Attention, Perception, & Psychophysics (2022) 84:2002–2015
https://doi.org/10.3758/s13414-022-02500-8

Implicit and explicit learning in talker identification

Jayden J. Lee[1] · Tyler K. Perrachione[1]

Accepted: 23 April 2022 / Published online: 9 May 2022

Abstract
In the real world, listeners seem to implicitly learn talkers' vocal identities during interactions that prioritize attending to the content of talkers' speech. In contrast, most laboratory experiments of talker identification employ training paradigms that require listeners to explicitly practice identifying voices. Here, we investigated whether listeners become familiar with talkers' vocal identities during initial exposures that do not involve explicit talker identification. Participants were assigned to one of three exposure tasks, in which they heard identical stimuli but were differentially required to attend to the talkers' vocal identity or to the verbal content of their speech: (1) matching the talker to a concurrent visual cue (*talker-matching*); (2) discriminating whether the talker was the same as the prior trial (*talker 1-back*); or (3) discriminating whether speech content matched the previous trial (*verbal 1-back*). All participants were then tested on their ability to learn to identify talkers from novel speech content. Critically, we manipulated whether the talkers during this post-test differed from those heard during training. Compared to learning to identify novel talkers, listeners were significantly more accurate learning to identify the talkers they had previously been exposed to in the talker-matching and verbal 1-back tasks, but not the talker 1-back task. The correlation between talker identification test performance and exposure task performance was also greater when the talkers were the same in both tasks. These results suggest that listeners learn talkers' vocal identity implicitly during speech perception, even if they are not explicitly attending to the talkers' identity.

Keywords Attention in learning · Perceptual categorization and identification · Psycholinguistics

During speech perception, listeners encode not only the meaning of the words they hear but also cues about the speaker's identity from their voice. Deconvolving the verbal and indexical information in the speech signal is an essential part of linguistic and social communication. The voice conveys a host of indexical information such as the speaker's sex, age, native language status, and social or regional dialect (Bent & Holt, 2017) and is often characterized as an embodiment of the speaker's personhood (Lavan et al., 2019). The ability to distinguish voices from one another is influenced by a variety of factors, including the cognitive, perceptual, and linguistic abilities of the listener (Best, Ahlstrom, et al., 2018a; Garrido et al., 2009), the acoustic properties of the talker (Kreiman & Sidtis, 2011; Latinus & Belin, 2011; Perrachione et al., 2019) and their interaction (Bregman & Creel, 2014; Perrachione, 2019; Sumner et al., 2014). Many phenomena have been discovered that demonstrate the extent to which processing the linguistic and indexical features of speech are bidirectionally intertwined (Scott, 2019). For example, in the *familiar talker advantage*, speech comprehension is better when hearing familiar compared to unfamiliar talkers (Kreitewolf et al., 2017; Levi et al., 2011; Nygaard & Pisoni, 1998; Palmeri et al., 1993; Schweinberger et al., 2014; Souza et al., 2013) while in the *language-familiarity effect*, listeners are better able to recognize voices in their native language compared to a non-native one (Hollien et al., 1982; Goggin et al., 1991; Perrachione & Wong, 2007; Thompson, 1987). However, little is known about how listeners develop representations of voice identity when their primary communication goal is to understand the content of talkers' speech. In this study, we investigate the extent to which listeners learn talkers' identities from exposure tasks that vary whether listeners' attention is directed to talkers' voices vs. the content of their speech.

An abundance of research has pushed our understanding of the synergistic relationship between linguistic processing and talker identification using various types of speech stimuli (Fleming et al., 2014; Johnsrude et al., 2013; Levi, 2019; Nygaard & Pisoni, 1998; Souza et al., 2013). Recent studies of talker identification have attempted to ascertain how this

✉ Tyler K. Perrachione
tkp@bu.edu

[1] Department of Speech, Language, & Hearing Sciences, Boston University, 635 Commonwealth Ave, Boston, MA 02215, USA

higher-level cognitive ability is supported by encoding various low-level acoustic or linguistic features. As a consequence, such studies have often employed slates of voices that vary in controlled ways (e.g., Latinus et al., 2013; Latinus & Belin, 2011), and are therefore frequently limited to paradigms where listeners are explicitly trained on the identity of the slate of talker stimuli during the experiment (e.g., Bregman & Creel, 2014; Levi et al., 2011; McLaughlin et al., 2019; Nygaard et al., 1994; Orena et al., 2015). Fewer studies have looked into the more ecological aspects of talker identification, wherein listeners are assessed for their ability to learn or recognize talker identities outside of the laboratory (cf. Lavan et al., 2018; Schweinberger et al., 1997). In real world communicative interactions, a talker's identity is not learned through short, scaffolded attempts at explicit talker identification analogous to prior laboratory training tasks. Instead, listeners implicitly accumulate talker identity knowledge over time while engaging in ecological communicative activities, such as conversations, during which attention is directed not at determining the identity of a talker, but at understanding the content of their speech (Sidtis & Kreiman, 2012). The paradigmatic divide between how talker identification abilities have been trained and assessed in the laboratory versus how they tend to be learned in real life motivates the need for laboratory research directed at understanding the implicit learning of talker identity.

