Talker-Specificity and Token-Specificity in Recognition Memory

William Clapp[a]*, Charlotte Vaughn[b], Simon Todd[c], and Meghan Sumner[a]

[a]Department of Linguistics
Stanford University
Margaret Jacks Hall, Bldg. 460
Stanford, CA 94301
sumner@stanford.edu
wsclapp@stanford.edu

[b]Language Science Center
University of Maryland
2130 H.J. Patterson Hall
College Park, MD 20742
cvaughn@umd.edu

[c]Department of Linguistics
University of California Santa Barbara
South Hall 3432
Santa Barbara, CA 93106
sjtodd@ucsb.edu

*Corresponding Author

Abstract [Word count: 233]

Given any feasible amount of time, a talker would never be able to produce the same word twice in an identical manner. Yet recognition memory experiments have consistently used identical tokens to demonstrate that listeners recognize a word more quickly and accurately when it is repeated by the same talker than by a different talker. These *talker-specificity effects* have served as the foundation of decades of research in speech perception, but the use of identical tokens introduces a confound: Is it the talker or the physical stimulus that drives these effects? And consequently, to what extent do listeners encode the high-level acoustic characteristics of a talker's voice? We investigate the roles of token and talker repetition in two continuous recognition memory experiments. In Exp. 1, listeners heard the voice of one talker, with either Identical or Novel repeated tokens. In Exp. 2, listeners heard two demographically matched talkers, with same-voice repetitions being either Identical or Novel. Classic talker-specificity effects were replicated in both Identical and Novel tokens, but recognition of Identical tokens was in some cases stronger than recognition of Novel tokens. In addition, recognition memory varied across demographically matched talkers, suggesting stronger episodic encoding for one talker than for the other. We argue that novel tokens should serve as the default design for similar studies and that consideration of talker variation can advance our understanding of encoding and memory differences more broadly.

Keywords: speech, recognition memory, episodic memory, talker specificity, perception

1. Introduction

The observation that listeners remember words better when repeated in the same voice than when repeated in a different voice has been one of the most influential findings in the speech perception literature in recent decades. Early theories of speech perception assumed that acoustic events are mapped to abstract representations, with phonologically irrelevant phonetic detail discarded by the listener (e.g., Jakobson, Fant, & Halle, 1952; Kiparsky, 1973; Stampe, 1979; Stevens, 2002). A landmark series of recognition memory studies provided empirical evidence that phonetic detail is not discarded by the listener, but is rather retained in memory and facilitates lexical access for future perception of speech by the same talker (Craik & Kirsner, 1974; Goldinger, 1996; Palmeri, Goldinger, & Pisoni., 1993). These studies paved the way for a new set of episodic or exemplar-based theories of speech perception according to which memory traces of acoustic events are not only retained, but form the substrate of the speech processing mechanism in cognition (Goldinger, 1998; Johnson, 1997; Pierrehumbert, 2001).

The evidence for the preservation of detailed acoustic traces of spoken words in memory stems largely from a series of experiments in the continuous recognition memory paradigm, in which participants hear a series of words, one at a time, and must indicate for each word whether it is “Old” (i.e., a repetition of a word heard earlier in the series) or “New” (i.e., not a repetition). In seminal work, Craik & Kirsner (1974) found that participants were faster and more accurate to identify a repeated word as “Old” when it was heard in the same voice as its original presentation than when it was heard in a different voice. The effect was consistent across intervening trials: even with many intervening trials between a word and its repetition, words repeated in the same voice were still recognized more accurately than words repeated in a different voice. Taken together, these results suggest that a same-voice recognition boost exists and that it is consistent across time delays, which in turn suggests that acoustic information is encoded in long-term memory of spoken words, rather being stripped away or limited to working memory. This boost has since been referred to as a *talker-specificity effect.*

Following Palmeri et al. (1993), who reproduced Craik & Kirsner’s (1974) earlier findings in a scaled-up replication, talker-specificity effects were widely recognized in the field. The attention of researchers in speech perception largely shifted to identifying the characteristics of acoustic information that are or are not encoded in memory. Some of these studies were conducted outside the recognition memory paradigm (e.g., Church & Schacter, 1994; Nygaard, Sommers, & Pisoni., 1995), and sought to investigate the roles of variables such as voice, intonation contour, overall f0, amplitude, and speech rate. Bradlow, Nygaard, & Pisoni(1999) reported findings showing that in the context of a continuous recognition memory experiment, a repeated word was more likely to be recognized as Old when it matched the first presentation in speech rate, but not when it matched in amplitude. In addition to replicating previously described talker-specificity effects, these findings helped demonstrate that not all sources of stimulus variability are encoded equally well in memory. Another line of inquiry has sought to investigate the encoding of non-speech sounds in memory traces of spoken words, ranging from broadband aperiodic noise to environmental sounds such as barking dogs and ringing telephones (e.g., Cooper, Brouwer, & Bradlow, 2015; Creel, Aslin, & Tanenhaus, 2012; Pufahl & Samuel, 2014; Strori, Zaar, Cooke, & Mattys, 2018). Results from this body of work indicate that while memory representations of words can certainly contain non-speech acoustic information, such information is encoded more strongly when it is integral to the speech signal. This contrast may help elucidate the distinction between memory for physical acoustic events and memory for higher-level properties of talker attributes. Further, preliminary evidence suggests that talker-specificity effects may be found asymmetrically across individual talkers (Clapp, Vaughn, & Sumner, 2023), but to date, this hypothesis has not been tested directly. If talkers are encoded with different levels of specificity, it also follows that the type of token heard at a word's repetition (novel vs. identical) would influence recognition performance asymmetrically across talkers. The present paper examines these questions as a secondary goal.

1.1 Memory for acoustic patterns

Despite the great strides that have been made in research on talker-specificity effects and related phenomena in recent decades, several fundamental questions have either gone uninvestigated or remain unanswered. One of these is the question of whether the talker-specificity effect reflects encoding of high-level voice characteristics and talker identity or low-level acoustic characteristics of the tokens themselves. In other words, is listener performance better on same-talker than different-talker trials because the *talker* is the same or because the *token* is the same? In all of the studies cited thus far, the tokens heard at first and second presentation of variable-matched pairs (e.g. same talker, same speech rate, same amplitude, and so on, depending on the independent variable in question) have been identical, even though this methodological decision creates an experimental confound in favor of the repetition benefit. That is, studies have conflated repetition of the same talker (or speech rate, etc.) with repetition of the exact physical stimulus[1]. This confound leaves open the possibility that rather than demonstrating *talker-specificity* effects, past research has shown *token-specificity* effects. In this case, the same-voice recognition benefit may not be the result of listeners encoding talker-level acoustic information (i.e., all the things that make voices recognizable and unique: spectral characteristics, VOT distributions, degrees of nasal leak, intonation contours, etc.), but rather of listeners imprinting a highly specific acoustic event in memory and recognizing it as a statistical match to their previous experience. Increased memory performance for exact acoustic matches would be consistent with the finding that listeners are capable of encoding even linguistically meaningless acoustic patterns in both non-linguistic contexts (Viswanathan, Rémy., Bacon-Macé, & Thorpe, 2016; Winkler, Korzykov, Gumenyuk, Cowan, Linkenkaer-Hansen, Alho, Ilmoniemi, & Näätänen, 2002) and linguistic contexts (Cooper et al., 2015; Pufahl & Samuel, 2014). If it were to turn out that listeners respond to different tokens produced by the same talker in the same way as they respond to tokens produced by different talkers (i.e. without a recognition boost), then a core foundational piece of episodic models of speech perception would be weakened. These models rely on the ability of listeners to leverage acoustic memory traces to facilitate interpretation of similar acoustic patterns. If it were found that recognition only benefited from *identical* rather than *similar* acoustic matches, it would be difficult to argue that episodic encoding was the basis of a functional speech perception system, given that such a system would have a hard time accounting for generalization across and within talkers, social groups, contexts, exemplars, and so on.

The repetition of an identical token also creates a context impossible in natural speech scenarios. The ecological validity of using identical tokens in examinations of exemplar effects has been raised in the literature before, in the context of repetition priming experiments (Hanique, Aalders, & Ernestus, 2014; Morano, ten Bosch, & Ernestus, 2019). These studies have not treated memory for novel and identical tokens as the central object of study and have yielded mixed results, perhaps due to the use of different designs and stimulus construction methods.

[1] Interestingly, Bradlow et al. (1999) may have demonstrated a token-specificity effect. While stimuli reflecting different speech rates were recorded separately, stimuli differing in amplitude were generated by rescaling a single token. Thus, repetitions differing in amplitude in fact consisted of the same acoustic footprint, whereas repetitions differing in speech rate consisted of entirely unique tokens. Given the confound we point out here, the result that speech rate but not amplitude is subject to a specificity effect may be at least in part a result of tokens used.

These mixed findings highlight the need for a dedicated study designed specifically to test how memory for novel and identical tokens differs. Our study provides a straightforward investigation of these factors in the context of recognition memory.

1.2 Talker-level asymmetries

The first goal of this study is to examine whether talker identity (as opposed to simply token identity) is encoded in memory traces of spoken words, as discussed above. The second goal of this study to examine the role of talker identity more closely by exploring asymmetries in responses to individual talkers and to investigate the effect of switching repeated-token identity (*identical* vs. *novel*) on responses to the individual talkers (see also Hanique et al., 2014; Mattys & Liss, 2008; McLennan & Gonzalez, 2012 for talker variation in other paradigms). It has been typical in the analysis of recognition memory results to treat responses to all voices as equivalent, using the *same-voice*/*different-voice* distinction as the only talker-based independent variable (e.g., Palmeri et al., 1993). Some studies have gone further in organizing voices into a two-dimensional space of perceptual similarity and analyzing responses to voices based on proximity within this plane (Goh, 2005; Goldinger, 1996). Other studies have been designed to investigate group-level encoding, for example by Clopper, Tamati, & Pierrehumbert (2017), who addressed asymmetrical encoding across two dialects of North American English. While such studies have been illuminating for the field and helped highlight the high-level group dynamics at play in memory for speech, it has long been known that variation is a hallmark of language production (e.g., Bell, 1984; Labov, 1966), and there is no one-to-one correspondence between the use of a particular variant and membership in a particular macro-demographic category (Eckert, 2008; 2012). Under a theory of speech perception that treats exemplars as asymmetrically weighted based on social characteristics (e.g., Clapp et al., 2023; Sumner, Kim, King, & McGowan, 2014), and given that the use of sociolinguistic variables ranges widely among individuals, it follows that the strength of lexical encoding and recall may vary between talkers even when they are matched for macro-demographic category membership. The present study is only a small step in squaring the psycholinguistic study of speech perception with recent developments in sociolinguistics, but takes seriously the need for theories of episodic representations to draw on these findings.

