Laboratory Phonology
Journal of the Association for Laboratory Phonology

Cooper, A. and Bradlow, A. R. 2017 Talker and background noise specificity in spoken word recognition memory. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 8(1): 29, pp.1–15, DOI: https://doi.org/10.5334/labphon.99

JOURNAL ARTICLE

Talker and background noise specificity in spoken word recognition memory

Angela Cooper[1] and Ann R. Bradlow[2]

[1] Department of Psychology, University of Toronto, Mississauga, CA

[2] Department of Linguistics, Northwestern University, Evanston, US

Corresponding author: Angela Cooper (angela.cooper@utoronto.ca)

Prior research has demonstrated that listeners are sensitive to changes in the indexical (talker-specific) characteristics of speech input, suggesting that these signal-intrinsic features are integrally encoded in memory for spoken words. Given that listeners frequently must contend with concurrent environmental noise, to what extent do they also encode signal-extrinsic details? Native English listeners' explicit memory for spoken English monosyllabic and disyllabic words was assessed as a function of consistency versus variation in the talker's voice (talker condition) and background noise (noise condition) using a delayed recognition memory paradigm. The speech and noise signals were spectrally-separated, such that changes in a simultaneously presented non-speech signal (background noise) from exposure to test would not be accompanied by concomitant changes in the target speech signal. The results revealed that listeners can encode both signal-intrinsic talker and signal-extrinsic noise information into integrated cognitive representations, critically even when the two auditory streams are spectrally non-overlapping. However, the extent to which extra-linguistic episodic information is encoded alongside linguistic information appears to be modulated by syllabic characteristics, with specificity effects found only for monosyllabic items. These findings suggest that encoding and retrieval of episodic information during spoken word processing may be modulated by lexical characteristics.

Keywords: speech perception; recognition memory; noise

1. Introduction

For successful spoken word recognition to take place, listeners must match the incoming auditory input to the appropriate lexical representation stored in memory. This is a complex process, as individual instances of a given word vary as a result of changes in talker, speaking style, or a whole host of other linguistic, paralinguistic, and situation-specific characteristics. Previous research has posited that one way listeners could handle this variability is by encoding the idiosyncratic characteristics of a particular speech event into memory and then retrieving such rich representations for processing of subsequent speech events with the same or similar instance-specific details (e.g., Goldinger, 1998). An extreme version of this notion might suggest that listeners encode all of the perceptual details of a speech event, everything from talker information and background noise to even information about the location where the speech event occurred. While listeners may not encode the color of the speaker's clothing along with a given speech exemplar (Sheffert & Fowler, 1995), recent work has provided evidence of perceptual integration and encoding of background noise that is concurrent with target speech (Cooper et al., 2015; Creel et al., 2012; Pufahl & Samuel, 2014). Thus, as a step towards delimiting which perceptual dimensions external to the speech signal listeners are encoding into memory, the present work uses a recognition memory paradigm (e.g., Goldinger, 1996) to examine the degree to which two types of extra-linguistic information, namely talker identity and environmental noise, are integrally encoded alongside linguistic information.

1.1. Integration of linguistic and signal-intrinsic non-linguistic information

Traditional accounts of spoken word recognition have posited that speakers map words that they hear onto abstract lexical representations from which all non-linguistic information has been stripped (e.g., see Pisoni, 1997 for a discussion of this position). A strict version of this notion has been challenged by a burgeoning body of evidence over the past couple of decades demonstrating that linguistic processing is influenced by non-linguistic features of the speech signal, that is, by indexical information, including gender, talker identity, speaking rate, and the speaker's affective state (e.g., Bradlow & Pisoni, 1999; Goh, 2005; Goldinger, 1996; Johnsrude et al., 2013; Kaganovich et al., 2006; Mullennix & Pisoni, 1990; Palmeri et al., 1993; Schacter & Church, 1992; Sheffert & Fowler, 1995, and many others). This work has demonstrated that listeners are sensitive to changes in the indexical features of the input, such that listeners were found to be less accurate at identifying or recalling items when the surface characteristics changed from their initial exposure to the items relative to when the surface characteristics remained consistent. This indexical specificity effect, where linguistic processing is influenced by instance-specific information, has been found across a variety of different tasks, including continuous recognition memory (Bradlow et al., 1999; Palmeri et al., 1993), delayed recognition memory (Goh, 2005; Goldinger, 1996; Mattys & Liss, 2008), cued re-call (Church & Schacter, 1994), long-term repetition priming, and lexical decision (González & McLennan, 2007; McLennan & Luce, 2005). Moreover, perceptual experience with a talker's voice characteristics has been found to transfer to enhanced linguistic processing, with higher shadowing or word recognition accuracy for familiar talkers relative to unfamiliar ones (Johnsrude et al., 2013; Newman & Evers, 2007; Nygaard & Pisoni, 1998).

In light of this empirical evidence, models of spoken word recognition have proposed that listeners encode into memory detailed episodic information (e.g., Goldinger, 1996, 1998; Johnson, 2006; Pierrehumbert, 2001). A strong version of exemplar-based accounts might predict that if all contextual details, including those that are related to the broader context in which the speech events occur as well as those that are related to talker variation, are encoded into memory, then they would all have an impact on linguistic processing. Some within-talker sources of variability, including speaking rate (Bradlow et al., 1999), speaking style, and emotional tone of voice (Krestar & McLennan, 2013; Sommers & Barcroft, 2006), have yielded similar effects on spoken word recognition as cross-talker variation. However, not all variation has been found to impact linguistic processing. Bradlow et al. (1999) reported finding significantly better recognition memory for items produced with the same voice or speech rate as the original presentation but no difference between items produced with the same or different amplitude. Similarly, variability in fundamental frequency, created by global shifts in pitch tracks, did not have a significant impact on English word recognition (Sommers & Barcroft, 2006). Based on these findings, these authors hypothesized that spoken word recognition will only be impaired by variability that affects linguistically-relevant features of the speech signal (the Phonetic Relevance Hypothesis, Sommers & Barcroft, 2006). For example, while duration can in and of itself serve as the primary, or even sole, acoustic correlate of a phonological contrast (e.g., many languages contrast long versus short vowels without a secondary contrast in vowel quality), amplitude is not a primary acoustic correlate of a phonological contrast. In this sense, then, variation in speaking rate (which is directly related to duration variation for particular speech sounds) is 'phonetically relevant,' and listeners should be sensitive to such variation. In contrast, amplitude variation is not phonetically relevant (at least, not as a primary cue to a phonological contrast), and listeners should therefore not encode this dimension of acoustic variation.