Conventional talker identification experiments have often followed a structured training paradigm of first explicitly familiarizing participants with a set of voices and subsequently testing their ability to recognize those trained voices (e.g., Goldstein et al., 1981; Hollien et al., 1982; Levi et al., 2011; McLaughlin et al., 2019; Nygaard & Pisoni, 1998; Orena et al., 2015; Zarate et al., 2015). Listeners may be introduced to each talker one by one, as each voice is presented speaking in turn and paired with some corresponding visual depiction of the talker's identity (e.g., a number, name, face, cartoon, etc.; Senior et al., 2018), while explicitly identifying one talker after another on many consecutive trials of the same task. Laboratory training usually provides feedback immediately after each trial indicating the correct answer, as listeners are tested later on their ability to maintain and recall the talker representations. While the nature of the speech stimuli varies from study to study (e.g., isolated vowels, words, sentences, or longer recordings; Cook & Wilding, 1997; Nygaard & Pisoni, 1998; Yarmey, 1995), the framework of explicitly training listeners on the identity of the voices used in each study predominates (reviewed in Perrachione, 2019, and Levi, 2019).

However, these approaches tend to gloss over a fundamental distinction between how listeners deploy selective attention to talker identity during laboratory familiarization of talkers versus how they appear to learn to associate voices with individuals in the real world. In ecological communication, listeners are primarily attending to the content of their interlocutors' speech. The structure of real-world conversations appears to limit the need to explicitly attend to vocal identity cues during communication (at least for sighted listeners) who know the identity of their interlocutor by seeing their face. An unresolved question, then, is how listeners learn a talker's identity when they are not required to actively or explicitly attend to the distinctive indexical features of a talker's voice during speech perception. Conversely, this also raises the question of the extent to which laboratory experiments of talker identification actually inform our understanding of how voices are encoded and recognized in real-world communication settings. Devising the training structure for perceptual learning tasks is critically important to shed light on these questions and to the carryover of the result from training-induced performance, the robustness of learning, and the likelihood of generalizing to novel speakers (Ahissar & Hochstein, 1997; Goldstone, 1998; Karni & Sagi, 1993; Ortiz & Wright, 2010; Tzeng et al., 2016).

Studies on perceptual learning have demonstrated that learning does not solely depend on explicit training but can also occur from unattended exposure to stimuli when the exposure phase has incorporated a sufficient number of trials to drive neural response to stimulation (Seitz et al., 2010; Seitz & Watanabe, 2009; Watanabe & Sasaki, 2015). Converging evidence in audition and vision suggests that extensive exposure to a stimulus can lead to learning, even in the absence of focused attention on the relevant feature of the stimulus, known as *task-irrelevant learning* (Ahissar & Hochstein, 1993; Watanabe et al., 2001). In this framework, perceptual learning can occur in situations of unattended and passive sensory stimulation as long as a threshold for learning is reached by a stimulus-related activity (Seitz & Dinse, 2007; Seitz & Watanabe, 2005). The process of learning can be distilled into two distinct phases: *acquisition* (the training period) and *consolidation* (the transfer of the learned information from a fragile neural state to a stable state); many studies have been performed to track this progression, including in the auditory and speech learning realm (Banai et al., 2010; Wright et al., 2010a, b; Wright & Zhang, 2009). Neural experiments have provided further evidence substantiating this learning model: rapid neural plasticity in sensory cortices occurs under task-irrelevant learning conditions, and additional sensory stimulation without the added benefit of trained performance also leads to neural changes (Ahissar et al., 2009; Banai et al., 2010; Desimone, 1996; Karni, 1996; Law & Gold, 2008).

Task-irrelevant learning can be incorporated into training paradigms to reduce the cognitive demands of perceptual learning. Perceptual skills in both audition and vision have been shown to improve even when reducing lengthy sessions of continuous practice of a specific task and interleaving practice with periods of mere exposure to the stimuli (e.g., Banai

et al., 2010; Wright et al., 2015). The effect was still present when exposure periods included performing an orthogonal task to the skill being learned, leading to testing various training paradigms seeking the model for optimal accuracy results. For example, interleaving periods of solely active practice on a task with additional stimulus exposure yielded the highest accuracy in performance of speech tasks like acquisition of a novel phonetic category and adaptation to foreign-accented speech (Wright et al., 2015). Implicit training has also been found to result in superior learning to explicit training in acquisition of nonnative phonetic contrasts (Luthra et al., 2019; Vlahou et al., 2012). However, such models have not yet been considered in the context of how listeners learn to identify talkers. It is currently unknown to what extent conventional laboratory paradigms that explicitly train listeners to identify a slate of talkers by the sound of their voice adequately reflect the cognitive, perceptual, and mnemonic processes that underlie talker identity learning during real life social interactions.