1.3 Current study

In this study, we investigated whether listeners encode high-level voice characteristics and leverage those memory traces to facilitate word recognition. Specifically, we compared the recognition of identical tokens to the recognition of novel tokens within and across talkers. Given that a growing body of research suggests that response patterns may be at least partially conditioned based on the voices themselves, this study also investigates the possibility of asymmetric encoding across talkers. Two continuous recognition memory experiments were conducted using stimuli produced by two white, male, middle-class, 28-year-old talkers living in Chicago but originally from smaller Midwestern cities (henceforth referred to as *M1* and *M2*). Words repeated in the same voice were presented with either *identical* tokens (i.e. exactly the same audio file heard twice) or *novel* tokens (i.e., two different recordings of the same talker producing the same word during the same recording session). As a first step, in Exp. 1, we directly compared memory for identical tokens with memory for novel tokens by presenting each participant with only one repetition type and stimuli from only one talker, either M1 or M2. We also asked whether the influence of token type on responses was equivalent for listeners hearing each of the voices. To directly address talker-specificity effects, in Exp. 2, all listeners heard one repetition type (either identical or novel tokens) but both voices, which created a context where talker-specificity effects could emerge. Exp. 2 was designed to address whether talker-specificity effects are replicable when same-voice repetitions are heard with novel tokens, whether asymmetries between the encoding of each talker's voice emerge more clearly with novel than with identical repetitions, and whether the presence of multiple talkers in the same context enhances the encoding asymmetries across voices.

2. Experiment 1

Exp. 1 investigates the role of specific token identity in the memory encoding and recall of spoken words, and the stability of these effects across two talkers matched in terms of macro-demographic categories. This experiment allows us to investigate whether there are differences in recognition memory for identical versus novel tokens for a single talker. If high-level voice characteristics are not encoded at a word's first presentation, and recognition is instead dependent on fine-grained acoustic properties of the token itself, we would expect that performance would be stronger when an identical token is repeated than when the repetition is a novel token. On the other hand, if listeners do encode high-level voice characteristics, hearing a repetition with the natural variability present in a single talker's voice would not inhibit recognition, and performance would be similar regardless of the type of token heard at repetition. Participants were placed in one of two token conditions (*Identical* or *Novel*) and one of two talker conditions (*M1* or *M2*). Using stimuli from two separate talkers served the dual purpose of ensuring that effects were not the result of idiosyncratic facts about a single talker's voice and allowing an analysis of asymmetric encoding and recall of distinct voices.

2.1 Methods

2.1.1 Participants

A total of 444 participants completed Exp. 1 via the participant recruitment platform Prolific. All participants provided informed consent. Based on Prolific's built-in pre-screening tool, the experiment was made available only to individuals who reported that they lived in the United States and were American by nationality, were English-speaking monolinguals, and had never had hearing loss or hearing-related difficulties. Responses from 89 of these individuals were removed from analysis on the basis that they either failed to respond to at least 85% of trials or had a miss rate of 10% or more on trials where a word was repeated either 1 or 2 trials after its first presentation, leading to a total sample size of 355 participants. The sample size was determined in advance by using the *pwr* package (Champely, Ekstrom, Dalgaard, Gill, Weibelzahl, Anandkumar, Ford, Volcic, & De Rosario, 2017) in R to estimate the number of participants needed to achieve an effect size of 0.06, given significance threshold $a = 0.02$ and power $\beta = 0.9$. Participants were compensated $3.77, and the procedure took 15.27 minutes on average to complete.

2.1.2 Stimuli & Design

Stimuli. Stimulus words were selected from a subset of the SUBTLEXus corpus (Brysbaert & New, 2009), which was filtered to include only monosyllabic words falling between the 33rd and 90th percentile of frequency according to the corpus's Lg10CD measure. From this list, 183 words were randomly selected to act as stimuli. The mean word frequency of this final subset was 28.0 per million.

Auditory stimuli were produced by two white males living in Chicago (originally from the Indianapolis and Dubuque metropolitan areas), aged 28 at the time of recording. Each talker read through the list of all 183 words three times, each time in a different randomized order. Out of these productions, two tokens were selected for use in the experiment. Tokens were excluded if the recording contained extraneous noise, a slip of the tongue, or an anomalous production. If all tokens of a word were possible candidates for inclusion in the experiment, the second and third were selected by default. By design, we wanted novel tokens to exhibit naturalistic degrees of variation, and thus we did not control them for acoustic similarity. Acoustic analysis indicated natural inter-speaker variation. For example, M1's tokens were slightly longer in duration than M2's (M1: $\bar{x} = 628$ ms, $\sigma = 94$ ms; M2: $\bar{x} = 481$ ms, $\sigma = 90$ ms). M1's tokens were also consistently produced with a higher mean *f*0 (in semitones, re: 1 Hz, M1: $\bar{x} = 84$ ST, $\sigma = 0.65$ ST; M2: $\bar{x} = 75$ ST, $\sigma = 0.81$ ST).

Audio was captured in a quiet room through an Electro-Voice RE320 dynamic microphone at a sampling rate of 48 kHz and a 24-bit depth. After recording, ambient noise was removed using the iZotope RX 8 audio editing software, and all tokens were normalized to a common mean intensity using Praat (Boersma & Weenink, 2021).

Design. Participants were randomly placed in one of two TokenConditions (Identical or Novel) and in one of two TalkerConditions (M1 or M2), leading to a 2 x 2 matrix of between-subjects conditions. In the Identical condition, repeated words were presented with exactly the same audio file as was heard at the word's first presentation. In the Novel condition, the second presentation was a different recording of a different production of the same word with all else held constant. In the M1 condition, all words were produced by talker M1 and in the M2 condition, all words were produced by talker M2.

Each participant completed three phases: practice, memory load, and test (see Figure 2.1). Each phase was identical except that participants received feedback on responses in the practice phase but not thereafter. The memory load phase differed from the main test phase only insofar as its data were not included in the analysis. Participants were not made aware of any distinction between the memory load and test phases, and it was not indicated to them when the test phase had begun. No word from the memory load phase was repeated after the first word of the test phase was presented. Every word in the experiment was played exactly twice, with no filler words. We refer to the combination of each stimulus word's Old and New presentations as a *word pair*. There were 8 word pairs in the practice round, 16 pairs in memory load, and 140 pairs in the test phase. This led to a total of 164 word pairs, or a total of 328 trials. The 164 words used in each participant's procedure were a randomly selected subset of the total 183 recorded words. More words were recorded than were necessary in case any word was found to be unusable after processing (e.g., because only one production was valid or a talker failed to produce the correct word). All words were usable and therefore included in the experiment.

The number of intervening trials between a given word and its repetition—henceforth referred to as *Lag*—was used as an independent variable in analysis and ranged from 1 to 65. The distribution of Lags was heavily left-modal, with a mean 18.54 and a median of 11. Note that because previous studies in the continuous recognition memory paradigm used ANOVAs for analysis, Lag was traditionally treated as a categorical variable, where each value fell into one of several discrete groups (e.g. Lag could be 2, 4, 8, 16, 32, or 64, but not any of the intermediate values). Because the present study's results were modeled with regression, Lag was converted into a continuous variable. The overall distribution was similar to previous studies', but intermediate values were included as well (i.e. any value between 1 and 65 was a valid Lag). The precise distribution was generated pseudorandomly and uniquely for each participant with an algorithm written in JavaScript designed to roughly approximate the distribution used by Palmeri et al. (1993) but without excluding intermediate values, leading to the left-modal distribution described above.

Participants made responses by pressing either the "D" or "K" key on their computer keyboard. For half of participants, the "D" key was associated with a response of "New" and the "K" key was associated with a response of "Old", and for the other half of participants, this was reversed. A visual prompt remained on the screen throughout the experiment to remind participants which key was associated with which response.

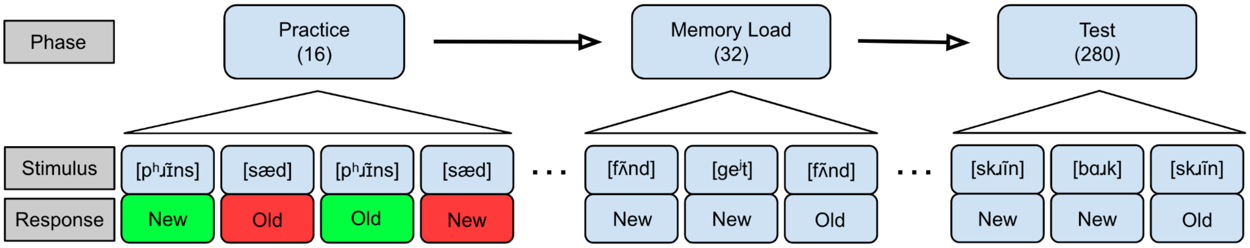

Fig. 2.1: Structure of the experiment. Schematic diagram of the structure of the experiment, showing phase, sample stimuli, and sample responses. Participants received feedback on their responses only during practice. Correct practice responses are

shown in green and incorrect responses are shown in red. From left to right, the responses shown in the practice phase (had they been given in the test phase and thus included in the analysis) would be coded as a Correct Rejection, Miss, Hit, and False Alarm. Procedures for the memory load and test phases were identical for participants, but data from the Memory load phase were not included in analysis.

2.1.3 Procedure

The experiment was conducted online through participants' web browsers. The procedure was coded in JavaScript, drawing heavily from the jsPsych library (de Leeuw, 2015). Before beginning the main experiment, participants completed an audio check inspired by Woods, Siegel, Traer, & McDermott (2017) to ensure that they were wearing headphones. After the audio check, participants completed the practice round, memory load phase, and test phase. In all phases, each trial consisted of an auditory stimulus (i.e., a single spoken word), to which participants responded "New" or "Old," depending on whether they believed it was their first time hearing the word or whether they believed they had heard it already. In the memory load and test phases, there was a 1-second silent interval between the time of response and the onset of the subsequent stimulus. It was not possible to provide a response before the stimulus had played fully. If no response was made within 4 seconds of the stimulus offset, the subsequent trial began automatically.

