scheme for generating the integral sounds applied to word *tiger* and sound "cat." Left panel, from top to bottom: The word (orange) and its envelope (purple); the sound cropped to length of the word; integral masker and its envelope; mixture of word (orange) and integral masker. Right panel: Corresponding spectrograms

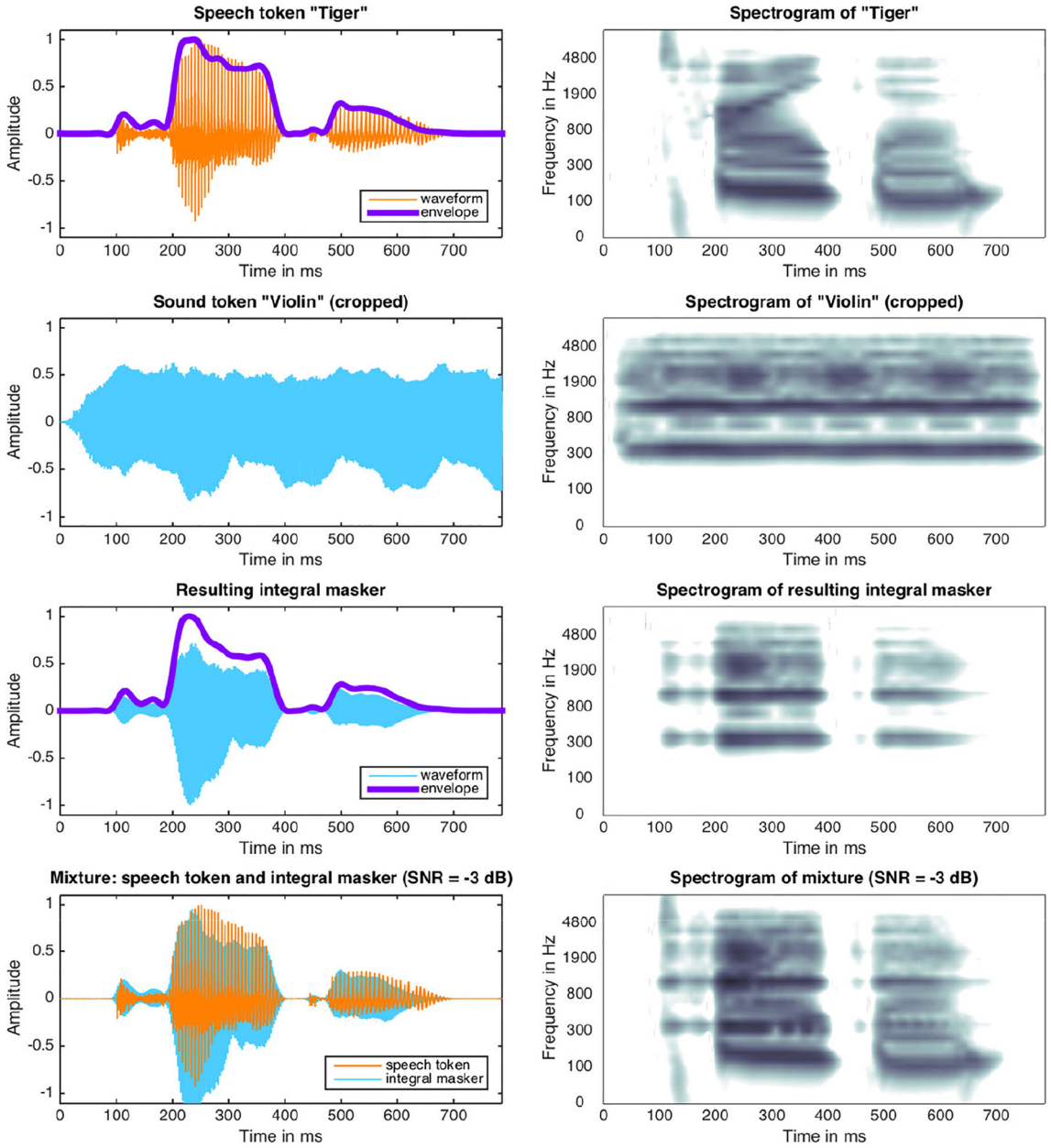

Fig. 3 Processing scheme for generating the integral sounds applied to the word *tiger* and the sound "violin." Left panel, from top to bottom: The word (orange) and its envelope (purple); the sound cropped to length of the word; integral masker and its envelope; mixture of word (orange) and integral masker. Right panel: Corresponding spectrograms