The present study makes a first step towards understanding implicit talker identification learning during communication by investigating how the principles of task-irrelevant learning apply to talker identification abilities. We examined to what extent exposing listeners to talkers' voices, without explicitly directing their attention to those talkers' identities, could lead to developing perceptual representations of talker identity. We tested this effect by comparing listeners' ability to learn the identities of the exposed versus novel talkers in a subsequent explicit talker identification task. In a mixed between/within-subjects design, we assigned participants to one of three familiarization task conditions, in which the amount of selective attention allocated to processing talker identity was parametrically manipulated. The three familiarization tasks in this experiment modulated the level of selective attention that listeners had to pay to the indexical features of the talkers' voices: (1) a *talker matching* task, in which listeners actively practiced assigning identities to voices; (2) a *talker 1-back* task, in which listeners decided whether the talker was the same as in the previous trial, but without needing to distinguish the talkers as individuals; and (3) a *verbal 1-back* task, in which listeners decided whether a portion of the speech content was the same as in the previous trial, but without needing to attend to talker characteristics at all. Thus, the talker matching task required the most focused attention on the vocal identity of the talkers in order to perform the familiarization task, while the verbal 1-back task required the least.

To assess the effect of exposure, we compared performance between two versions of each task: one in which the voices heard during the familiarization phase were the same as those during a subsequent talker identification test, and one in which listeners heard different voices during the test than those they had been familiarized with. Perceptual learning frameworks have described how learning one stimulus feature can occur while attending to a different feature during training (Seitz et al., 2010; Watanabe et al., 2001; Watanabe & Sasaki, 2015). Correspondingly, we hypothesized that talker identification performance would always be better for familiarized vs. unfamiliarized voices, regardless of the task during the familiarization phase. However, we also hypothesized that, when comparing across the task groups, there would be a marked difference in talker identification accuracy between familiar and unfamiliar talkers depending on which exposure task the participant had performed: The talker matching task most resembles explicit talker identification and, therefore, should directly lead to the greatest effect of familiarity in the subsequent talker identification test; the effect of familiarity should then be correspondingly smaller for the talker 1-back and then verbal 1-back tasks, respectively. Overall, we aimed to understand how the amount of attention listeners paid to talkers' voices during a familiarization task affected their ability to later identify those voices.

Methods

Participants

Native speakers of American English (N = 96; 74 female, 22 male, ages 18–31; mean = 20.5 years) who reported no history of speech, hearing, or language disorder completed this study. All participants provided informed, written consent prior to undertaking this experiment. Additional participants were recruited, but those who declined to complete the hour-long study or who responded to fewer than 80% of the trials in the exposure phase were excluded from analysis (n = 31). The study was approved and overseen by the Internal Review Board at Boston University.

Stimuli

Ten spoken digits ("zero" through "nine") were recorded by 20 adult native speakers of American English (10 male, 10 female). Stimuli were recorded in a sound attenuated chamber with a MX153 microphone and Roland Quad Capture sound card sampling at 44.1 kHz and 16 bits. Recordings of each digit were isolated and normalized to 70 dB SPL RMS amplitude in Praat (Boersma, 2001). On each trial, participants heard a sequence of five unique digits, all spoken by the same talker. Recordings of the digits 1, 3, 4, 7, and 8 were used as stimuli exclusively during the exposure phase, and recordings of the digits 2, 5, 6, 9, and 0 were used exclusively during the test phase. These groupings were chosen to maximize the phonetic diversity from each talker during exposure and test, and to ensure that listeners' performance at test was based on a generalizable knowledge about the talkers' voices, not specific memories for the speech tokens during familiarization. Although stimuli based on concatenated recordings lack some of the coarticulatory or prosodic qualities

of natural, connected speech, the phonetic and indexical variability in such stimuli have nonetheless been shown to be effective for studying talker-specific effects in speech perception (e.g., Bolia et al., 2000; Bressler et al., 2014; Kidd et al., 2008; Lim et al., 2019, 2021).

Talkers were organized into four groups of five (two groups of five male talkers each, two groups of five female talkers each). Gross acoustic features of the talkers' recordings are shown in Table 1. The assignment of talkers to groups was extensively piloted to obtain roughly equal talker identification performance by listeners for each group of talkers. Each talker was associated with a distinct cartoon avatar.

Design

Each run of the experiment comprised an *exposure phase* and a *test phase*. During the exposure phase, participants performed one of three *exposure tasks* (talker matching, talker 1-back, verbal 1-back) designed to focus their attention on the audio stimuli in different ways. Participants were randomly assigned to each exposure task in a between-subjects design (see below). During the test phase, all participants performed an identical explicit talker identification test. Each participant completed two runs of the experiment (Fig. 1a), with the same exposure task in both runs. Recordings from different talkers were used during each run of the experiment, and the particular talkers that listeners heard during the exposure and test phases formed the basis of our within-subjects *familiarity condition* manipulation. In the *exposed condition*, participants heard the same voices during the exposure and test phases; in the *novel condition*, participants heard recordings from one group of talkers during the exposure phase and from a new group of talkers during the test phase. The order of conditions and the talkers assigned to each condition were counterbalanced across participants.