Moreover, recent research has investigated differences in the presence of specificity effects and hypothesized that they may arise from differences in processing speed at retrieval (e.g., McLennan & Luce, 2005), attention during encoding (Theodore et al., 2015), or overall attention levels (Tuft et al., 2016). The processing speed account is based on observations of greater indexical specificity effects for dysarthritic speech (Mattys & Liss, 2008), foreign-accented speech (McLennan & González, 2012), and other imposed task difficulties (McLennan & Luce, 2005) relative to normal, native-accented speech in easier task conditions. Linguistic and indexical information are posited to be processed at different rates, with indexical information arriving relatively later in processing (McLennan & Luce, 2005), which is why indexical specificity effects are hypothesized to be particularly salient in slower processing contexts (e.g., dysarthritic or foreign-accented speech).

However, recent work has suggested that attention is the mediating factor in observing specificity effects. Theodore et al. (2015) found talker-specificity effects only when listeners were asked to specifically attend to talker characteristics (i.e., gender) during the exposure phase but not when they attended to linguistic features of the speech signal (i.e., lexical or syntactic characteristics), despite equivalent processing times at retrieval. They postulated that heightened attention to indexical (or non-linguistic) dimensions of the speech signal enhances the salience of those dimensions within the encoded representation, thereby increasing the activation of the relevant episodic traces, resulting in stronger specificity effects. Similarly, Tuft et al. (2016) found that intermixing taboo words in certain contexts served to increase listeners' overall attention, resulting in enhanced talker specificity effects. Overall, the accumulation of evidence suggests that specificity effects in speech recognition and encoding can be somewhat fragile and variable across tasks and stimuli (see Pufahl & Samuel, 2014, and Strori, 2016, for related discussion).

1.2. Noise in speech processing

The integral encoding of linguistic and non-linguistic features, such as talker information, speaking rate, and speaking style, may also be related to the fact that these features are inherent to the speech signal itself, stemming from the same sound source. However, speakers must frequently contend with environmental noise that co-occurs with the speech signal (see Mattys et al., 2012 for a review), which raises the question as to the extent to which listeners encode contextual details that are external to the speech sound source. Given the evidence for the Phonetic Relevance Hypothesis (Sommers & Barcroft, 2006), one might predict that variation in environmental noise would not have a substantive impact on linguistic processing, as changes in environmental noise do not apparently cue any phonetic contrasts, similar to amplitude variation, and are thus not phonetically relevant for listeners. On the other hand, when noise spectrally overlaps with speech, it may hinder listeners' access to certain phonetic cues. Additionally, though not phonetically relevant, noise is part of the broader auditory context in which speech is uttered and may not be completely irrelevant to the overall communicative situation. For example, background noise could conceivably be relevant for ambiguity resolution: One might speculate that background sounds of flowing water may bias listeners towards one meaning of the ambiguous lexical item, *bank*, whereas background sounds of cash registers may bias listeners to the alternate meaning. Moreover, the presence of background noise may influence overall processing speed and/or attention, other factors that have been suggested to influence encoding specificity and retrieval (e.g., McLennan & Luce, 2005; Mattys & Liss, 2008; Theodore et al., 2015; Tuft et al., 2016).

Recent work has begun to investigate the notion that listeners encode more than just speech-intrinsic perceptual details of a particular speech event into an integrated cognitive representation (Cooper et al., 2015; Creel et al., 2012; Pufahl & Samuel, 2014; Strori, 2016). Creel et al. (2012) trained listeners on nonsense word-meaning associations, with training either in the clear or in white noise and the test phase either matching or mismatching the initial exposure conditions. Listeners were found to be faster and more accurate in matching exposure conditions (e.g., exposure and test in the clear or exposure and test in noise) than mismatching conditions (e.g., exposure in noise and test in the clear). Similarly, in a series of experiments, Pufahl and Samuel (2014) exposed listeners to words paired with unique environmental sounds (e.g., a dog bark, a phone ring) in an animate/inanimate judgment task. This was followed by an identification task where the environmental sound was either the same or different as the initial exposure. The frequency of exposure was also manipulated, either 1 or 8 repetitions, over the course of the exposure phase. Consistent with Creel et al. (2012), listeners were found to be more accurate at identifying the words if the noise matched the initial exposure. However, contrary to the effect of lexical frequency on imitation of talker-specific characteristics found by Goldinger (1998), no effect of repeated presentation was found, such that hearing a particular episodic pairing 8 times rather than just once did not have a substantive impact on the strength of the exemplar specificity effect. These studies suggest that listeners encode the perceptual details of a speech event, including details that are extrinsic to the speech signal, such as environmental noise.