2.2 Results & Discussion

For full model specifications and summaries, see Appendix A. Models reported in that section are numbered 1.1–1.4 and reflect analyses of the dependent variables of correct Old responses on repeated items (**Hit**s, Model 1.1) based on **TokenCondition** (**Novel** vs. **Identical**), **TalkerCondition** (**M1** vs. **M2**), number of intervening trials (**Lag**, 1–65, mean-centered and re-scaled), and **TrialNumber** (1–280, mean-centered and re-scaled); log-transformed response latency on Hits (**logRT**, Model 1.2) by TokenCondition, TalkerCondition, Lag, TrialNumber, and the token's **Duration** (measured in ms, mean-centered and rescaled); incorrect Old responses on New items (*false alarms*, **FA**s, Model 1.3) based on TokenCondition, TalkerCondition, and TrialNumber; and z-transformed Hit-rate minus FA-rate (**d'**, Model 1.4) based on TokenCondition and Talker. RTs more than 2 standard deviations from the mean were discarded (approximately 4% of data). Interactions were included for all fixed effects except the controls of TrialNumber and Duration, where relevant. Categorical independent variables were treatment coded with reference levels of **Novel** for TokenCondition and **M1** for TalkerCondition. Treatment as opposed to sum coding was selected in order to facilitate direct comparisons across levels without making reference to grand means, which would not be relevant under the present design. Using treatment coding allowed us to replicate the analyses of traditional approaches, which used only identical tokens, while also showing how the results of the analysis may differ based on talker and token type. For more information on treatment coding and sum coding, see Brehm & Alday (2022). Post-hoc estimated marginal mean (EMM) tests were still necessary to draw all relevant comparisons and were conducted using the *emmeans* package in R (Russell, 2022).

All models were mixed-effects regression models fitted in R using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) except the model for d', which was fitted as a fixed-effects linear regression model using R's built-in *lm* function[2]. For mixed-effects models, p-values were obtained via Satterthwaite's method using the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017). Random effects structures were determined by beginning with the maximal structure and removing slopes and intercepts until the model converged (Barr, Levy, Scheepers, & Tily, 2013). The maximal random effects structures were determined uniquely for each model and included random intercepts by Item and by Subject along with random slopes by each within-unit variable. If the maximal model did not converge, the random slope associated with the lowest variance was removed. The same procedure continued until the model converged. Fixed effects were determined in advance and not pruned in any way. Full model specifications, including final random effects structures, are reported in Appendix A.

We hypothesized that if recognition of repeated words was facilitated by identical tokens, performance would be stronger in the Identical condition than in the Novel condition. Stronger performance would be characterized by higher Hit rates, faster RTs, lower FA rates, and higher D' in the Identical than in the Novel condition. We also tested whether results would be asymmetrical across the two Talker conditions, M1 and M2, though we did not explicitly predict the directions of these effects, which is beyond the scope of the current study.

[2] All data and code, including models with maximal random effects structures specified, are available on Mendeley Data: https://data.mendeley.com/datasets/cgrphxc976/2

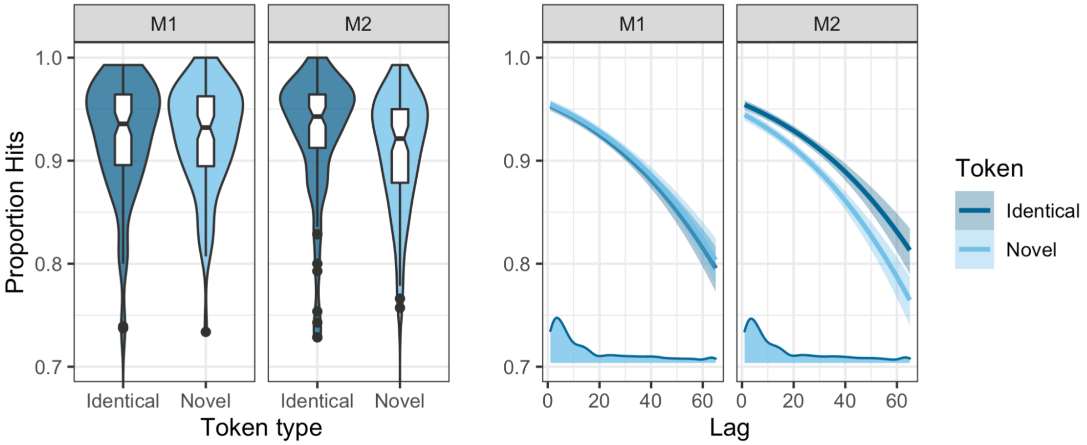

Fig. 2.2: Accuracy by talker and token type. Left: Proportions of hits based on token condition (Identical vs. Novel) and talker condition (M1 vs. M2) with each point representing a participant's mean. Right: Proportions of hits across lags with smoothers drawn by a generalized linear model. Densities at the bottom of the plots represent the number of repetitions at each Lag. Note that plots are based on raw data rather than model predictions.

Hit rates are visualized in Fig. 2.2. Accuracy decreased as Lag increased (Model 1.1: β = –1.78, SE = 0.13, z = –14.18, p < 0.001), as is visible in Fig. 2.2. Accuracy increased as TrialNumber increased (Model 1.1: β = 0.50, SE = 0.061, z = –8.09, p < 0.001). For the reference level of M1, there was not a statistically significant difference in accuracy rates between participants in the Novel and Identical conditions (Model 1.1: β = –0.0048, SE = 0.12, z = –0.04, p > 0.1). There was also no effect of Talker: Within the reference level of Novel, there was no difference between M1 and M2 (Model 1.1: β = –0.20, SE = 0.13, z = –1.60, p > 0.1). Although neither of these simple effects was significant alone, we observed a marginal interaction between TokenCondition and Talker (Model 1.1: β = 0.30, SE = 0.17, z = 1.75, p = 0.08), suggesting that accuracy may have been higher when TokenCondition was Identical and the Talker was M2 than the cumulative effects of these variables would otherwise suggest. This possibility was partially supported by a marginal effect in the post-hoc test showing that accuracy may have been higher for M2 in the Identical than in the Novel TokenCondition (β = –0.30, SE = 0.13, z = –2.39, p = 0.08).

Latency data were analyzed for correct responses on Old trials. As expected, responses slowed as the number of intervening trials increased (Model 1.2: β = 0.12, SE = 0.009, t = 14.00, p < 0.001). Responses were also faster as TrialNumber increased (Model 1.2: β = –0.054, SE = 0.003, t = –18.16, p < 0.001). Faster response times were associated with durationally longer stimuli (Model 1.2: β = –0.32, *SE* = 0.01, *t* = –26.70, *p* < 0.001). There was no simple effect of Talker or TokenCondition, and neither was present in interactions.

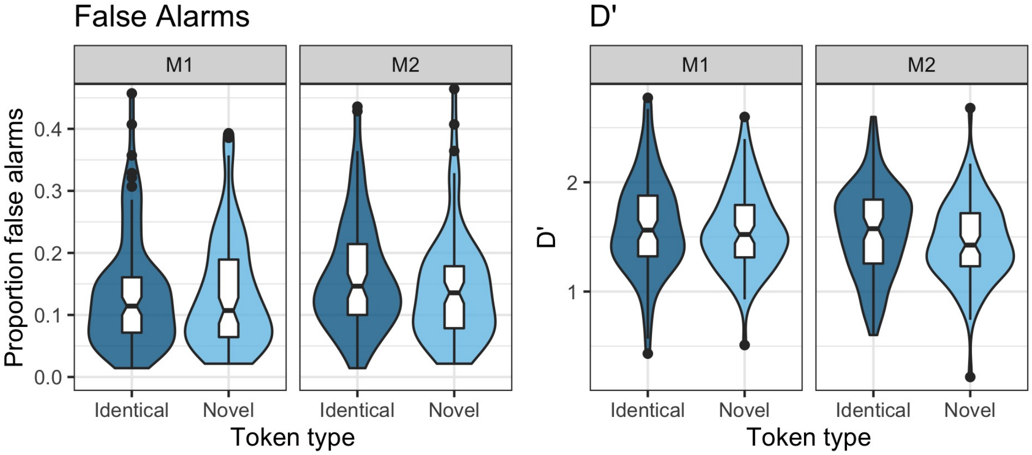

Fig. 2.3: False alarms and d'. Left: Proportions of FAs (Old responses to New items) based on TokenCondition (Identical vs. Novel) and Talker (M1 vs. M2) with each data point reflecting a participant's mean. Right: D' values (z-transformed FAs subtracted from Hits, calculated at the participant level) based on TokenCondition and Talker.

The analysis of FAs (incorrect Old responses on New items) demonstrated that listeners produced more FAs as the experiment progressed (Model 1.3: β = 1.23, *SE* = 0.05, *z* = 25.06, *p* < 0.001), as demonstrated by a main effect of TrialNumber. No simple effect was observed for either TokenCondition or Talker, and neither were involved in interactions, but pairwise comparisons revealed one pattern based on Talker. Within the Identical condition, listeners produced significantly more FAs when hearing M2 than when hearing M1 (β = –0.34, *SE* = 1.31, *z* = –2.61, *p* < 0.05). This was the only significant pairwise comparison and is visible in Fig. 2.3 (left).

Lastly, the analysis of D' (Fig. 2.3, right) showed that listeners in the Novel condition (reference level) were more sensitive to words produced by M1 than words produced by M2 (Model 1.4: β = –0.23, *SE* = 0.11, *t* = –2.15, *p* < 0.05). While we did not observe a significant simple effect of TokenCondition within the reference level of M1 (Model 1.4: β = –0.087, *SE* = 0.10, *t* = –0.83, *p* > 0.1), we did observe an interaction between TokenCondition and Talker (Model 1.4: β = 0.32, *SE* = 0.15, *t* = 2.12, *p* < 0.05). Given the negative coefficient associated with the effect of Talker and the null effect associated with TokenCondition, this interaction suggests that sensitivity in the M2, Identical condition is higher than would be expected based on the independent effects.

The primary purpose of Exp. 1 was to test whether recognition of words produced by a single talker would be inhibited when repetitions contained natural phonetic variability, and if so whether the memory effects are asymmetrical between talkers. Few effects of TokenCondition were found, particularly for listeners hearing M1, whose Hit rates, RTs, FA rates, and D' were all stable between the Identical and Novel conditions. However, there were several notable effects of Talker, and interactions between Talker and TokenCondition, suggesting that this stability is not broadly generalizable. Crucially, the analysis of D' showed greater sensitivity among listeners hearing M1 than M2, as well as an interaction between TokenCondition and Talker, indicating that the repetition type bore a stronger influence on responses to M2 than to M1. Within the Identical condition, listeners who responded to M2's voice produced more FAs than those hearing M1's voice. The analysis of Hits showed a marginal effect of TokenCondition among listeners in the M2 condition. Although the influence of repetition type was subtle in Exp. 1, these effects are enough to suggest that variation in listener behaviors is patterned to some extent by the combination of talker and token repetition type and provide a baseline for subsequent comparisons in more complex contexts. Exp. 2 expands on these findings by examining the roles of token type and talker in a context where it is possible for classic talker-specificity effects to emerge, i.e. a context where both talkers are heard together.