For this mixed 2 (*familiarity condition*; within-subjects) × 3 (*exposure task*; between-subjects) factorial design, the stimulus structure of the exposure phase was identical for all participants. This ensured that all participants, regardless of their exposure task, had exactly the same amount of exposure to the talkers during the exposure phase. That is, the trial-by-trial composition of stimuli (both audio recordings and visual graphics) was identical for all participants regardless of which task they were performing or which talkers they were hearing (see Fig. 1b).

Table 1 Global acoustic characteristics of talkers

Group	Talker	Mean f_0	Mean HNR	Est. VTL (cm)
Male-1	M0	109.98	14.17	17.15
	M1	97.84	12.96	16.20
	M2	99.73	9.32	15.28
	M3	109.53	11.39	15.33
	M4	102.87	9.52	14.82
Male-2	M5	116.96	12.95	16.63
	M6	83.07	8.26	17.10
	M7	110.69	11.22	15.81
	M8	116.72	9.38	16.17
	M9	97.26	10.91	16.10
Female-1	F0	156.82	10.42	15.70
	F1	232.34	18.93	14.44
	F2	186.31	14.68	15.65
	F3	181.76	14.28	14.45
	F4	216.37	18.91	14.74
Female-2	F5	196.53	16.55	14.82
	F6	200.22	15.50	15.78
	F7	219.2	17.75	14.51
	F8	191.06	17.12	14.91
	F9	190.46	16.26	15.06

HNR harmonics-to-noise ratio, *Est. VTL* estimated vocal tract length

Procedure

Participants performed an assigned task on every trial during the exposure phase (200 trials/run), in which participants were either explicitly or implicitly familiarized with talkers' voices. In the *talker matching task*, participants decided on every trial whether the talker they heard on that trial matched a simultaneously presented cartoon avatar. In the *talker 1-back task*, participants decided on every trial whether the talker they heard on that trial was the same as the talker they heard on the preceding trial. In the *verbal 1-back task*, participants decided whether a portion of the speech content they heard on that trial was the same as the preceding trial. The structure of each task is described in detail below. After completing the assigned exposure task, all participants performed an identical *explicit talker identification test* (50 trials).

Every trial of the experiment consisted of a single talker speaking a five-digit sequence, along with five avatars presented on screen uniquely representing the group of talkers in that phase. Participants were given feedback after every trial indicating whether their response was "correct" or "incorrect" according to the assigned task, as depicted in Fig. 1b–c. On each trial, the ordinal position of the five avatars seen on the screen and the sequence of five digits heard spoken by the taker was randomized. The order of stimuli was such that one out of every five trials was a "hit" in each task, but which trials were hits was orthogonal across tasks.

Exposure tasks

Talker matching In this task, participants learned to match each talker to their corresponding cartoon avatar. One talker