While Pufahl and Samuel (2014) and Creel et al. (2012) make a case for the claim that the cognitive representation of words may involve integrated representations of the environmental noise along with the word exemplar, the fact that their stimuli involved speech and noise that spectrally overlapped with one another—that is the frequency range of the speech and noise signals overlapped—allows for another possible explanation. It is plausible that listeners segregate concurrent speech and noise into separate representations in memory, such that, in the case of the Pufahl and Samuel (2014) and Creel et al. (2012) stimuli, the encoded word exemplars would have spectro-temporal gaps as a result of masking from the noise. For example, if the word "cat" were presented concurrently with white noise at 3–4 kHz at 0 dB signal-to-noise ratio (SNR), then a listener might segregate this noise from the speech signal and store them separately. The speech sample's exemplar would thus contain a spectral gap, or at least a band of degraded speech, at 3–4 kHz as a result of the noise masking the speech signal at these frequencies. This type of representation would also be expected to yield the word recognition and word learning findings from Pufahl and Samuel (2014) and Creel et al. (2012). The lower accuracy rates arising from a change in environmental noise from first to second exposure could have stemmed from the fact that the acoustic characteristics of the segregated speech exemplar from the first exposure would not be an identical match to the characteristics of the presented speech input on the second (test) exposure.[1] While this account still involves highly detailed representations of words heard in the context of simultaneous non-linguistic auditory events, it leaves open the possibility that only those auditory

[1] Note that a possible process of phoneme restoration might lead listeners to perceptually restore 'missing' or degraded speech information in noise-masked signals, in which case the mismatch from first to second exposure may be in terms of processing (i.e., mismatch between signals that engage a process of restoration versus those that do not).

events that are closely linked to the auditory word stimulus at the stimulus level will be integrated into the word's lexical representation.

Therefore, a stronger test for the hypothesis that listeners store integrated representations of co-occurring auditory events would be to spectrally separate the speech and noise, such that the acoustic characteristics of the speech would be identical in both same-exemplar and different-exemplar repetitions, with only the characteristics of the noise changing in the case of different-exemplar repetitions. Cooper et al. (2015) tested this hypothesis using disyllabic words with spectrally-segregated and spectrally-overlapping noise in a classification task (Garner speeded classification) as well as a continuous recognition memory paradigm. The Garner task assessed the extent to which listeners could ignore irrelevant background noise variation when asked to classify speech samples (e.g., male/female classification). They reported perceptually integrated processing of speech and noise information at the relatively low (pre-lexical) level of processing tapped by the Garner task, in that classification along a speech dimension was slowed down by variation in the task-irrelevant noise dimension and vice versa, regardless of whether the speech and noise were spectrally overlapping or separated. However, in the recall-based task of continuous recognition memory, they found a specificity effect only in the spectrally-overlapping condition but not in the spectrally segregated condition.

1.3. Current study

The present work builds on this prior research by investigating the extent to which indexical (talker identity) and noise information are integrally encoded in memory with linguistic information using (a) a delayed recognition memory paradigm, (b) spectrally segregated speech and noise signals, and (c) word stimuli with different inherent lexical characteristics. Does the integral encoding of speech and noise persist after a delay? What are the constraints on this joint encoding? To investigate these questions, the current study followed prior work (Goldinger, 1996; Mattys & Liss, 2008) and employed a delayed recognition memory paradigm that was comprised of two phases: An exposure phase and a recognition phase. Listeners first completed a word identification task, where they were exposed to a set of items (divided between two female talkers in the talker condition or combined with two kinds of noise in the noise condition). This was followed by a recognition memory task, which included new words not heard in the word identification task and old words (i.e., words heard in the word identification task). Half of the old items were the same exemplars provided in the word identification task, and half were different exemplars (with either a change in talker or noise). Listeners were asked to recall whether or not they had heard the word in the first task.

By using a delayed recognition memory paradigm rather than a continuous recognition memory paradigm (as in Cooper et al., 2015), the present work sought to examine the encoding of this speech-extrinsic feature (e.g., noise) at a level beyond what is tapped into by the Garner and continuous recognition memory tasks, both of which required minimal contact with the mental lexicon and limited linguistic processing. The degree of delay between initial and repeated tokens is substantially larger in a delayed recognition memory task than in a continuous recognition memory task, with a maximum of 16 words intervening between initial and repeated items in the latter case (approximately 40 seconds) but several hundred words in addition to a 3-minute task delay in the former case. Moreover, the initial exposure phase in the present work is a word identification task, which involves greater contact with the mental lexicon during encoding.

If listeners encode all concurrent perceptual details of a speech event, including speech-intrinsic information (e.g., talker identity) as well as information extrinsic to the speech signal (e.g., environmental noise), then listeners should be more accurate at recognizing that they had encountered a word previously if the surface details (talker or noise) of an item match the surface details of the original presentation. As described above, previous work has strongly suggested exactly this specificity of encoding for spectrally overlapping speech and noise signals (Pufahl & Samuel, 2014), but perhaps not for spectrally separated speech and noise (Cooper et al., 2015), at least at the level tapped by a continuous recognition memory paradigm. The present study asks if this specificity of encoding extends to a test of delayed recognition memory with spectrally separated speech and noise signals, providing a stronger test for the extent (and limitations) of integrated cognitive representations.

As discussed above, earlier work (Cooper et al., 2015) with a continuous recognition memory task indicated that specificity effects may be limited to signals with spectrally-overlapping speech and noise. Moreover, other work (e.g., McLennan & Luce, 2005) has suggested that increasing cognitive demand yields more robust specificity effects, as it slows processing and allows specificity effects to emerge. In light of these prior studies, we ask whether noise and speech need to be spectrally integrated in order to be integrally encoded, or whether spectrally-segregated speech and noise items are faster and easier to process, thus attenuating any specificity effect.