3. Experiment 2

Exp. 2 has two main goals. One goal of Exp. 2 is to test whether talker-specificity effects are replicable when listeners hear novel as opposed to identical token repetitions on same-talker trials. In cases where voice information—as opposed to solely token-level acoustic information—is encoded, it would be expected that even with novel-token repetitions, listeners would be faster and more accurate in responding to words repeated in the same voice than in a different voice. The other goal is to test whether these effects are equivalent for the two talkers. Our hypothesis was that there would be significant differences in response patterns based on the talker heard at each word's first presentation.

3.1 Methods

3.1.1 Participants

A total of 462 participants completed the experiment after recruitment through Prolific. All participants provided informed consent. The same exclusion criteria were applied to Exp. 2 as Exp. 1: participants were monolingually English-speaking American nationals living in the United States who had reported no hearing loss. Responses from 87 participants were removed from analysis because they either had a miss rate of over 10% on trials where a token was repeated with a lag of either 1 or 2 trials, or failed to respond to at least 85% of trials, leading to a sample size of 375. Power analysis was conducted in the same manner as in Exp. 1. Participants were again compensated $3.77, and took an average of 15.46 minutes to complete the experiment.

3.1.2 Stimuli & Design

Stimuli. The words and tokens used as stimuli in Exp. 2 were identical to those used in Exp. 1. The experiment was again coded in JavaScript using the jsPsych library.

Design. The design of Exp. 2 was nearly identical to that of Exp. 1 with the exception that all participants heard both talkers rather than one talker individually. Thus, TokenCondition (Identical vs. Novel) was the only between-subjects variable.

The words in each phase had a 50% chance of being heard in the voice of each talker. In other words, of the 140 first presentations heard in the test phase, 70 were produced by M1 and 70 were produced by M2. Among the repeated words, 50% were produced by the same talker as the first presentation and 50% were produced by the other talker. For example, among repetitions of the 70 words initially produced by M1, 35 were repeated by M1 and 35 were repeated by M2. In the identical token condition, all words repeated in the same voice as the initial presentation were identical recordings to the initial presentation. In the novel token condition, all repeated words in the same voice as the initial presentation consisted of previously unheard audio. The exact distribution of words and talkers was generated randomly and uniquely for each participant in-browser using an algorithm written in JavaScript. There were again 8 practice word pairs, 16 memory load word pairs, and 140 test word pairs. Here, *word pair* refers to each individual stimulus word's Old and New presentations. There were 164 total word pairs, and therefore 328 total trials. Lags again ranged from 1 to 65 with approximately the same distribution as in Exp. 1.

3.1.3 Procedure

The procedure of Exp. 2 was identical to that of Exp. 1. The response keys, the number of trials in each phase, and the audio check procedure were all the same.

3.2 Results & Discussion

Mixed-effects regression models were again fitted in R using the *lme4* package (Bates et al., 2015). Random effects structures were determined in the same manner as in Exp. 1. For full model specifications and summaries, see Appendix B. Models in that section are numbered 2.1–2.7. These models were fitted to the proportion of correct Old responses on both talker's stimuli pooled (**Hit**s, Model 2.1), based on **TokenCondition** (**Novel** vs. **Identical**), the voice of the repetition (**RepVoice**, **DIFF** vs. **SAME**), **Lag** (1–65, mean-centered and rescaled), and **TrialNumber** (1–280, mean-centered and rescaled); and the proportion of Hits with words first presented in each talker's voice separated (Model 2.2) based on TokenCondition, talker heard on first presentation (**FirstTalker**, **M1** vs. **M2**), talker heard at second presentation (**SecondTalker**, **M1** vs. **M2**), and Lag. Models with the same fixed-effects structures and the additional control of **Duration** (measured in ms, mean-centered and rescaled) were fitted to **logRT** with the talkers pooled (Model 2.3) and separated (Model 2.4). As in Exp. 1, RT was measured from the offset of each stimulus. Models were also fitted to the proportion of incorrect Old responses on New words (*false alarms*, **FA**s, Model 2.5) based on TokenCondition, **Talker** (**M1** vs. **M2**), and TrialNumber, to **D'** (Model 2.6) based on TokenCondition, and to **By-Talker D'** (Model 2.7) based on TokenCondition and Talker. By-Talker D' was calculated using Hits on words heard in each talker's voice at the first presentation. As in Exp. 1, categorical independent variables were treatment coded with reference levels of **Novel** for TokenCondition, **DIFF** for RepVoice, and **M1** for FirstTalker, SecondTalker, and Talker. Models with a dependent variable of Hits or FAs were fitted using binomial logistic regression and models with a dependent variable of logRT or d' were fitted using linear regression. Interaction terms were included for all fixed effects except the controls of TrialNumber and Duration. Given the large number of fixed effects, higher order interactions involving Lag were excluded from Models 2.2 and 2.4.

3.2.1 Replication of talker-specificity effects

We first asked whether talker-specificity effects were observable in the context of novel token repetitions. In the Novel condition (the reference level), listeners were more accurate in correctly recognizing Old words when the word was repeated in the SAME voice than in a DIFF voice (Model 2.1: $\beta = 0.12$, $SE = 0.05$, $z = 2.18$, $p < 0.001$), as visualized in Fig. 3.1 (left). Pairwise comparisons revealed that the same was true in the Identical condition ($\beta = –0.34$, $SE = 0.56$, $z = –6.11$, $p < 0.001$), replicating the classic talker-specificity effect. As expected, listeners were decreasingly accurate as Lag increased (Model 2.1: $\beta = –1.75$, $SE = 0.11$, $z = –16.42$, $p < 0.001$), and more accurate as TrialNumber increased (Model 2.1: $\beta = 0.28$, $SE = 0.053$, $z = 5.28$, $p < 0.001$).

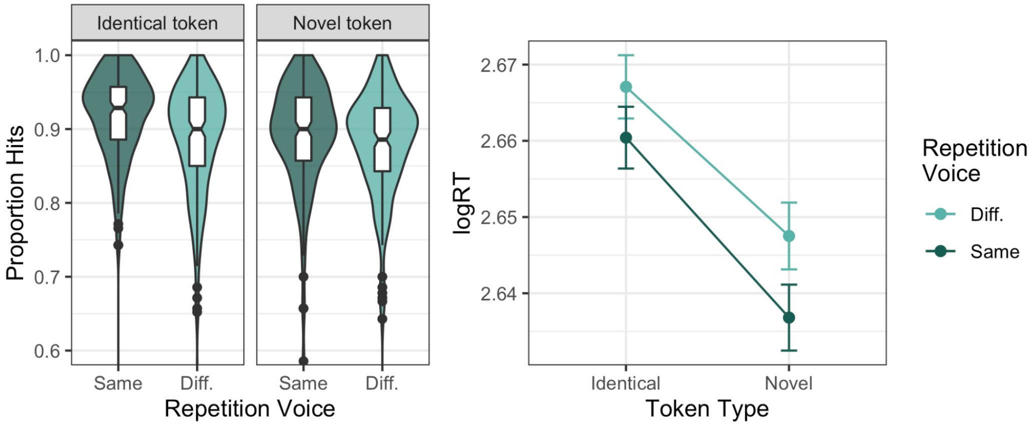

Fig. 3.1: Accuracy and latency by RepVoice and token type. Left: Proportions of hits based on TokenCondition (Identical vs. Novel) and RepVoice (SAME vs. DIFF) with each point representing a participant's mean. Right: Mean logRT values based on TokenCondition (Identical vs. Novel) and RepVoice (Same vs. Diff) with 95% CIs. For reference, logRT values of 2.64 and 2.67 are equivalent to 437 ms and 468 ms respectively. Note that both plots are based on raw data rather than model estimates.

Latency data further confirmed that talker-specificity effects were observable in the context of Novel tokens (Fig. 3.1, right). In the Novel condition (reference level), responses were faster on SAME trials than on DIFF trials (Model 2.3: $\beta = –0.010$, $SE = 0.002$, $t = –4.085$, $p < 0.001$). A post-hoc EMM test showed that this effect was also significant in the Identical condition ($\beta = 0.0067$, $SE = 0.0024$, $t = 2.96$, $p < 0.05$), replicating the classic talker-specificity effect. As Lag increased, responses on DIFF trials (the reference level) slowed (Model 2.3: $\beta = 0.12$, $SE = 0.0078$, $t = 15.84$, $p < 0.001$). Responses were faster at higher than at lower TrialNumbers (Model 2.3: $\beta = –0.037$, $SE = 0.0030$, $t = –12.54$, $p < 0.001$) and faster when tokens were longer than when they were shorter (Model 2.3: $\beta = –0.34$, $SE = 0.0061$, $t = –54.86$, $p < 0.001$).

Some effects suggested differences across TokenConditions. While responses were only marginally more accurate in the Identical than in the Novel condition within the reference level of DIFF (Model 2.1: $\beta = 0.14$, $SE = 0.085$, $z = 1.67$, $p = 0.1$), there was an interaction between TokenCondition and RepVoice (Model 2.1: $\beta = 0.22$, $SE = 0.067$, $z = 3.37$, $p < 0.001$), suggesting that on SAME repetitions in the Identical condition, listeners were more accurate than would be expected based on the independent effects of TokenCondition and RepVoice. The post-hoc test confirmed that on SAME trials, responses were more accurate in the Identical than in the Novel condition ($\beta = –0.37$, $SE = 0.089$, $z = –4.15$, $p < 0.01$). We did not observe effects of TokenCondition in the latency data. However, the analysis of d' showed that when accounting for decision bias, listeners were more discriminate in the Identical condition than in the Novel condition (Model 2.6: $\beta = 0.11$, $SE = 0.037$, $t = 3.01$, $p < 0.01$).

Taken together, these effects suggest that talker-specificity effects are observable even when listeners are unable to match precise acoustic patterns. In the Novel condition, listeners were faster and more accurate to recognize Old words when they were repeated in the same voice than when they were repeated in the alternate voice, suggesting that this group of effects surfaces not only through the precise repetition of a distinct acoustic event but also through the encoding of higher level talker voice characteristics. However, we also observed effects based on TokenCondition, suggesting that recognition may be bolstered by the repetition of an exact phonetic match. The difference in accuracy between SAME and DIFF trials was greater when listeners heard identical tokens than when they heard novel tokens. Additionally, overall discriminability was higher in the Identical than in the Novel condition. We explore these effects further via analyses of talker-based asymmetries.

3.2.2 Asymmetric talker encoding

Our second set of analyses in Exp. 2 aimed to investigate the hypothesis that encoding and recall of the two talkers' voices would be asymmetrical. We hypothesized that accuracy and latency would differ based on the talker heard at each word's first presentation. The remainder of this section uses a shorthand convention where, for example, M1-M2 refers to trials where the first presentation was heard in the voice of M1 and the repetition was heard in the voice of M2. Thus M1-M1 and M2-M2 refer to SAME trials.