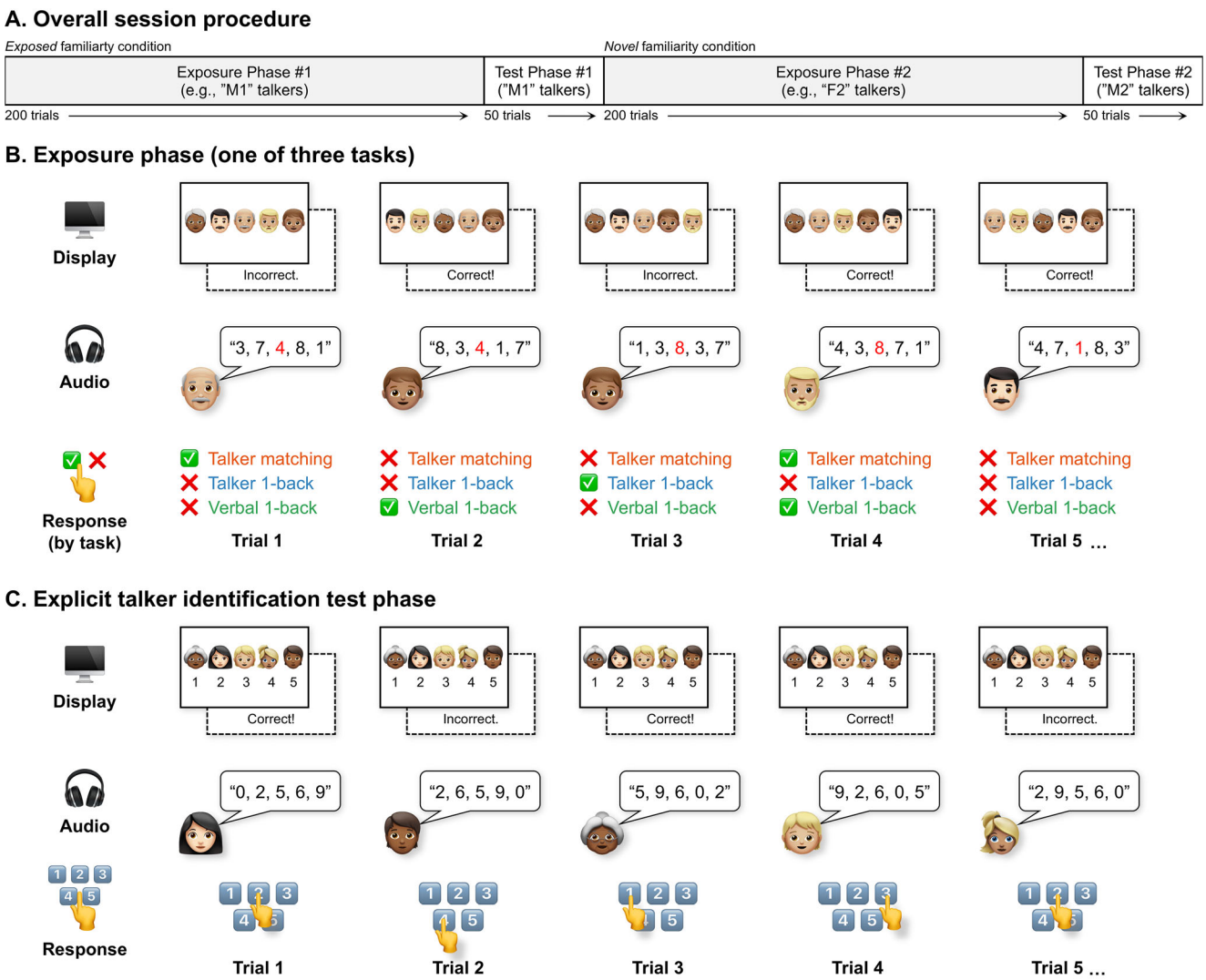

Fig. 1 Task design and procedure. **a** Participants completed two runs, each comprised of an *exposure phase* and *test phase*. The order of *familiarity conditions* (exposed vs. novel) and the talkers heard during each phase were counterbalanced across participants. **b** Participants were assigned to one of three task conditions during the exposure phase. The visual and audio stimuli on each trial were the same across all tasks, but the response demands differed. For each trial, the target (correct) response for each task is shown. In the *talker matching* task (shown in orange), participants were to indicate "yes" when the heard voice matched the central avatar (as shown in Trials 1 and 4), and "no" when they did not match (as in Trials 2, 3, and 5). In the *talker 1-back* task (shown in blue), participants were to indicate "yes" when the talker was the same as the previous trial (as in Trial 3) and "no" when the talker was not the same (as in Trials 1, 2, 4, and 5). In the *verbal 1-back* task (shown in green), participants were to indicate "yes" when the middle digit of the audio stimulus was the same as the previous trial (as in Trials 2 and 4) and "no" when the digits differed (as in Trials 1, 3, and 5). Response trials were orthogonal across conditions. **c** In the explicit talker identification test, participants matched the identity of the voice they heard on that trial to the corresponding avatar. During the *exposed* familiarity condition the talkers were the same as those heard during the exposure phase; during the *novel* familiarity condition (shown), new talkers were heard during the test. Participants received corrective feedback on every trial during both the exposure and test phases. (Color figure online)

would speak a single five-digit string (e.g., "3-7-1-8-4") and the participant would indicate whether the target talker was designated by the middle avatar displayed on screen by pressing the corresponding button (i.e., "Is the middle person speaking? Press 1 for yes, 2 for no."). After providing their response, participants received corrective feedback: If they correctly indicated that the middle avatar corresponded to the talker they heard (*hits*, 20% of trials)—or if they correctly indicated that the middle avatar did not match the talker they heard (*correct rejections*, 80% of trials)—the feedback would state, "Correct!" However, if participants responded with an incorrect association (*misses* or *false alarms*), the feedback would state, "Incorrect." Participants were provided with no *a priori* knowledge about the voice-avatar pairs and learned these associations via feedback during the task. This task maximized listeners' attention to talkers' vocal identity and was most similar to the explicit talker identification training used in many laboratory tasks of talker identification (e.g., Fecher

& Johnson, 2018; Levi, 2015; Perea et al., 2014; Winters et al., 2008; Xie & Myers, 2015).

Talker 1-back In this task, participants performed a 1-back working memory recognition task in which they indicated whether the talker they were hearing on the present trial was the same as the talker they had heard on the immediately preceding trial (i.e., "Is this voice the same as the one before? Press 1 for yes and 2 for no.") This task required participants to attend to the talkers' voice characteristics, but without explicitly needing to learn the individual talkers' identities. Thus, the structure of this task is roughly similar to how talkers are familiarized during laboratory tasks of talker discrimination (e.g., Fecher & Johnson, 2018; Latinus & Belin, 2011; Wester, 2012). As before, participants received corrective feedback on every trial, indicating whether they had correctly determined that the present talker was the same as (or different from) the one from the preceding trial ("Correct!") or had made a miss or false alarm response ("Incorrect."). As before, on 20% of the trials the talker was the same as the preceding trials (hits), and on 80% of the trials the two talkers differed. Importantly, whether the talker repeated on each trial was fully orthogonal to the order of the avatars or the content of the talkers' speech, so listeners could not use a correlated cue for performing the task or learning the talkers' identities. Listeners were not told how many voices there were or given any other individuating information about the voices.