Finally, the current study presented listeners with both monosyllabic and disyllabic items. Monosyllables, being from high density lexical neighbourhoods and containing limited bottom-up input due to shorter durations, require relatively more cognitive resources to identify than disyllables (e.g., Pisoni et al., 1985). Thus, monosyllables may be lexically 'harder' to recognize and therefore processed slower relative to disyllables which may, in turn, yield stronger specificity effects. The literature investigating specificity effects has not been consistent in the syllabic characteristics of the stimuli, with some studies using monosyllabic items (e.g., Goldinger, 1996; Matty & Liss, 2008; Theodore et al., 2015), others using disyllabic items (e.g., McLennan & Luce, 2005), and yet others using a combination of both mono- and disyllables (Pufahl & Samuel, 2014), nor has it included an examination of the potential influence of this stimulus characteristic on specificity effects. Given that monosyllabic and disyllabic words differ in a way that could have a significant impact on their processing and encoding, which may interact with the observation of specificity effects, both types of words were included and compared in the present work.

Taken together, the three main features of the current study—a delayed recognition memory task, spectrally separated speech and noise signals, and the inclusion of both monosyllabic and disyllabic lexical items—provide a more stringent test of specificity in spoken word recognition memory than prior literature, which should help resolve open questions regarding the extent (and limits) to which speech-intrinsic (talker-specific) and speech-extrinsic (noise-specific) indexical information are retained in memory for spoken words.

2. Method

2.1. Participants

Sixty-four native American English listeners, self-reporting no speech or hearing deficits at the time of testing, were included in this experiment and received course credit for their participation. Listeners were randomly assigned to either the talker (n = 32; Female = 18; M_{age} = 20 years, *SD* = 1.36) or noise (n = 32, Female = 22; M_{age} = 20 years, *SD* = 1.35) conditions. This sample size was selected to satisfy counterbalancing requirements and is approximately consistent with prior work (e.g., Cooper et al., 2015; Krestar & McLennan, 2013).

Table 1: Mean number of phonemes, neighbourhood density, and frequency for monosyllabic and disyllabic stimuli. Standard deviations provided in parentheses.

	# of phonemes	Neighbourhood Density	SubtLexUS frequency

Monosyllables	3 (0.5)	25 (7.8);
		range = 8–40
			197 (308.3);
			range 0.06–1959

Disyllables	5 (0.8)	6 (5.2);
		range = 0–26
			13 (15.5);
			range = 0.43–86

2.2. Stimuli

The stimulus materials were 296 English words, including 96 disyllabic items (from Cooper et al., 2015) and 200 monosyllabic items (**Table 1**).[2] Mono- and disyllabic words differed significantly on both frequency and neighbourhood density ($p < 0.001$). Lexical characteristics were obtained from IPhOD (Vaden et al., 2009).

Words were produced in citation form by two female American English talkers (one from the Midwest, the other from the Pacific Northwest) and recorded at a 44,100 Hz sampling rate. The disyllable productions were taken from Experiment 1B of Cooper et al. (2015). The stimuli were normalized for duration (490 ms for monosyllables and 542 ms for disyllables), low-pass filtered at 5 kHz, and normalized for root-mean-square (RMS) amplitude in Praat (Boersma & Weenink, 2013). For the noise condition, the productions of a single female talker from the talker condition were used to create two sets of stimuli by combining the speech files with narrow band-pass filtered white noise from 7–10 kHz and, for the other set, a 6 kHz pure tone. The relative amplitude of the speech and noise signals was set to 0 dB SNR, and, while temporally concurrent, the speech and noise did not spectrally overlap with each other.

2.3. Procedure

Each experiment session was composed of two phases: An exposure phase and a recognition phase. Stimuli were presented over Sony MDR-V700 headphones at a comfortable listening volume in sound-attenuated booths. In the exposure phase, listeners were presented with half of the total number of stimuli (148 items: 48 disyllables, 100 monosyllables). Each trial consisted of an individually-presented word, and listeners were required to type the word they heard and press the enter key in order to initiate the next trial. In the talker condition, items were presented in the clear (i.e., no added noise or tone). Seventy-four words were produced by one talker and 74 by the other talker, which were randomly-presented to listeners. In the noise condition, productions from one of the female talkers in the talker condition were presented, where half of the items were white noise-combined and the other half were pure-tone combined. Participants were not asked to memorize the items they were identifying nor were they informed that they would be asked to recall these items later in the session.

In order to avoid recency effects for the recognition memory task (Goh, 2005), a 3-minute math filler task was administered between the exposure and recognition phases, with questions such as “$6 + 4 - 3 \times 1 - 7$.” Participants were given the 20-question

[2] The asymmetry in the number of monosyllables and disyllables stems from the fact that the original experiment design included a frequency manipulation (100 high frequency and 100 low frequency monosyllables). Analyses revealed frequency did not significantly interact with any relevant factors and, for brevity, this factor has not been included here.

math sheet and were asked to work their way through as many questions as they could within 3 minutes.

During the recognition phase, participants were presented with all 296 stimuli, including the 148 exposure items as well as 148 novel items. For each word, they were asked to indicate as quickly and accurately as possible whether or not they had heard that item during the encoding phase. If they recalled hearing the word, they pressed a button labeled "Old"; however, if they did not think the word had been previously presented, they pressed a button labeled "New." They were provided 3.5 seconds to respond and a 500 ms delay after each button press before the next trial began. In the talker condition, half of the exposure items were produced by the same talker as in the exposure phase, and the other half of the exposure items were produced by a different talker than in the exposure phase. In the noise condition, half of the exposure items were presented with the same noise as in the exposure phase, the other half with different noise, while the talker was held constant. Talkers and noise types were randomly presented in their respective conditions.

Which words appeared as exposure and novel items as well as which words were same or different exemplars at test were counterbalanced across participants, such that each word appeared with each talker or noise combination in both the exposure and recognition phases and occurred as a same and different exemplar. Furthermore, the button order in the recognition phase was also counterbalanced across participants.