We found substantial differences in accuracy based on the talkers (Fig. 3.2). Relative to M1-M1 trials in the Novel condition (the reference level), participants produced fewer correct responses on M1-M2 trials (Model 2.2: β = –0.23, *SE* = 0.070, z = –3.36, $p < 0.001$) and on M2-M1 trials (Model 2.2: β = –0.65, *SE* = 0.073, z = –9.050, $p < 0.001$), reflecting the across-the-board lower accuracy on DIFF trials relative to SAME trials described above. However, the post-hoc EMM test revealed these two categories of DIFF trials were not equivalent. The order of talkers influenced responses significantly in the Novel condition, with significantly higher Hit rates for M1-M2 than for M2-M1 (β = –0.42, *SE* = 0.073, z = –5.80, $p <$ 0.001). The same relationship was found in the Identical condition (β = –0.43, *SE* = 0.074, z = –5.85, $p < 0.001$). This indicates that words first presented in the voice of M1 were remembered better than words first presented in the voice of M2. The decrease in accuracy relative to M1-M1 was less pronounced for the SAME trials M2-M2 than the combined effects of DIFF trials M1-M2 and M2-M1 would suggest, as indicated by a significant interaction between FirstTalker and SecondTalker (Model 2.2: β = 0.34, *SE* = 0.084, z = 4.03, $p < 0.001$). However, a post-hoc EMM test confirmed that participants were significantly less accurate on M2-M2 trials than M1-M1 trials in the Novel condition (β = 0.55, *SE* = 0.079, z = 6.95, $p < 0.001$), indicating asymmetrical accuracy on the two types of SAME trial within the Novel condition. As in previous models, response accuracy decreased as lag increased (Model 2.2: β = –1.89, *SE* = 0.12, z = –16.46, $p < 0.001$) and increased as TrialNumber increased (Model 2.2: β = 0.28, *SE* = 0.084, z = 4.03, $p < 0.001$).

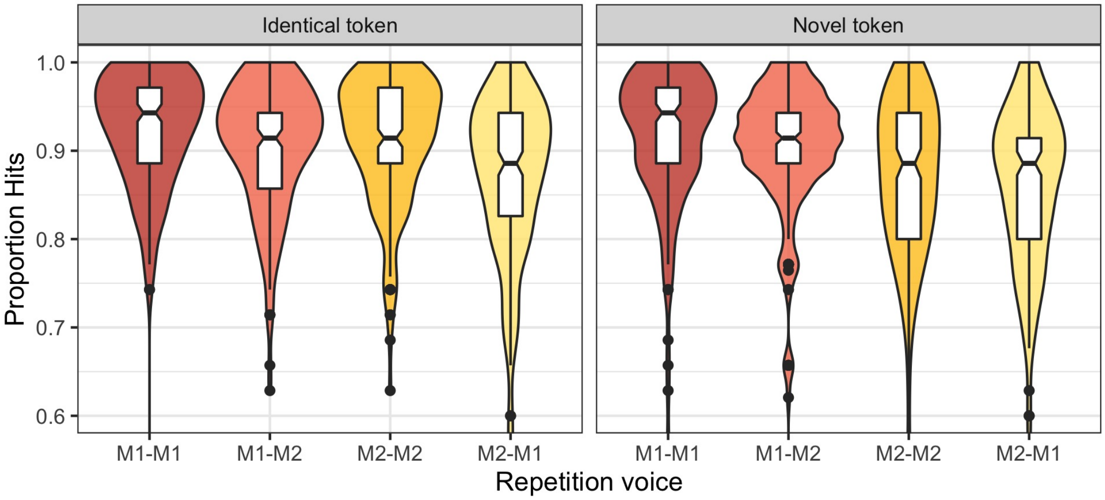

Fig. 3.2: Accuracy by talker and token type. Left: Plots of proportions of Hits (correct Old responses) based on token condition (Identical vs. Novel) and voice of each presentation on SAME trials. The talker label refers to the talker heard at both the first and second presentation. Right: Plots of proportions of hits on DIFF trials. The voice order naming schema shows the voice of the initial presentation separated by a hyphen from the voice of the repetition. For example, the group M1-M2 includes all trials where the word was first heard in the voice of M1 and repeated in the voice of M2.

Latency data from the Novel condition were largely consistent with accuracy data. Relative to M1-M1 trials, listeners were slower to respond on M1-M2 trials (Model 2.4: β = 0.030, *SE* = 0.0053, *t* = 5.65, *p* < 0.001) as well as on M2-M1 trials (Model 2.4: β = 0.018, *SE* = 0.0037, *t* = 4.73, *p* < 0.001), reflecting the previously described RT cost on DIFF trials relative to SAME trials. The increase in RT was less pronounced on M2-M2 trials, as indicated by an interaction between FirstTalker and SecondTalker (Model 2.4: β = –0.020, *SE* = 0.0049, *t* = –4.08, *p* < 0.001), but the EMM test indicated that responses were slower on M2-M2 trials than on M1-M1 trials (β = –0.028, *SE* = 0.0054, *t* = –5.12, *p* < 0.001). As observed previously, responses were slowed as Lag increased (Model 2.4: β = 0.12, *SE* = 0.0077, *t* = 15.82, *p* < 0.001), and sped as TrialNumber increased (Model 2.4: β = –0.037, *SE* = 0.0030, *t* = –12.62, *p* < 0.001), and as Duration increased (Model 2.4: β = –0.30, *SE* = 0.012, *t* = –25.037, *p* < 0.001).

Talker-based asymmetries were also evident in the FA data (Fig. 3.3, left) and in the By-Talker D' data (Fig. 3.3, right). Listeners had higher error rates for New items in the Novel condition (reference level) when responding to M2 than to M1 (Model 2.5: β = 0.29, *SE* = 0.051, *z* = 5.78, *p* < 0.001). The EMM test revealed that the same was true in the Identical condition (Model 2.5: β = –0.14, *SE* = 0.050, *z* = –2.88, *p* < 0.05). They were also more likely to produce a

FA as the experiment progressed (Model 2.5: $\beta = 1.24$, $SE = 0.046$, $z = 26.87$, $p < 0.001$). The analysis of By-Talker D' revealed that listeners in the Novel condition showed greater sensitivity to words first presented in the voice of M1 compared to M2 (Model 2.7: $\beta = –0.26$, $SE = 0.026$, $t = –9.96$, $p < 0.001$). Similarly, we observed higher d' values for M1 than M2 in the Identical condition in the EMM test ($\beta = 0.15$, $SE = 0.026$, $t = 6.08$, $p < 0.001$).

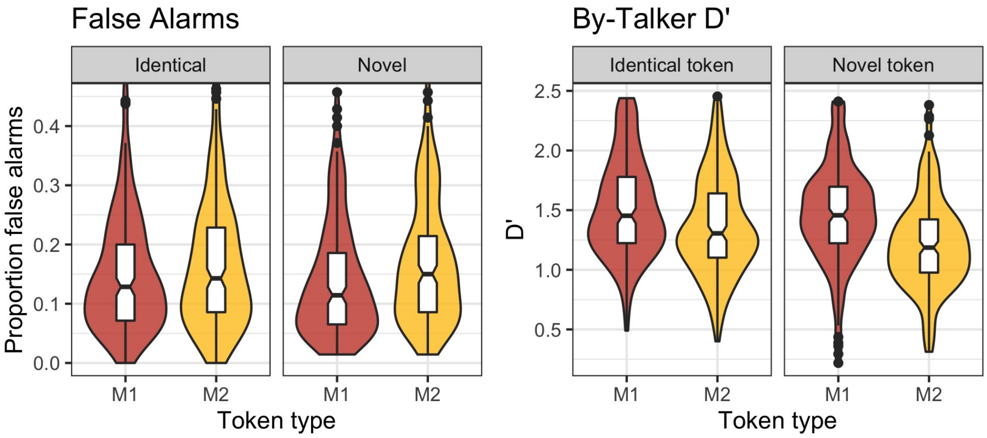

Fig. 3.3: False alarms and d' by talker and token type. Left: Proportion of FAs (incorrect Old responses on New trials) based on Talker and TokenCondition. Right: Overall sensitivity or D' (Hits – FAs) calculated at the talker level.

We also found substantial evidence that TokenCondition influenced responses to the talkers asymmetrically. On M1-M1 trials (reference level), listeners in the Identical condition had only marginally higher Hit rates than listeners in the Novel condition (Model 2.2: $\beta = 0.18$, $SE = 0.10$, $z = 1.78$, $p = 0.08$), but the EMM test showed significantly higher accuracy on M2-M2 in the Identical than in the Novel condition (Model 2.2: $\beta = –0.56$, $SE = 0.093$, $z = –6.01$, $p < 0.001$). This asymmetry is further supported by a three-way interaction between TokenCondition, FirstTalker, and SecondTalker (Model 2.2: $\beta = 0.42$, $SE = 0.12$, $z = 3.50$, $p < 0.001$). This is clearly visible in Fig. 3.2 (left), when comparing responses on M1-M1 trials across conditions to M2-M2 trials across conditions. A post-hoc test also showed that listeners were not measurably more accurate on M2-M2 trials than M2-M1 trials in the Novel condition ($p > 0.1$), indicating that in the absence of a precise physical match between tokens, the specificity effect was entirely dependent on the voice heard at encoding (appearing for M1 but not for M2).

We also observed asymmetries based on TokenCondition in the latency data. The latency data showed that listeners were slightly slower to respond to M1-M1 trials in the Identical condition than in the Novel condition (Model 2.4: $\beta = 0.028$, $SE = 0.014$, $t = 2.03$, $p < 0.05$). The

RT increase associated with M1-M2 trials relative to M1-M1 was less pronounced in the Identical condition than in the Novel condition, as indicated by a significant interaction between TokenCondition and SecondTalker (Model 2.4: $\beta = -0.016$, $SE = 0.0054$, $t = -2.94$, $p < 0.01$).

Token-based asymmetries across talkers were also present in the FA data, although the simple effect of TokenCondition was not significant in the reference level of M1, Novel (Model 2.5: $\beta = 0.12$, $SE = 0.093$, $z = 1.30$, $p > 0.1$). M2 trials were characterized by more FAs than M1 trials in both conditions, but to a lesser extent in the Identical condition than in the Novel condition, as demonstrated by an interaction between TokenCondition and Talker (Model 2.5: $\beta = -0.15$, $SE = 0.0062$, $z = -2.431$, $p < 0.05$). The D' data showed a similar pattern, where responses to M2 were characterized by less of a decrease relative to M1 in the Identical condition relative to the Novel condition, as demonstrated by an interaction between TokenCondition and Talker (Model 2.7: $\beta = 0.10$, $SE = 0.036$, $t = 2.85$, $p < 0.01$). A post-hoc EMM test confirmed that D' for M2's tokens was significantly higher in the Identical condition than in the Novel condition ($\beta = -0.16$, $SE = 0.42$, $t = -3.82$, $p < 0.001$).