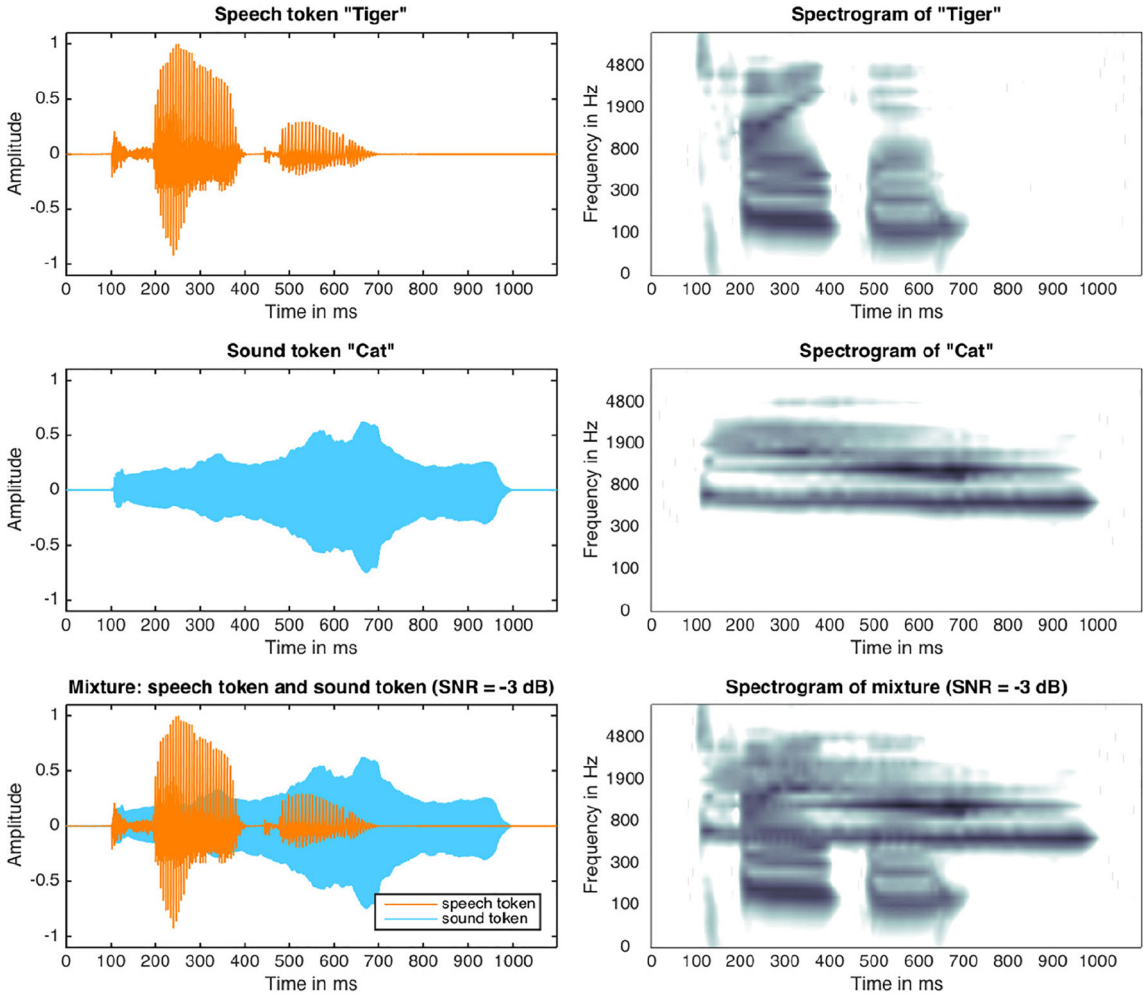

Fig. 4 Processing scheme for mixing words and nonintegral sounds applied to the word *tiger* and the sound "cat." Left panel, from top to bottom: The word (orange), the sound (light blue) with 100-ms silence appended, the corresponding mixture. Right panel: corresponding spectrograms