3. Results

The mean proportion of correctly identified words in the exposure phase was tabulated for each participant. A correct identification entailed an exact match with the target item. Spelling and obvious typing errors were also considered to be accurate. Homophones (e.g. *wait*, *weight*) were flagged but included as accurate items for the overall word identification accuracy measure. Mean identification accuracy was 93.9% for the talker condition and 93.7% for the noise condition, indicating that the presence of non-spectrally overlapping noise in the noise condition did not inhibit identification of the items relative to talker changes in the talker condition.

For the recognition phase, words that were inaccurately identified by participants in the exposure phase were excluded from the analysis of their test phase. For example, if a participant misidentified the word *theme* as *seem* during the exposure phase, the trial containing *theme* in the test phase would be excluded for that participant. If they had not accurately perceived the identity of the item initially, then they might not recall the appropriate item during the recognition memory task. No participant scored below 90% accuracy on the word identification task, and as such, all participants were included in the analysis of the recognition memory task. Overall recognition memory accuracy was calculated. To account for any potential response bias, mean percentages of hits (defined as correctly responding "old" when the item was "old") and false alarms (responding "old" when the item was "new") were calculated and used to estimate d' and β by condition and syllable type.[3] Computing d' entailed the subtraction of the normalized probability of false alarms from the normalized probability of hits (**Table 2**).

[3] Beta analysis showed that listeners have a significantly looser (more liberal) criterion (lower beta scores) in the talker condition ($M = 0.55$, $SD = 0.58$) than in the noise condition ($M = 0.88$, $SD = 0.59$), indicating that listeners were overall more likely to respond "old" than "new" in the talker condition compared to in the noise condition. It is unclear exactly why this is the case, but it may indicate greater interference of speech-extrinsic noise with spoken word encoding and retrieval. In view of these divergent criterion settings, separate hit rate analyses for each condition were run, yielding a pattern of results that mirrors the pattern reported above. That is, for both conditions, specificity effects are observed for monosyllables but not for disyllables.

Following Goldinger (1996), exemplar-specificity effects were analyzed by examining hit rates to same- and different-exemplar repetitions. **Figure 1** depicts the mean proportion of hit rates for same- and different-exemplar repetitions by condition and syllable. From **Figure 1**, we see that listeners were overall more accurate at identifying disyllabic relative to monosyllabic items ($M = 72\%$ vs. 57%) and more accurate in the talker relative to the noise condition ($M = 64\%$ vs. 60%). For disyllabic items, there appears to be no difference in accuracy between same- and different-exemplar trials; however, for monosyllabic items, a difference between these trial types appears to emerge, with higher accuracy on same-exemplar trials as compared to different-exemplar trials.

To investigate these patterns, the data of the "old" trials were analyzed with a generalized linear mixed-effects model with a logistic linking function (Baayen et al., 2008) with hits as the binary dependent variable. Contrast-coded fixed effects included Exemplar Match (Same, Different), Condition (Talker, Noise), and Syllable (Monosyllable, Disyllable), as well as their 2- and 3-way interactions. The maximal random effects structure that would converge was employed, which included random intercepts for Participant and Item, as well as random slopes for Exemplar Match, Syllable and Exemplar Match × Syllable by

Table 2: Calculated *d′* values for the recognition memory task by condition and syllable type. Standard error provided in parentheses.

Condition	**Syllable**	***d′***
Noise	Monosyllable	1.56 (0.06)
	Disyllable	2.04 (0.08)
Talker	Monosyllable	1.36 (0.07)
	Disyllable	1.92 (0.09)

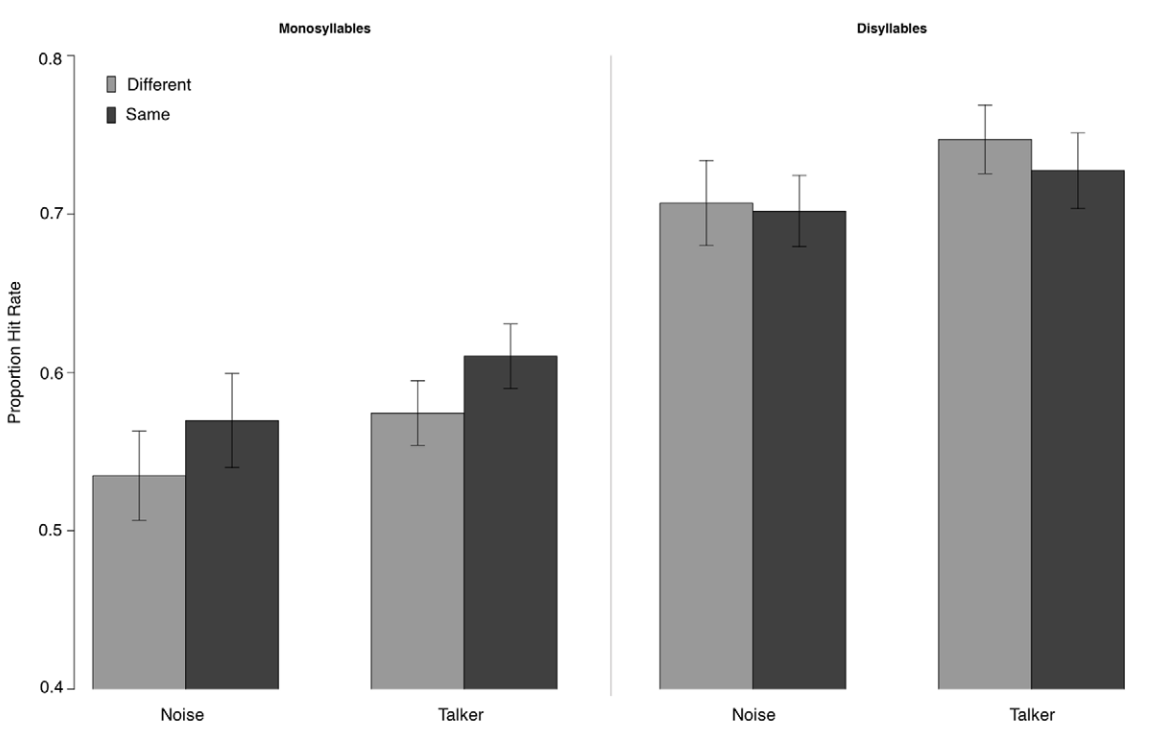

Figure 1: Mean proportion hit rate for Same trials and Different trials by Condition and Syllable. Errors bars denote +/− 1 standard error.

participant, and Condition by item. Model comparisons were performed to determine whether the inclusion of each of these fixed effects and their interactions made a significant contribution to the model.