Verbal 1-back In this task, participants also performed a 1-back working memory recognition task in which they indicated whether the middle digit of the sequence they were hearing in the present trial was the same as the middle digit of the sequence in the previous trial (i.e., "Is the middle digit the same as the one before? Press 1 for yes and 2 for no."). A hit was a correct response in indicating the middle digit was a repeat at a rate of 20% as in the other two tasks, and participants were provided the same feedback of "Correct!" for hits and correct rejections and "Incorrect" for misses and false alarms. This task directed participants' attention to focus on the content of speech rather than any distinguishing acoustic properties of the talkers' voices, and listeners were not told about the number of talkers or given any other information about their identity. As before, the relationship between the target dimension (speech content) and other dimensions (talker's voice, avatar) were orthogonal, so listeners could not use a correlated cue to perform the task or discern the talkers' identities. This task is most similar to real-world speech processing demands, where listeners' attention is primarily focused on understanding the verbal content.

Postexposure explicit talker identification phase

After completing their assigned exposure task, all participants undertook an identically structured talker identification task (Fig. 1c). This task consisted of 50 trials in which listeners heard new recordings of five-digit sequences and selected which of the five new cartoon avatars on screen depicted who was speaking. Participants indicated the identity of the current talker by pressing the number designating that talker's avatar. All of the digit stimuli in the test phase were different from those used in the exposure phase (Fig. 1b vs. 1c). This ensured that the test phase measured listeners' generalized knowledge of the talkers' voices, not their specific memories for the speech content from the exposure phase. Furthermore, all of the avatars used during the test phase were different than those seen during the exposure phase. This manipulation ensured that all participants started with an equal knowledge (i.e., none) of the talker-avatar correspondences during the test phase. Feedback was given immediately after each trial to indicate whether participants had identified the talker correctly, including showing the correct avatar for the talker on that trial.

Participants performed the explicit talker identification test two times—once following the exposure task for the *exposed* familiarity condition, in which the voices at test were the same as those in the immediately preceding exposure phase, and once following the exposure task for the *novel* familiarity condition, in which a new set of voices were used during the test phase. During the test phase of the novel condition, the new voices were of the opposite sex to those that had been trained (e.g., if male voices were heard during the exposure phase, female voices were heard at test) to make the change in talkers obvious. This was done so that the exposure-test manipulation was not obvious to participants: Those who underwent the *novel* condition first would not expect that the voices during familiarization were relevant to those at test, and those who underwent the *exposed* condition first could not draw upon their expectations about the task to improve their performance during the second test phase when new voices were encountered. In effect, the *novel* condition served as an unbiased, within-subject baseline for talker identification learning ability during the test phase across tasks and exposure conditions.

Data analysis

The primary outcome variable for this study was talker identification accuracy. Accuracy was calculated in two ways according to the different response demands of the two phases of the study: (1) task accuracy during the exposure phase was identified by the sensitivity index (d'), and (2) talker identification accuracy during the test phase was calculated as the number of trials on which participants correctly identified

the talker out of the total number of trials. We chose to evaluate accuracy for the exposure tasks in terms of *d′* because this measure more effectively captures the dual match/mismatch contingency between participants' responses and the stimuli presented. Furthermore, to measure change in task performance over time, each phase was subdivided into five blocks of equal length, and the corresponding accuracy measurement was calculated separately for each block. Blocks in the exposure phase comprised 40 trials each, while those in the test phase comprised 10 trials each. Participants performed both tasks continuously and were unaware of the blocking structure, which was pertinent only to data analysis.

All data were analyzed in R (Version 4.0.3) using (generalized) linear mixed-effects models implemented in the package *lme4* (Version 1.1.26). A Type-III analysis of variance (ANOVA) using Wald chi-squared tests was used to analyze significance of fixed factors with the packages *lmerTest* (Version 3.1.3) and *car* (Version 3.0.10). Significance was determined based on $\alpha = 0.05$, utilizing the Satterthwaite approximation of the degrees of freedom.

Results

Exposure phase

Performance during the exposure tasks is illustrated in Fig. 2. We submitted participants' *d′* scores on each block of the exposure phase to a linear mixed-effects model with categorical fixed factors of *exposure task* (talker matching vs. talker 1-back vs. verbal 1-back) and *familiarity condition* (exposed vs. novel talkers), as well as *block* (1–5) as a de-meaned, continuous fixed factor. Random factors included by-participant intercepts and slopes for the within-subject fixed factors of *familiarity condition* and *block*. Contrasts on the categorical fixed factors in the model design matrix were coded for successive differences (exposure task) or deviation (familiarity condition).