These patterns of results strongly suggest that encoding and recall of spoken words is not symmetrical across talkers. Responses differed substantially even to two talkers matched along macro-demographic lines. Words spoken by M1 were encoded more strongly and recognized more accurately than words spoken by M2, as indicated by the higher rate of Hits, faster RTs, lower FA rates, and higher D' values. Crucially, the influence of token type was also asymmetrical across talkers. While responses to M1's stimuli were relatively stable in the Novel and Identical conditions, responses to M2's stimuli were generally characterized by stronger performance (higher Hits, higher D') in the Identical than in the Novel condition. The better recognition of Old items first presented in the voice of M1 than M2 in the Novel condition may indicate that high-level voice information was better encoded for M1 than for M2. This pattern suggests that the stronger performance on SAME trials in the Identical as opposed to Novel condition when talkers were pooled (described in the previous section, 3.2.1) must have been driven largely by the decreased sensitivity to M2's stimuli in the Novel condition rather than an across-the-board decrease in sensitivity when same-talker trials were repeated with novel tokens.

4. General Discussion

In this study, we compared recognition of identically repeated tokens to recognition of novel tokens within and across talkers. By including word repetitions that contained the natural phonetic variability present in a talker's voice (novel tokens), we tested whether listeners used high-level voice characteristics to facilitate recognition. Aside from informing our understanding of what kinds of acoustic information are stored in memory traces of spoken words, this allowed us to investigate the possibility that previously observed talker-specificity effects arose from an experimental confound where same-talker repetitions were presented with identical tokens. We also investigated data at the talker level, a mode of analysis which has previously been absent in recognition memory studies, even as there is reason to believe that encoding and recall of spoken words is not consistent across talkers. We found that memory for identical tokens was in some cases stronger than memory for novel tokens, although this effect was dependent on the talker. Although specificity effects were replicated with novel token repetitions, an analysis of talker-level patterns revealed substantial richness in the data not captured by that generalization. We argue that these asymmetries were the result of variable encoding strengths of each voice (i.e., listeners encoded one voice more strongly than the other). We also argue that examining data at the talker level is a crucial step in speech research, noting that the conclusions of this study would have been very different if we had only analyzed the pooled data.

One important finding was that talker-specificity effects were replicated when same-talker repetitions were presented with non-identical tokens. Even when these repetitions consisted of tokens that listeners had not previously heard in the procedure, responses were faster and more accurate when the voice was the same as the first presentation. This finding suggests that previously observed talker-specificity effects are unlikely to have been purely the result of the widespread experimental confound of using same-talker, identical-token repetition trials. This means that to at least some extent, listeners must be capable of encoding higher level voice characteristics and using those encodings actively in word recognition. However, rich and nuanced by-talker effects emerged that add nuance to this interpretation.

Although classic talker-specificity effects were replicated, the results of the two experiments together gesture toward a system where recognition is facilitated by identical token matching, but only in some cases and in a way that is largely contingent on the talker. While effects of TokenCondition were observed in both experiments, these were largely overshadowed by effects of Talker and often came in the form of interactions with Talker. For M1, results were largely consistent across TokenConditions in both experiments. In Exp. 1, listeners hearing M1 performed equally well in the Novel and Identical conditions, but this level of consistency was not observed for listeners hearing M2, whose recognition performance was somewhat poorer in the Novel condition than in the Identical condition. This suggests that listeners hearing M2 were more dependent on the precise repetition of the physical stimulus than were listeners hearing M1.

A similar but more robust pattern of results was observed in Exp. 2, with TokenCondition having a more substantial influence on responses for M2 than for M1. While the only observed benefit of identical repetitions for M1 came in the form of marginally higher accuracy on M1-M1 trials in the Identical than in the Novel condition, benefits of Identical repetitions were widespread for M2. In the Identical condition relative to the Novel condition, Hit rates on M2-M2 trials were higher, false alarm rates on M2 trials were lower, and d' was higher for repetitions of words first presented in M2's voice. As in Exp. 1, performance benefited more from the repetition of identical tokens when M2 rather than M1 was the talker in question. Given that humans have been demonstrated to retain even meaningless acoustic information quite well (e.g. Viswanathan et al., 2016; Winkler et al., 2002), it is logical that listeners would recruit this information for word recognition when relevant. It follows that when listeners jointly encode word and high-level voice information well (as in the case of M1), fine-grained acoustic information does not improve performance, but when voice information is more weakly encoded (as in the case of M2), this otherwise meaningless acoustic information becomes advantageous, leading to a relative boost in performance.

Importantly for this account, not all talker-level asymmetries were contingent on the type of token heard at repetition. This was particularly evident in the analysis of D' in Exp. 2, which showed that even within the Identical condition, participants were more accurate when the talker was M1 than when the talker was M2. The analysis of Hits showed that on different-talker trials in the Identical condition, participants were more accurate when the talker order was M1-M2 than when it was M2-M1. This pattern indicates that encoding may have been overall stronger for words spoken by M1 than words spoken by M2. While the current study was not designed to tease apart the specific factors that lead to and condition asymmetric encoding and recognition of voices, we explore several possibilities here.

One account may appeal to the typicality of each voice. Some past work has demonstrated that atypical forms draw more attention at encoding, leading to easier access at recall. This effect has been found for speech rates (Nygaard, Burt, & Queen, 2000), typical and atypical male and female voices (Johnson, 2006), and typical and atypical phonological forms (Sumner & Samuel, 2005). The account of the present results appealing to typicality would need to demonstrate that M2's voice is more typical than M1's voice, leading to weaker episodic encoding and therefore worse recall. (Although defining and quantifying typicality in this case may prove difficult.)

A second explanation may appeal to the ideological idealization of voices. Sumner & Kataoka (2013) found in a false-memory study that among General American accent (GA) listeners, New York City accents (NYC) and Southern Standard British English accents (BE) were represented asymmetrically in memory, even though the two variants are in principle equally atypical for GA listeners. False recall was higher for NYC talkers than for BE talkers, suggesting that the BE voices were somehow privileged in encoding. This increased allocation of cognitive resources to BE rather than NYC talkers may be the result of BE having a more idealized status than NYC among many GA listeners. In the present study, if responses to M1's voice were stronger than responses to M2's it may be the case that M1's voice is in some way more idealized than M2's. However, there is currently a wide range of possible explanations for the asymmetries seen in encoding of the two voices used in the present experiments, and it is not the aim of this study to disentangle the source of these asymmetries. Rather, our purpose is to demonstrate that encoding is not equivalent across talkers and to encourage researchers to tease apart the factors that may contribute to variable degrees of encoding. Future studies will need to target specific sources of variance directly. Research in this area could prove fruitful in contributing to understandings of how social phenomena emerge in speech perception, how idiolects emerge for individuals, how the veridicality of ear-witness testimonies varies depending on the talkers in question, and so on.

The fact that high-level voice characteristics seem to have been encoded well for listeners hearing M1 leads to questions about the nature of memory traces of voices as something separable from the specific tokens in which they were encountered. It has been a fruitful line of research in auditory cognitive science and cognitive neuroscience to investigate how humans are able to associate voices with individuals (e.g., Lavan, Burston, & Garrido, 2019; van Lancker, Kreiman, & Cummings, 1989; von Kriegstein & Giraud, 2006; Winters, Levi, & Pisoni, 2008). However, identifying specific characteristics of speech that listeners use to associate voices with individuals has been difficult. For example, van Dommelen (1990) found that f0 was treated as a useful cue to talker identity only when talkers had unusually high or low pitch averages, suggesting that the roles of acoustic parameters in identifying speakers by voice are not hierarchically fixed. In an fMRI study, Hasan, Valdes-Sosa, Gross, & Belin (2016) found cross-classification of facial and vocal information in several areas of the temporal lobe, suggesting that voice information is part of a larger identity-classification mechanism, where voice characteristics are entangled or interlocking with other sources of individual differentiation. Much of this recognition system must be divorced to at least some extent from the types of acoustic regularities that are present only in identical tokens, but much work remains to be done to better understand memory representations of individual voices, how these representations are active in speech perception, and why they are apparently encoded more strongly for some voices than for others.

Another important point regarding the observed asymmetries relates to the degree to which we observed talker-based asymmetries in each experiment. Notably, these effects were subtle in Exp. 1, but quite robust in Exp. 2. When M1 and M2 were heard in isolation, response patterns barely differed across talkers, but when the two talkers were heard together, the differences increased substantially. We posit that context strongly conditions the degree to which listeners encode the voices of particular talkers. When the talkers were heard separately, listeners allocated all available resources to the task at hand, but when the talkers were heard together, listeners may have (subconsciously) allocated more cognitive resources to the recognition and encoding of tokens spoken by M1 than M2. This observation is consistent with a rich history of context effects in speech perception research (Barreda, 2012; Sommers, Nygaard, & Pisoni, 1994). Kim and Sumner (2015), for example conducted two experiments probing the extent to which emotionally neutral words (e.g., *pineapple*) uttered with phonetically-cued emotional information prime semantically related words. Crucially, in Exp. 1, each listener heard only one type of emotional prosody for all words (*happy*/*neutral*/*angry*), and in Exp. 2, participants heard a mix of prosody types (*happy & neutral*/*angry & neutral*). When neutral primes were presented alone in Exp. 1, they successfully primed semantically related targets, but when they occurred in a mixed context in Exp. 2, the neutral primes did not prime semantically related targets. The authors posit that this asymmetry emerged as a result of the additional attention allocated to emotionally uttered primes relative to neutral primes when the two types co-occur. This schema may be analogous to our own observations in the present study, where M1's voice may have drawn additional attention primarily in the context of M2's voice.

A final point relating to the observed asymmetries in responses based on talkers is that this level of analysis has not been common in speech perception, but had it not been conducted in this case, the study's conclusions would have been very different. This is clear from the outcomes of the two strategies employed for modeling the results of Exp. 2. One set of models followed the traditional approach for analysis by pooling results for both talkers while the other set included predictors reflecting the talker heard in each trial. If our analysis had consisted solely of the models with pooled talkers, the conclusion would have been that performance is characterized by a tripartite system where recognition is strongest for identical-token same-talker repetitions, followed by novel-token same-talker repetitions, followed by different-talker repetitions. The analysis of responses by talker revealed an altogether different schema where novel token repetitions are recognized just as well as identical repetitions for some talkers, but not for others. This demonstrates the importance of considering the specific talker producing the stimuli and suggests that the encoding of spoken words is not equivalent from talker to talker. Individual talkers are encoded differently and induce unique patterns of results. Future studies may benefit from using stimuli from as many talkers as is feasible, and where possible, analyzing results with consideration of talker-level asymmetries.