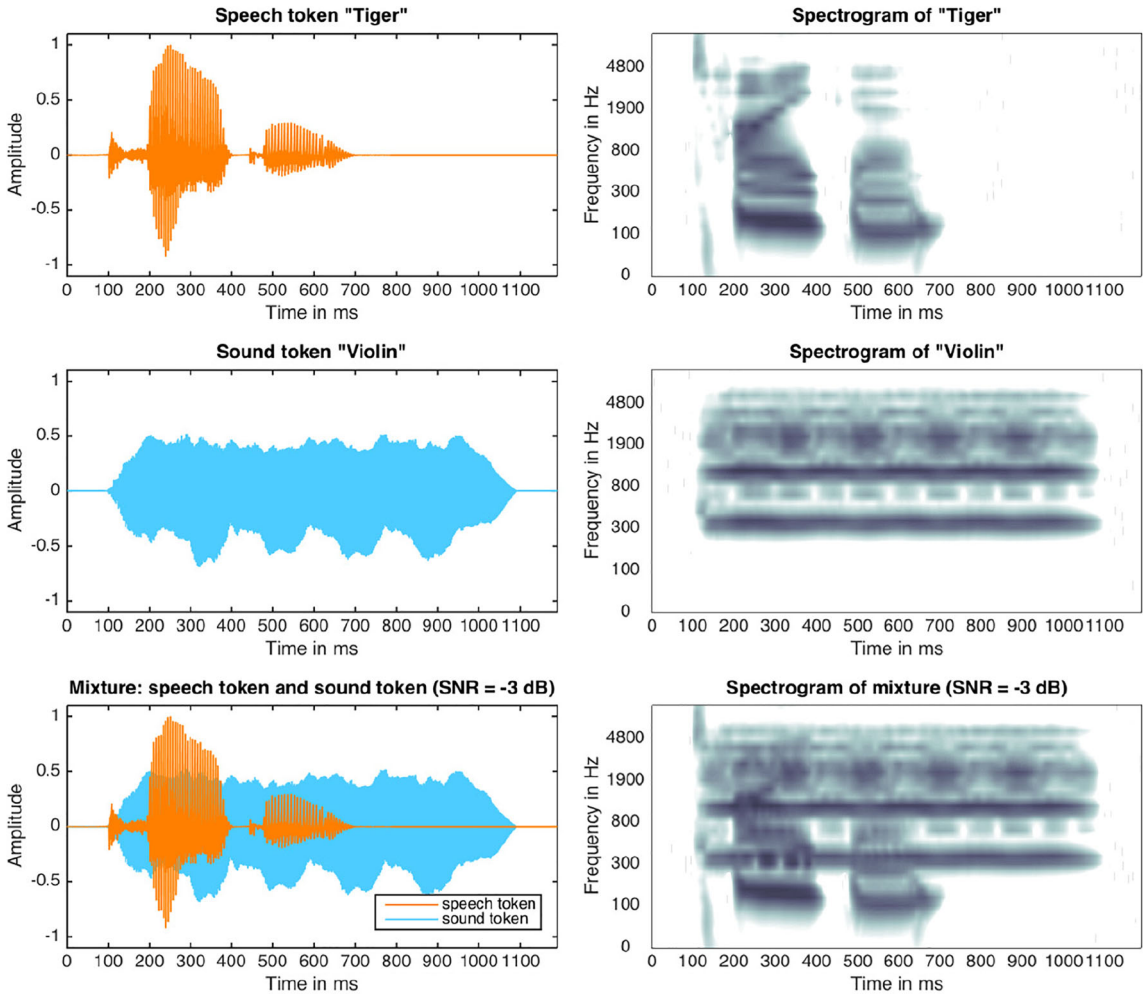

Fig. 5 Processing scheme for mixing words and nonintegral sounds applied to the word *tiger* and the sound "violin." Left panel, from top to bottom: The word (orange), the sound (light blue) with 100-ms silence appended, the corresponding mixture. Right panel: Corresponding spectrograms

References

Baayen, R. H. (2007). Storage and computation in the mental lexicon. In G. Jarema & G. Libben (Eds.), *The mental lexicon: Core perspective* (pp. 81–104). Amsterdam, Netherlands: Elsevier.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Barr, D. J., Levy, R., Scheepers, C., & Tilly, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*. doi:10.18637/jss.v067.i01

Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer [Computer software]. Retrieved from www.praat.org

Bonin, P., Gelin, M., & Bugaiska, A. (2014). Animates are better remembered than inanimates: Further evidence from word and picture stimuli. *Memory & Cognition*, *42*, 370–382.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics, 61*(2), 206–219.

Bregman, A. S. (1990). *Auditory scene analysis.* Cambridge, MA: MIT Press.

Cooke, M. (2003). Glimpsing speech. *Journal of Phonetics, 31,* 579–584.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America, 119,* 1562–1573.

Cooper, A., Brouwer, S., & Bradlow, A. R. (2015). Interdependent processing and encoding of speech and concurrent background noise. *Attention, Perception, & Psychophysics, 77*(4), 1342–1357.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heading the voice of experience: The role of talker variation in lexical access. *Cognition, 108*, 633–664.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2012). Word learning under adverse listening conditions: Context-specific recognition. *Language and Cognitive Processes, 27,* 1021–1038.

Church, B. A., & Schacter, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 496–509.

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences, 8*(7), 301–306.

Elman, J. L. (2009). On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive Science, 33*(4), 547–582.