The ANOVA on this model revealed significant main effects of *exposure task* [$\chi^2(2) = 124.70$, $p \ll .0001$], with highest performance on the talker 1-back task, then the verbal 1-back, and lowest performance on the talker matching task. There was also a main effect of *block* [$\chi^2(1) = 52.39$, $p \ll 0.0001$], though its interpretation varies by condition (see below). There was no main effect of *familiarity condition* [$\chi^2(1) = 0.07$, $p = .79$], confirming that participants' performance during the exposure phase did not differ in a systematic way that might confound measuring the effect of the familiarity condition manipulation during the subsequent talker identification test.

The *exposure task* × *block* interaction was significant [$\chi^2(2) = 81.68$, $p \ll .0001$]. All other two- and three-way interactions were not significant [*exposure task* × *familiarity condition*: $\chi^2(2) = 1.86$, $p = .39$; *familiarity condition* × *block*: $\chi^2(1) = 0.18$, $p = .68$; *exposure task* × *familiarity condition* × *block*: $\chi^2(2) = 1.96$, $p = .38$].

To understand the *exposure task* × *block* interaction, we conducted a series of reduced models separately for each task, with only the continuous fixed factor of *block*. (The absence of any interactions involving the fixed factor *familiarity condition* did not motivate its inclusion in these models.) Random factors included by-participant intercepts and slopes for the within-subject fixed factor *block*. For the talker matching task, there was a significant improvement in participants' accuracy across blocks [$\chi^2(1) = 65.16$, $p \ll .0001$]. However, performance did not increase as a function of *block* in either the talker 1-back task [$\chi^2(1) = 0.0062$, $p = .94$] or the verbal 1-back task [$\chi^2(1) = 1.56$, $p = .21$].

Test phase

Performance during the talker identification test is illustrated in Fig. 3. We submitted participants' accuracy (0 for incorrect, 1 for correct trials) to a generalized linear mixed-effects model for binomial data, with categorical fixed factors of *exposure task* (talker matching vs. talker 1-back vs. verbal 1-back) and *familiarity condition* (exposed vs. novel talkers) and a continuous, de-meaned covariate of *block* (1–5). Random factors

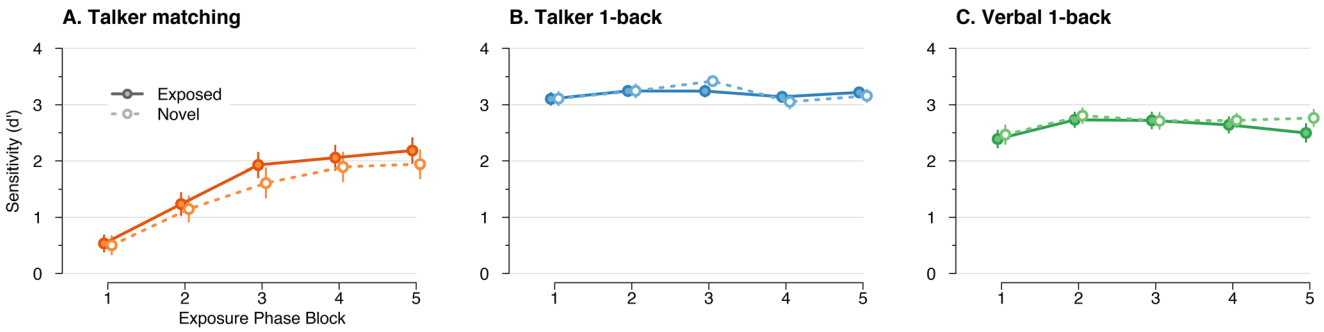

Fig. 2 Mean *d′* across participants for the five blocks (40 trials each) of the exposure phase for the two familiarity conditions (*exposed*: solid lines; *novel*: dashed lines) for each of the three exposure tasks **a** talker-matching, **b** talker 1-back, and **c** verbal 1-back. Error bars show ± 1 standard error of the mean (*SEM*) across participants