The results of these analyses revealed a significant main effect of Syllable ($\beta = -0.75$, *SE* $\beta = 0.12$, $\chi^2(1) = 36.512$, $p < 0.001$), where overall listeners were more accurate at recalling disyllables relative to monosyllables. Additionally, the 2-way interaction of Exemplar Match x Syllable was also significant ($\beta = 0.27$, *SE* $\beta = 0.104$, $\chi^2(1) = 6.472$, $p = 0.011$). No other effects reached significance ($\chi^2 < 1.22$, $p > 0.269$). To further investigate the 2-way interaction, separate models were constructed for the monosyllable and disyllable data.

A model with monosyllable accuracy as the dependent variable was constructed, with a contrast-coded fixed effect of Exemplar Match (Different, Same), random intercepts for Participant and Item, and a random slope for Exemplar Match by participant. A significant effect of Exemplar Match was found, whereby listeners were significantly more accurate on same-exemplar repetitions relative to different-exemplar repetitions ($\beta = 0.19$, *SE* $\beta = 0.059$, $\chi^2(1) = 9.852$, $p = 0.002$). Accuracy rates for "old" trials with disyllabic words were similarly analyzed, with Exemplar Match as a contrast-coded fixed effect and the same random effects structure as the previous model. Unlike the results with the monosyllabic items, model comparisons did not reveal a significant main effect of exemplar match ($\chi^2 = 0.829$, $p = 0.363$).[4]

Similar analyses were run on response latencies, with only correct responses to old words included. Additionally, latencies that were 3 standard deviations from the mean were removed. Log-transformed latencies were submitted to a linear mixed effects regression model with contrast-coded fixed effects for Exemplar Match, Condition and Syllable, random intercepts for Participant and Item, and random slopes for Exemplar Match, Syllable and Exemplar Match × Syllable by participant, and Condition by item. A significant effect of Syllable was found ($\beta = 0.04$, *SE* $\beta = 0.008$, $\chi^2(1) = 16.86$, $p < 0.001$), with faster responses to disyllabic items ($M = 1587$ ms) relative to monosyllabic items ($M = 1640$ ms). No other main effects or interactions were significant ($\chi^2 < 2.96$, $p > 0.090$).

4. Discussion

The present study investigated the extent to which talker identity and noise information are integrally encoded with linguistic information in memory. In prior work, initial evidence for the perceptual integration of speech and noise was found at a relatively low-level of processing through the Garner speeded classification task and in a recall-based task (Cooper et al., 2015). Here, we sought to examine the integrality of speech and noise encoding at a level of processing beyond the level tapped into by the Garner and continuous recognition memory tasks. In order to further investigate the constraints on joint encoding, participants completed a task requiring delayed recognition memory. The results of the present experiments were consistent with prior work on talker effects on the processing of linguistic information (e.g., Bradlow et al., 1999; Goh, 2005; Goldinger, 1996), such that listeners were found to be more accurate in a recognition memory task when the talker was consistent from exposure to test relative to when the talker changed.

[4] We acknowledge that there were fewer disyllabic than monosyllabic items and so this lack of an effect could be attributed to the lower statistical power. However, specificity effects have been found for the same number of disyllabic items in Cooper et al. (2015), for the spectrally-overlapping speech and noise items, and also for a group of L2 listeners performing the exact same task in the present study (not reported here). We also conducted the same set of analyses reported here using half of the monosyllables and still found a significant specificity effect for only the monosyllables.

Moreover, the current study extended previous findings on listeners' sensitivity to talker variation to include a dimension not inherent to the speech signal, namely background noise. Prior work has primarily examined the encoding of noise and linguistic information where the noise was spectrally-overlapping with the speech signal (Creel et al., 2012; Pufahl & Samuel, 2014). However, a stronger test of the hypothesis that listeners store integrated speech and noise representations was to utilize items where the speech and noise were spectrally segregated, so that any change in noise would not yield a concomitant change in the speech signal due to a change in masking pattern. Indeed, results from the current study using spectrally-segregated speech and noise items revealed a significant exemplar specificity effect, in that retaining the same noise from exposure to test resulted in a benefit in recognition accuracy relative to when the noise changed. Because the speech and noise were spectrally non-overlapping in the stimulus, the speech portion of the signal was always identical in both same-noise and different-noise trials, with no portions of the speech degraded by the effects of the spectrally non-overlapping noise masker. Thus, the same-noise benefit in this recognition memory task indicates that, in certain contexts, listeners encode both within-signal speech features (e.g., talker identity) as well speech-extrinsic information (e.g., background noise) into integrated cognitive representations, critically even when the two auditory streams are spectrally non-overlapping.

The present findings also point to an interaction between a lexical characteristic, namely monosyllable versus disyllable, and the specificity of encoding, as this same-exemplar benefit only emerged for monosyllabic items. One critical difference between monosyllabic and disyllabic words is the density of their lexical neighbourhoods. In the current study, the monosyllabic words had an average of 25 phonological neighbours, while the disyllabic words had an average of only 5 neighbours. Monosyllables, being more difficult to identify during the encoding phase ($M = 91\%$ for monosyllables, $M = 99\%$ for disyllables) and more easily confused with other items during the recognition phase, will slow processing in listeners relative to disyllables. Indeed, an examination of listeners' reaction times revealed that, even despite being of longer duration, disyllabic items were responded to faster ($M = 1586$ ms) than monosyllabic items ($M = 1639$ ms), suggesting that listeners needed longer to process the monosyllables, as a result of their confusability with other lexical neighbours.