Several other outstanding questions were apparent in the data. Contrary to our hypothesis that identical token repetitions would facilitate recognition and therefore be characterized by faster responses relative to novel tokens, the only significant effect of TokenCondition on RT (Model 2.4) showed that in some cases, responses were in fact somewhat *slower* for identical tokens than novel tokens in Exp. 2. Because this was not the central focus of analysis in the current study, additional experiments would need to be conducted to confirm the validity of this finding. In the case that this finding is replicated, it may point to differences in the nature of the tasks of recognizing identical and novel tokens. While the most typical pattern described across chronometric research is that high accuracy is associated with low RTs (a negative correlation), it has been observed that in some cases, more difficult tasks requiring more attention to the stimulus may induce a positive correlation, or high accuracy associated with high RTs (see De Boeck & Jeon, 2019 for discussion). This pattern of results is consistent with findings in Exp. 2, where slower RTs were accompanied by higher accuracy in the Identical condition relative to the

Novel condition. In this case, the positive correlation would theoretically emerge because the nature of the task encourages listeners to attend to subtle acoustic characteristics of the stimuli rather than broad voice characteristics, making the process of correctly identifying a match more often successful, but also more cognitively burdensome than when fine-grained acoustic characteristics are not available. However, for reasons described above, the interpretation of patterns in the latency data is not straightforward and more convincing explanations of these patterns will require their own studies.

5. Conclusion

In this study, we asked whether listeners encoded higher level voice characteristics in addition to low-level acoustic properties of spoken words, whether widely replicated talker-specificity effects were bolstered by an experimental confound involving the repetition of a physical stimulus, whether encoding was stable across demographically matched talkers, and whether talker-level asymmetries in recognition were influenced by the type of token heard at a word's repetition. We replicated talker-specificity effects in the context of novel token repetitions but found that these effects were less stable than when the exact physical stimulus was repeated. Specifically, when novel token repetitions were heard, it was clear that listeners encoded the voice of one talker more strongly than the voice of the other talker, suggesting that the degree to which listeners encode voice characteristics may depend in part on the voice itself. We draw on this observation to suggest that future research in talker specificity use novel rather than identical token repetitions, given that this approach provides a more sensitive measure of memory, a context where asymmetrical encoding strengths are more readily observable, and a more direct test of the encoding of high-level voice characteristics as opposed to idiosyncratic statistical patterns present in the stimuli. It will also be important for researchers to consider the voices used for stimulus creation, given that the cognitive processes we aim to study are in part dependent on the characteristics of the voices. Future research will be needed to determine the extent to which various cognitive processes are or are not contingent on the voices heard as well as the characteristics of voices that encourage or discourage various processing strategies.

Declaration of Competing Interests

No competing interests to declare.

CRediT Author Contributions

William Clapp: Conceptualization, Methodology, Software, Data Curation, Visualization, Formal Analysis, Writing – Original Draft, Writing – Review & Editing; *Charlotte Vaughn*: Conceptualization, Methodology, Formal Analysis, Writing – Review & Editing; *Simon Todd*: Conceptualization, Methodology, Formal Analysis, Writing – Review & Editing; *Meghan Sumner*: Conceptualization, Methodology, Formal Analysis, Writing – Review & Editing, Funding Acquisition

Acknowledgements

We appreciate members of the Stanford Phonetics Lab and attendees of LabPhon 18 for helpful comments on this work.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278. doi: 10.1016/j.jml.2012.11.001

Barreda, S. (2012). Vowel normalization and the perception of speaker changes: an exploration of the contextual tuning hypothesis. *The Journal of the Acoustical Society of America*, *132*, 3453–3464. doi: 10.1121/1.4747011

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bell, A. (1984). Language style as audience design. *Language in Society*, *13*(2). 145–204. doi: 10.1017/S004740450001037X

Boersma, P. & Weenink, D. (2021). Praat: doing phonetics by computer. [Computer program]. Version 6.1.22.

Bradlow, A. R., Nygaard, L. C., and Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics, 61*(2), 206–219. doi: 10.3758/BF03206883

Brehm, L., & Alday, P. M. (2022). Contrast coding choices in a decade of mixed models. *Journal of Memory and Language*, *125*. 104334. doi: 10.1016/j.jml.2022.104334

Brysbaert, M. & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990. doi: 10.3758/BRM.41.4.977

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., Volcic, R., & De Rosario, H.. (2017). *pwr*: Basic functions for power analysis. [Computer program]. https://cran.r-project.org/web/packages/pwr/

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), 521–533. doi: 10.1037/0278-7393.20.3.521

Clapp, W., Vaughn, C., & Sumner, M. (2023). The episodic encoding of talker voice attributes across diverse voices. *Journal of Memory and Language, 128.* 104376. doi: 10.1016/j.jml.2022.104376

Clopper, C. G., Tamati, T. N., & Pierrehumbert, J. B. (2017). Variation in the strength of lexical encoding across dialects. *Journal of Phonetics, 58*, 87–103. doi: 10.1016/j.wocn.2016.06.002

Cooper, A., Brouwer, S., & Bradlow, A. R. (2015). Interdependent processing and encoding of speech and concurrent background noise. *Attention, Perception, & Psychophysics, 77*, 1342–1357. doi: 10.3758/s13414-015-0855-z

Craik, F. I. M. & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology, 26*(2), 274–284. doi: 10.1080/14640747408400413

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: 10.1080/01690965.2011.610597

De Boeck, P. & Jeon, M. (2017). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology, 10.* 102. doi: 10.3389/fpsyg.2019.00102

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods, 47*(1), 1–12. doi: 10.3758/s13428-014-0458-y

Eckert, P. (2008). Variation and the Indexical Field. *Journal of Sociolinguistics, 12*(4), 453–476. doi: 10.1111/j.1467-9841.2008.00374.x

Eckert, P. (2012). Three Waves of Variation Study: The Emergence of Meaning in the Study of Sociolinguistic Variation. *Annual Review of Anthropology, 41,* 87–100. doi: 10.1146/annurev-anthro-092611-145828

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: 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), 1166-1183. doi: 10.1037/0278-7393.22.5.1166

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychol. Rev.* 105, 251–279. doi: 10.1037/0033-295x.105.2.251

Hanique, I., Aalders, E., & Ernestus, M. (2014). How robust are exemplar effects in word comprehension? *The Mental Lexicon, 8*(3), 269–294. doi: 10.1075/ml.8.3.01han

Hasan, B.A.S., Valdes-Sosa, M., Gross, J., Belin, P., 2016. Hearing faces and seeing voices: Amodal coding of person identity in the human brain. *Scientific Reports, 6*, 37494. doi: 10.1038/srep37494

Jakobson, R., Fant, G., and Halle, M. (1952). *Preliminaries to Speech Analysis*. MIT Press, Cambridge, MA.

Johnson, K. (1997). Speech perception without speaker normalization: an exemplar model, in K. Johnson and J. W. Mullennix (Eds.), *Talker Variability in Speech Processing* (pp. 145–165). Academic Press.

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

Kim, SK & Sumner, M. (2015). Effects of emotional prosody and attention on semantic priming. Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.) *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, pp. 1099 - 1104.

Kiparsky, P. (1973) How abstract is phonology? In O. Fujimura (Ed.), *Three Dimensions of Linguistic Theory* (pp. 57–86). Tokyo Inst. Adv. Stud. Lang.

Kuznetsova A., Brockhoff P. B., Christensen R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. doi: 10.18637/jss.v082.i13

Labov, W. (1966) *The Social Stratification of English in New York City*. Washington, DC: Center for Applied Linguistics.

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), 1235–1242. doi: 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: 10.3758/s13414-012-0315-y

Morano, L., ten Bosch, L., Ernestus, M. (2019). Looking for exemplar effects: Testing the comprehension and memory representations of r'duced words in Dutch learners of French. In S. Fuchs, J. Cleland, & A. Rochet-Capellan (Eds.), *Speech Production and Perception: Learning and Memory* (pp. 245–277). Peter Lang.

Nygaard, L. C., Burt, S. A., & Queen, J. S. (2000). Surface form typicality and asymmetric transfer in episodic memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(5), 1228. doi: 10.1037/0278-7393.26.5.1228

Nygaard, L. C., Sommers, M., and Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception & Psychophysics, 57*(7), 989–1001. doi: 10.3758/BF03205458

Palmeri, T. J., Goldinger, S. D., and Pisoni, D. B. (1993). Episodic Encoding of Voice Attributes and Recognition Memory for Spoken Words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(3), 309–328. doi: 10.1037//0278-7393.19.2.309

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

Pufahl, A. & Samuel, A. G. (2014). How Lexical is the Lexicon? Evidence for Integrated Auditory Memory Representations. *Cognitive Psychology, 70*, 1–30. https://psycnet.apa.org/doi/10.1016/j.cogpsych.2014.01.001

Russell V. L. (2022). emmeans: Estimated Marginal Means, aka Least-Squares Means. [Computer program]. Version 1.7.5.

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition: Effects of variability in speaking rate and overall amplitude. *The Journal of the Acoustical Society of America*, *96*(3), 1314– 1324. doi: 10.1121/1.411453

Stampe, D. (1979). *A Dissertation on Natural Phonology*. Garland, New York.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *Journal of the Acoustical Society of America, 111*:1872–1891. doi: 10.1121/1.1458026

Strori, D., Zaar, J., Cooke, M., & Mattys, S.L. (2018). Sound specificity effects in spoken word recognition: The effect of integrality between words and sounds. *Attention, Perception, & Psychophysics, 80*, 222–241. doi: 10.3758/s13414-017-1425-3

Sumner, M. & Kataoka, R. (2013). Effects of phonetically-cued talker variation on semantic encoding. *Journal of the Acoustical Society of America, 134*(6), EL485. doi: 10.1121/1.4826151

Sumner, M., Kim, S.K., King, E., & McGowan, K.B. (2014). The socially weighted encoding of spoken words: A dual-route approach to speech. *Frontiers in Psychology, 4,* 1015. doi: 10.3389/fpsyg.2013.01015

Sumner, M., & Samuel, A. G. (2005). Perception and representation of regular variation: the case of final-/t/. *Journal of Memory and Language, 52*, 322–338. doi: 10.1016/j.jml.2004.11.004

van Dommelen, W. (1990). Acoustic Parameters in Human Speech Recognition. *Language and Speech, 33*(3), 259–272. doi: 10.1177/002383099003300302

van Lancker, D. R., Kreiman, J., & Cummings, J. (1989). Voice perception deficits: Neuroanatomical correlates of phonagnosia. *Journal of Clinical and Experimental Neuropsychology, 11*(5), 665–674. doi: 10.1080/01688638908400923