Forster, K. I., & Forster, J. C. (2003). DMDX: A windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers, 35*, 116–124.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12*, 613–656.

Gaskell, M. G., & Marslen-Wilson, W. D. (1999). Ambiguity, competition, and blending in spoken word recognition. *Cognitive Science, 23*(4), 439–462.

Gaskell, M. G., & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology, 45*, 220–566.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and underwater. *British Journal of Psychology*, *66*(3), 325–331.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and cognition, 22*, 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes: An episodic theory of lexical access. *Psychological Review,105,* 251–279.

Hawkins, S., & Smith, R. (2001). Polysp: A polysystemic, phonetically rich approach to speech understanding. *Italian Journal of Linguistics-Rivista di Linguistica 13,* 99–188.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (pp. 77–109). Cambridge, MA: MIT Press.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,*411–428.

Jørgensen, S., & Dau, T. (2011). Predicting speech intelligibility based on the signal-to-noise envelope power ratio after modulation-frequency selective processing. Journal of the Acoustical Society of America, 130(6), 1475–1487.

Jørgensen, S., Ewert, S., & Dau, T. (2013). A multi-resolution envelope-power based model for speech intelligibility. *Journal of the Acoustical Society of America, 134*(1)*,* 436–446.

Jusczyk, P. W., & Luce, P. A. (2002). Speech perception. In S. Yantis & H. E. Pashler (Eds.), *Stevens' handbook of experimental psychology* (Vol. 1. 3rd ed., pp. 493–536). New York, NY: John Wiley & Sons.

Lachs, L., McMichael, K., & Pisoni, D. B. (2003). Speech perception and implicit memory: Evidence for detailed episodic encoding. In J. S. Bowers & C. J. Marsolek (Eds.), *Rethinking implicit memory* (pp. 215–235). Oxford, UK: Oxford University Press.

Luce, P. A., Goldinger, S. D., Auer, E. T., & Vitevitch, M. S. (2000). Phonetic priming, neighborhood activation, and PARSYN. *Perception & Psychophysics, 62,* 615–625.

Luce, P. A., & Lyons, E. (1998). Specificity of memory representation for spoken words. *Memory & Cognition, 26,* 708–715.

Luce, P. A., & McLennan, C. T. (2005). Spoken word recognition: The challenge of variation. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception,* (pp. 591–609). Malden, MA: Blackwell.

Mattys, S. L. & Liss, J. M. (2008). On building models of spoken-word recognition: When there is as much to learn from natural "oddities" as from artificial normality. *Perception & Psychophysics*, *70*, 1235–1242.

McClelland, J. L., & Elman, J. L. (1986). Interactive processes in speech recognition: The TRACE model. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition,* (pp. 58–121). Cambridge, MA: MIT Press.

Moore, B. C. J. (2003). Temporal integration and context effects in hearing. *Journal of Phonetics 31,* 563–574.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85,* 365–378.

Nairne, J. S., VanArsdall, J. E., Pandeirada, J. N. S., Cogdill, M., & LeBreton, J. M. (2013). Adaptive memory: The mnemonic value of animacy. *Psychological Science, 24,* 2099–2105.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52,* 189–234.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker- contingent process. *Psychological Science, 5,* 42–46.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 309–328.

Patterson, R. D., Holdsworth, J., Nimmo-Smith, I., & Rice, P. (1988). SVOS Final Report: The Auditory Filterbank. Technical Report 2341. Medical Research Council Applied Psychology Unit, University of Cambridge, Cambridge

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 9–32). San Diego, CA: Academic Press.

Pisoni, D. B., & Levi, S. V. (2007). Some observations on representations and representational specificity in speech perception and spoken word recognition. In G. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 3–18). Oxford, UK: Oxford University Press.

Pufahl, A., & Samuel, A. G. (2014). How lexical is the lexicon? Evidence for integrated auditory memory representations. *Cognitive Psychology, 70,* 1–30.

Schacter, D. L., & Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(5), 915–930.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304.

Sheffert, S. M. (1998a). Contributions of surface and conceptual information to recognition memory. *Perception & Psychophysics, 60,* 1141–1152.

Sheffert, S. M. (1998b). Format-specificity effects on auditory word priming. *Memory & Cognition, 26,* 591–598.

VanArsdall, J. E., Nairne, J. S., Pandeirada, J. N. S., & Blunt, J. R. (2013). Adaptive memory: Animacy processing produces mnemonic advantages. *Experimental Psychology, 60,* 172–178.

Vitevitch, M. S. (2003). Change deafness: The inability to detect changes between two voices. *Journal of Experimental Psychology: Human Perception and Performance, 29,* 333–342.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt II. *Psycologische Forschung*, *4*, 301–350.