These findings help to elucidate the results in Cooper et al. (2015), where no specificity effect was found in the continuous recognition memory task. Note that the stimuli utilized in that task were the identical spectrally-segregated disyllabic items used in the present experiment. The fact that listeners were slower to respond to the monosyllables than the disyllables, and that the specificity effect only emerged on the former, may be taken as evidence consistent with the processing speed account (e.g., McLennan & Luce, 2005). However, it is also conceivable that because monosyllables are more difficult to identify, listeners would have needed to attend to the signal more closely during encoding (i.e., the word identification task), which would support an attentional account (e.g., Tuft et al., 2016). The present results do not differentiate these theories, and it remains an open empirical question as to whether it is processing speed or attention that modulates the strength of exemplar specificity effects.

The current work provides support for a model of lexical access where lexical representations are, at least in part or at some level, episodic (indeed, there is evidence to suggest the existence of both episodic and abstract information within the lexicon, e.g., Cutler, 2008; Goldinger, 2007). The present results provide converging evidence with Pufahl and Samuel (2014) in support of the notion that listeners store detailed episodic information—not only speech-intrinsic indexical information but also speech-extrinsic auditory information—that co-occurs with particular lexical items. During the exposure phase, episodic traces are activated, priming listeners and enabling them to have more accurate recognition at test. This priming is enhanced when all dimensions (lexical, indexical, non-speech auditory information) of the integrated representation match from exposure to test. It is worth noting that these specificity effects are relatively smaller than what has been reported in certain studies previously (e.g., Theodore et al., 2015). One possible explanation might be that two female speakers were used (in the talker condition), while prior research has often used male and female speakers. However, the fact that we found smaller specificity effects for both talker and noise conditions suggests an alternative explanation, since the noise condition only involved a single female talker and two different noise types. The fact that listeners' attention was directed towards lexical information during encoding (as a consequence of performing a transcription task) may have contributed to these smaller effects. Theodore et al. (2015) posited that specificity effects are attenuated when listeners' attention is not focused on the relevant dimension of interest, such as talker identity.

Given that prior work has suggested that phonetically irrelevant features (e.g., amplitude, visible speaker information) are not encoded or retained in memory (Bradlow et al., 1999; Sheffert & Fowler, 1995), one might wonder why the current study found that background noise can be encoded, at least under some circumstances (e.g., when co-occurring with monosyllabic words from relatively dense lexical neighbourhoods). It could be the case that amplitude differences are such a simple change (uniform level adjustment) and that visible speaker information is from a different sensory domain (visual rather than auditory), that listeners can easily segregate these features during encoding. In contrast, as discussed in the Introduction, background noise may not be as phonetically irrelevant or as easy-to-segregate as unidimensional level adjustments and visual contextual information, given that it can in some cases impact the energetic properties of the signal (by masking or distorting them). Such masking or distortion could certainly affect access to certain phonetic cues. Moreover, background noise may contribute real-world, semantic-pragmatic, contextual information that impacts spoken language processing and comprehension. Thus, an association between or joint encoding of speech and background noise, even when spectrally non-overlapping, may be retained rather than lost as appears to be the case for uniform amplitude or visual contextual information.

In sum, the present work supports models of spoken word recognition that incorporate episodic information, including both speech-intrinsic (e.g., talker identity) as well as speech-extrinsic (e.g., noise) details, at some level(s) of the cognitive representation of speech. Moreover, the extent to which extra-linguistic episodic information is encoded alongside linguistic information appears to be modulated by the syllabic characteristic of the stimuli, with greater benefits of specific acoustic attribute matching across exposure and recognition phases in a delayed recognition memory task for monosyllabic items with relatively dense phonological neighbourhoods (compared to disyllabic words). It remains for future research to establish conclusively whether these modulating influences are determined by cognitive mechanisms related to processing speed, attention, and/or cognitive load.

Acknowledgements

This research was supported by NIH-NIDCD grant R01-DC005794 awarded to Ann R. Bradlow. We thank Chun Liang Chan, Emily Kahn, and Victoria Kuritza for their research and technical support. Portions of this work were presented at *Architecture and Mechanisms of Language Processing (AMLaP 2014)* in Edinburgh, UK.

Competing Interests

The authors have no competing interests to declare.

References

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*(4), 390–412. DOI: https://doi.org/10.1016/j.jml.2007.12.005

Boersma, P., & Weenink, D. 2013. Praat: Doing phonetics by computer. Software.

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. DOI: https://doi.org/10.3758/BF03206883

Bradlow, A. R., & Pisoni, D. B. 1999. Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*(4), 2074–85. Retrieved from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3468898&tool=pmcentrez&rendertype=abstract. DOI: https://doi.org/10.1121/1.427952

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*(3), 521–33. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/8207370. DOI: https://doi.org/10.1037/0278-7393.20.3.521

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. DOI: https://doi.org/10.3758/s13414-015-0855-z

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. 2012. Word learning under adverse listening conditions: Context-specific recognition. *Language and Cognitive Processes*, *27*(7/8), 1021–1038. DOI: https://doi.org/10.1080/01690965.2011.610597

Cutler, A. 2008. The abstract representations in speech processing. *Quarterly Journal of Experimental Psychology*, *61*(11), 1601–1619. DOI: https://doi.org/10.1080/13803390802218542

Goh, W. D. 2005. Talker variability and recognition memory: Instance-specific and voice-specific effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(1), 40–53. DOI: https://doi.org/10.1037/0278-7393.31.1.40

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*(5), 1166–83. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/8926483. DOI: https://doi.org/10.1037/0278-7393.22.5.1166

Goldinger, S. D. 1998. Echoes of echoes? An Episodic Theory of Lexical Access. *Psychological Review*, *105*(2), 251–79. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/9577239. DOI: https://doi.org/10.1037/0033-295X.105.2.251

Goldinger, S. D. 2007. A complementary-systems approach to abstract and episodic speech perception. In: *Proceedings of the 16th International Congress of Phonetic Sciences*, 49–54. Saarbrucken.