Viswanathan, J., Rémy, F., Bacon-Macé, N., & Thorpe, S. J. (2016). Long Term Memory for Noise: Evidence of Robust Encoding of Very Short Temporal Acoustic Patterns. *Frontiers in Neuroscience, 10*, 490. doi: 10.3389/fnins.2016.00490 von Kriegstein, K. & Giraud, A. (2006). Implicit Multisensory Associations Influence Voice Recognition. *PLOS Biology, 4*(10), 1809–1820. doi: 10.1371/journal.pbio.0040326

Winkler, I., Korzykov, O., Gumenyuk, V., Cowan, N., Linkenkaer-Hansen, K., Alho, K., Ilmoniemi, R. J., & Näätänen, R. (2002). Temporary and longer retention of acoustic information. *Psychophysiology, 39*, 530–534. doi: 10.1017/S0048577201393186

Winters, S. J., Levi, S. V., Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages. *The Journal of the Acoustical Society of America, 123*, 4524. doi: 10.1121/1.2913046

Woods, K. J. P., Siegel, M. H., Traer, K., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics, 79*(7), 2064–2072. doi: 10.3758/s13414-017-1361-2

Appendix A – Exp. 1 Model Summaries

Model 1.1

Hits ~ TokenCondition × Talker × Lag + TrialNumber + (1 + Lag | Participant) + (1 + Talker | Item)

	β	SE	z	p	Sig.
Intercept	2.859732	0.092228	31.007	<2e-16	***
TokenCondition–Identical	-0.004774	0.120129	-0.04	0.9683	
Talker–M2	-0.201397	0.125794	-1.601	0.1094	
Lag	-1.783588	0.125827	-14.175	<2e-16	***
TrialNumber	0.496981	0.061472	8.085	6.23E-16	***
TokenCondition–Identical : Talker–M2	0.302764	0.173012	1.75	0.0801	.
TokenCondition–Identical : Lag	0.036061	0.171911	0.21	0.8339	
Talker–M2 : Lag	-0.053959	0.171095	-0.315	0.7525	

Model 1.2

logRT ~ TokenCondition × Talker × Lag + TrialNumber + Duration + (1 + Lag | Participant) + (1 + TokenCondition + Talker + Lag | Item)

	β	SE	DF	t	p	Sig.
Intercept	2.62	0.0136	404	192.577	<2e-16	***
TokenCondition–Identical	0.00621	0.0183	355	0.339	0.735	
Talker–M2	0.0129	0.0191	376	0.677	0.499	
Lag	0.121	0.00861	359	13.998	<2e-16	***
TrialNumber	-0.0540	0.00297	42900	-18.16	<2e-16	***
Duration	-0.318	0.0119	715	-26.701	<2e-16	***
TokenCondition–Identical : Talker–M2	0.00629	0.0264	351	0.238	0.812	
TokenCondition–Identical : Lag	0.00433	0.0116	335	0.374	0.708	
Talker–M2 : Lag	-0.00195	0.0118	331	-0.165	0.869	
TokenCondition–Identical : Talker–M2 : Lag	-0.00794	0.0167	331	-0.477	0.634	

Model 1.3

FalseAlarms ~ TokenCondition × Talker + TrialNumber + (1 | Participant) + (1 + TokenCondition + Talker | Item)

	β	SE	z	p	Sig.
Intercept	-2.2512	0.10682	-21.074	<2e-16	***
TokenCondition–Identical	0.01537	0.1274	0.121	0.904	
Talker–M2	0.10678	0.13248	0.806	0.42	
TrialNumber	1.22763	0.04898	25.064	<2e-16	***
TokenCondition–Identical : Talker–M2	0.22693	0.18193	1.247	0.212	

Model 1.4

d’ ~ TokenCondition × Talker

	β	SE	t	p	sig.

Intercept	1.68999	0.07466	22.636	<2e-16	***
TokenCondition–Identical	-0.08663	0.10419	-0.831	0.4063	
Talker–M2	-0.23025	0.10713	-2.149	0.0323	*
TokenCondition–Identical : TalkerM2	0.31933	0.15053	2.121	0.0346	*

Appendix B – Exp. 2 Model Summaries

Model 2.1

Hits ~ TokenCondition × RepVoice × Lag + TrialNumber + (1 + RepVoice + Lag | Participant) + (1 + TokenCondition + RepVoice + Lag | Item)

	β	SE	z	p	Sig.
Intercept	2.28284	0.07699	29.652	<2e-16	***
TokenCondition–Identical	0.14207	0.08517	1.668	0.095312	.
RepVoice–Same	0.11617	0.0532	2.184	0.028999	*
Lag	-1.74645	0.10633	-16.424	<2e-16	***
TrialNumber	0.27868	0.05281	5.277	1.32E-07	***
TokenCondition–Identical : RepVoice–Same	0.22701	0.06736	3.37	0.000751	***
TokenCondition–Identical : Lag	-0.10846	0.14205	-0.764	0.445144	
RepVoice–Same:Lag	-0.07122	0.13342	-0.534	0.593452	
TokenCondition–Identical : RepVoice–Same : Lag	0.23889	0.19183	1.245	0.213001	

Model 2.2

Hits ~ TokenCondition × FirstTalker × SecondTalker + Lag + TrialNumber + Lag:TokenCondition + Lag:FirstTalker + Lag:SecondTalker + (1 + FirstTalker + SecondTalker + Lag | Participant) + (1 + FirstTalker + SecondTalker + Lag | Item)

	β	SE	z	p	Sig.
Intercept	2.72732	0.0868	31.419	<2e-16	***

TokenCondition–Identical	0.18445	0.10385	1.776	0.075703	.
FirstTalker–M2	-0.65633	0.07257	-9.045	<2e-16	***
SecondTalker–M2	-0.23333	0.06951	-3.357	0.000789	***
Lag	-1.89438	0.11509	-16.461	<2e-16	***
TrialNumber	0.27754	0.05299	5.238	1.63E-07	***
TokenCondition–Identical : FirstTalker–M2	-0.02967	0.0895	-0.332	0.740257	
TokenCondition–Identical : SecondTalker–M2	-0.01822	0.09075	-0.201	0.84084	
FirstTalker–M2 : SecondTalker–M2	0.33996	0.08438	4.029	5.60E-05	***
TokenCondition–Identical : Lag	-0.01077	0.10952	-0.098	0.92167	
FirstTalker–M2 : Lag	0.13755	0.10015	1.373	0.16965	
SecondTalker–M2 : Lag	0.03507	0.09902	0.354	0.723184	
TokenCondition–Identical : FirstTalker–M2 : SecondTalker–M2	0.42248	0.12063	3.502	0.000461	***

Model 2.3

logRT ~ TokenCondition × RepVoice × Lag + TrialNumber + Duration + (1 + Lag | Participant) + (1 + TokenCondition + Lag | Item)

	β	SE	DF	t	p	Sig.
Intercept	2.65	0.00977	446	271.084	<2e-16	***
TokenCondition–Identical	0.0164	0.0132	395	1.239	0.2162	
RepVoice–Same	-0.0100	0.00245	43700	-4.085	4.42E-05	***
Lag	0.123	0.00778	784	15.835	<2e-16	***
TrialNumber	-0.0372	0.00296	43700	-12.535	<2e-16	***
Duration	-0.335	0.00611	12200	-54.863	<2e-16	***
TokenCondition–Identical : RepVoice–Same	0.00340	0.00340	43800	1.000	0.3173	
TokenCondition–Identical : Lag	-0.0178	0.0106	882	-1.687	0.0919	.

RepVoice–Same : Lag	0.00398	0.00908	43200	0.438	0.6616	
TokenCondition–Identical : RepVoice–Same : Lag	0.00414	0.0126	43200	0.328	0.7426	

Model 2.4

logRT ~ TokenCondition × FirstTalker × SecondTalker + Lag + TrialNumber + Duration + Lag:TokenCondition + Lag:FirstTalker + Lag:SecondTalker + (1 + FirstTalker + SecondTalker + Lag | Participant) + (1 + TokenCondition + FirstTalker + SecondTalker + Lag | Item)

	β	SE	DF	t	p	Sig.
Intercept	2.63	0.0105	473	249.642	<2e-16	***
TokenCondition–Identical	0.0283	0.0139	393	2.033	0.04274	*
FirstTalker–M2	0.0175	0.00370	949	4.728	2.61E-06	***
SecondTalker–M2	0.0301	0.00532	1040	5.646	2.11E-08	***
Lag	0.121	0.00767	744	15.819	<2e-16	***
TrialNumber	-0.0373	0.00295	43600	-12.619	<2e-16	***
Duration	-0.301	0.0120	713	-25.037	<2e-16	***
TokenCondition–Identical : FirstTalker–M2	-0.00611	0.00496	1330	-1.232	0.21823	
TokenCondition–Identical : SecondTalker–M2	-0.0159	0.00541	936	-2.939	0.00338	**
FirstTalker–M2 : SecondTalker–M2	-0.0198	0.00486	43100	-4.079	4.54E-05	***
TokenCondition–Identical :Lag	-0.0158	0.00838	356	-1.882	0.06071	.
FirstTalker–M2:Lag	0.00113	0.00628	43000	0.179	0.85793	
SecondTalker–M2:Lag	0.00665	0.00629	43200	1.057	0.29034	
TokenCondition–Identical : FirstTalker–M2 : SecondTalker–M2	0.00589	0.00675	43100	0.873	0.38264	

Model 2.5

FalseAlarms ~ TokenCondition × Talker + TrialNumber + (1 + Talker | Participant) + (1 + Talker | Item)

	β	SE	z	p	Sig.
Intercept	-2.16869	0.08458	-25.642	<2e-16	***
TokenCondition–Identical	0.12055	0.09282	1.299	0.194	
TalkerM2	0.29472	0.05098	5.781	7.44E-09	***
TrialNumber	1.23554	0.04598	26.871	<2e-16	***
TokenCondition–Identical : TalkerM2	-0.15107	0.06214	-2.431	0.0151	*

Model 2.6

d’ ~ TokenCondition

	β	SE	t	p	Sig.
Intercept	1.29586	0.02646	48.972	<2e-16	***
TokenCondition–Identical	0.11169	0.03713	3.008	0.0028	**

Model 2.7

By-talker d’ ~ TokenCondition × Talker + (1 | Participant)

	β	SE	DF	t	p	Sig.
Intercept	1.46437	0.02973	539.1072	49.26	<2e-16	***
TokenCondition–Identical	0.05575	0.04165	539.1072	1.338	0.1813	
FirstTalker–M2	-0.25829	0.02593	373	-9.96	<2e-16	***
TokenCondition–Identical : FirstTalker–M2	0.10352	0.03634	373	2.849	0.00463	**