González, J., & McLennan, C. T. 2007. Hemispheric differences in indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology. Human Perception and Performance*, *33*(2), 410–24. DOI: https://doi.org/10.1037/0096-1523.33.2.410

Johnson, K. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, *34*(4), 485–499. DOI: https://doi.org/10.1016/j.wocn.2005.08.004

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. 2013. Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, *24*(10), 1995–2004. DOI: https://doi.org/10.1177/0956797613482467

Kaganovich, N., Francis, A. L., & Melara, R. D. 2006. Electrophysiological evidence for early interaction between talker and linguistic information during speech perception. *Brain Research*, *1114*(1), 161–72. DOI: https://doi.org/10.1016/j.brainres.2006.07.049

Krestar, M. L., & McLennan, C. T. 2013. Examining the effects of variation in emotional tone of voice on spoken word recognition. *Quarterly Journal of Experimental Psychology*, *66*(9), 1793–802. DOI: https://doi.org/10.1080/17470218.2013.766897

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. 2012. Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7/8), 953–978. DOI: https://doi.org/10.1080/01690965.2012.705006

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 artificial normality. *Perception & Psychophysics*, *70*(7), 1235–42. DOI: https://doi.org/10.3758/PP.70.7.1235

McLennan, C. T., & González, J. 2012. Examining talker effects in the perception of native- and foreign-accented speech. *Attention, Perception, & Psychophysics*, *74*(5), 824–830. DOI: https://doi.org/10.3758/s13414-012-0315-y

McLennan, C. T., & Luce, P. A. 2005. Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 306–21. DOI: https://doi.org/10.1037/0278-7393.31.2.306

Mullennix, J. W., & Pisoni, D. B. 1990. Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, *47*(4), 379–390. DOI: https://doi.org/10.3758/BF03210878

Newman, R. S., & Evers, S. 2007. The effect of talker familiarity on stream segregation. *Journal of Phonetics*, *35*(1), 85–103. DOI: https://doi.org/10.1016/j.wocn.2005.10.004

Nygaard, L. C., & Pisoni, D. B. 1998. Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–76. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/9599989. DOI: https://doi.org/10.3758/BF03206860

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*(2), 309–28. Retrieved from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3499966&tool=pmcentrez&rendertype=abstract. DOI: https://doi.org/10.1037/0278-7393.19.2.309

Pierrehumbert, J. B. 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In: Bybee, J., & Hopper, P. (eds.), *Frequency effects and the emergence of lexical structure*, 137–157. Amsterdam: John Benjamins Publishing Company. DOI: https://doi.org/10.1075/tsl.45.08pie

Pisoni, D. B. 1997. Some Thoughts on "Normalization" in Speech Perception. In: Johnson, K., & Mullennix, J. W. (eds.), *Talker variability in speech processing*, 9–32. San Diego: Academic Press.

Pisoni, D. B., Nusbaum, H. C., Luce, P. A., & Slowiaczek, L. M. 1985. Speech Perception, Word Recognition and the Structure of the Lexicon. *Speech Communication*, *4*(1–3), 75–95. DOI: https://doi.org/10.1016/0167-6393(85)90037-8

Pufahl, A., & Samuel, A. G. 2014. How lexical is the lexicon? Evidence for integrated auditory memory representations. *Cognitive Psychology*, *70*, 1–30. DOI: https://doi.org/10.1016/j.cogpsych.2014.01.001

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), 915–30. Retrieved from: http://www.ncbi.nlm.nih.gov/pubmed/1402716. DOI: https://doi.org/10.1037/0278-7393.18.5.915

Sheffert, S. M., & Fowler, C. A. 1995. The Effects of Voice and Visible Speaker Change on Memory for Spoken Words. *Journal of Memory and Language, 34*, 665–685. DOI: https://doi.org/10.1006/jmla.1995.1030

Sommers, M. S., & Barcroft, J. 2006. Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *The Journal of the Acoustical Society of America*, *119*(4), 2406–2146. DOI: https://doi.org/10.1121/1.2171836

Strori, D. 2016. *Specificity effects in spoken word recognition and the nature of lexical representations in memory*. University of York.

Theodore, R. M., Blumstein, S. E., & Luthra, S. 2015. Attention modulates specificity effects in spoken word recognition: Challenges to the time-course hypothesis. *Attention, Perception, & Psychophysics*. DOI: https://doi.org/10.3758/s13414-015-0854-0

Tuft, S. E., McLennan, C. T., & Krestar, M. L. 2016. Hearing taboo words can result in early talker effects in word recognition for female listeners. *Quarterly Journal of Experimental Psychology*, *218*(August). DOI: https://doi.org/10.1080/17470218.2016.1253757

Vaden, K. I., Halpin, H. R., & Hickok, G. S. 2009. Irvine Phonotactic Online Dictionary, Version 2.0. [online database]. Available from: http://www.iphod.com.

How to cite this article: Cooper, A. and Bradlow, A. R. 2017 Talker and background noise specificity in spoken word recognition memory. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 8(1): 29, pp. 1–15, DOI: https://doi.org/10.5334/labphon.99

Submitted: 24 June 2017 **Accepted:** 12 September 2017 **Published:** 21 November 2017

]u[*Laboratory Phonology: Journal of the Association for Laboratory Phonology* is a peer-reviewed open access journal published by Ubiquity Press. OPEN ACCESS