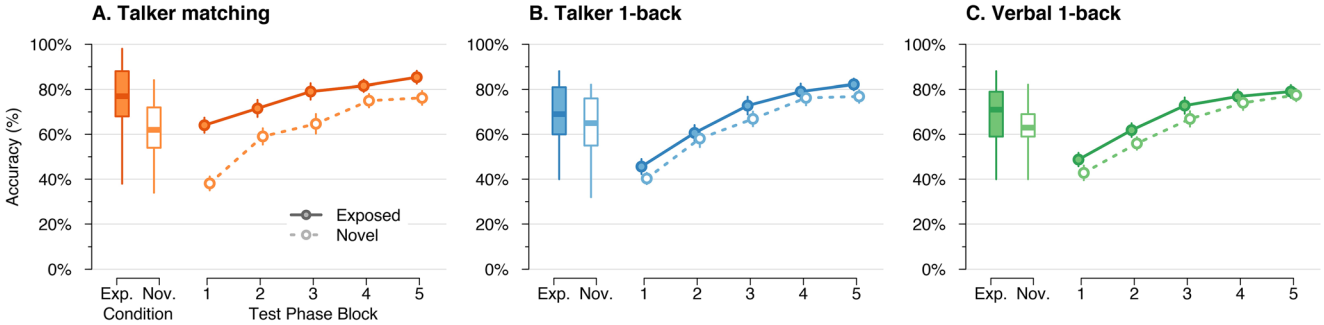

Fig. 3 Mean test accuracy overall (boxplots) and by block (10 trials each) for the test phase for the two familiarity conditions (*exposed*: filled / solid lines; *novel*: open / dashed lines) for each of the three exposure tasks **a** talker-matching, **b** talker 1-back, and **c** verbal 1-back. Error bars show ± 1 standard error of the mean (*SEM*) across participants. Participants were more accurate when learning to identify the voices that they had heard during the exposure phase than the novel voices, but the magnitude of this effect differed across exposure tasks; the greatest effect of exposure came from talker matching, followed by the verbal 1-back task, though the effect of exposure was not significant for the talker 1-back task

included by-participant intercepts and slopes for the within-subject fixed factors of *familiarity condition* and *block*, and random intercepts for each talker. Contrasts on the categorical factors in the model design matrix were coded for successive differences (*exposure task*) or deviation (*familiarity condition*).

The ANOVA on this model revealed significant main effects of *familiarity condition* [$\chi^2(1) = 28.75$, $p \ll .0001$] and *block* [$\chi^2(1) = 377.21$, $p \ll .0001$]. The main effect of *exposure task* was not significant for test accuracy [$\chi^2(2) = 3.78$, $p = .15$].

The *exposure task* × *familiarity condition* interaction was also significant [$\chi^2(2) = 9.97$, $p = .007$]. All other two- and three-way interactions were not significant [*exposure task* × *block:* $\chi^2(2) = 2.27$, $p = .32$; *familiarity condition* × *block*: $\chi^2(1) = 0.34$, $p = .56$; *exposure task* × *familiarity condition* × *block*: $\chi^2(2) = 2.57$, $p = .28$].

To explore the *exposure task* × *familiarity condition* interaction, we conducted a series of reduced models separately for each task, with the categorical fixed factor of *familiarity condition*. (The absence of any interactions involving the fixed factor *block* did not motivate its inclusion in these models.) Random factors included by-participant intercepts and slopes for *familiarity condition*, and random intercepts for each talker. Contrast on the categorical factor *familiarity condition* in the model design matrix was deviation coded.

The results of these models revealed that exposure to the talkers during the exposure phase facilitated talker identification accuracy during the test phase for two of the three tasks. For the talker-matching task, there was a significant effect of *familiarity condition* [$\chi^2(1) = 29.30$, $p \ll .0001$]. For the talker 1-back task, the effect of *familiarity condition* was not significant [$\chi^2(1) = 2.88$, $p = .09$). However, this factor was significant in the verbal 1-back task [$\chi^2(1) = 4.06$, $p < .05$].

Correlations between exposure and test

To determine whether performance during the different exposure tasks was related to participants' subsequent ability to learn to identify the talkers—and whether this relationship was contingent on the consistency in talkers between exposure and test—we correlated participants' *d′* scores in the exposure phase to their mean talker identification accuracy in the test phase (Fig. 4). A rationalized arcsine transformation was applied to test accuracy data (proportion of correct trials by participant by exposure condition) prior to analysis (Studebaker, 1985). Analysis of the correlations between participants' *d′* scores from the exposure phase and talker identification accuracy during the test phase showed there was a significant relationship following the talker-matching task, but not for the talker 1-back or verbal 1-back tasks (Table 2).

To determine whether there were differences in the relationship between exposure phase performance and test phase performance across exposure tasks or familiarity conditions, we analyzed these data using a linear mixed effects model with arcsine-transformed *test accuracy* as the dependent variable. This model included *d′ scores* from the exposure phase as a continuous fixed factor, and the *exposure task* (talker matching, talker 1-back, verbal 1-back) and *familiarity condition* (exposed, novel) as categorical fixed factors. Random factors included by-participant intercepts. (With only one value of the dependent variable per participant per level of the categorical fixed factors, this model cannot accommodate random slopes or item effects). Contrasts on the categorical factors in the model design matrix were coded for successive

Table 2 Correlations between exposure phase performance and talker identification accuracy

	Familiarity condition	
Exposure task	Exposed talkers	Novel talkers
Talker matching	$r_{30} = 0.603$, $p < .0003$ *	$r_{30} = 0.492$, $p < .005$ *
Talker 1-back	$r_{30} = 0.166$, $p = .37$	$r_{30} = -0.051$, $p = .78$
Verbal 1-back	$r_{30} = 0.296$, $p = .10$	$r_{30} = 0.162$, $p = .38$

*Significant after Holm-Bonferroni correction for multiple comparisons

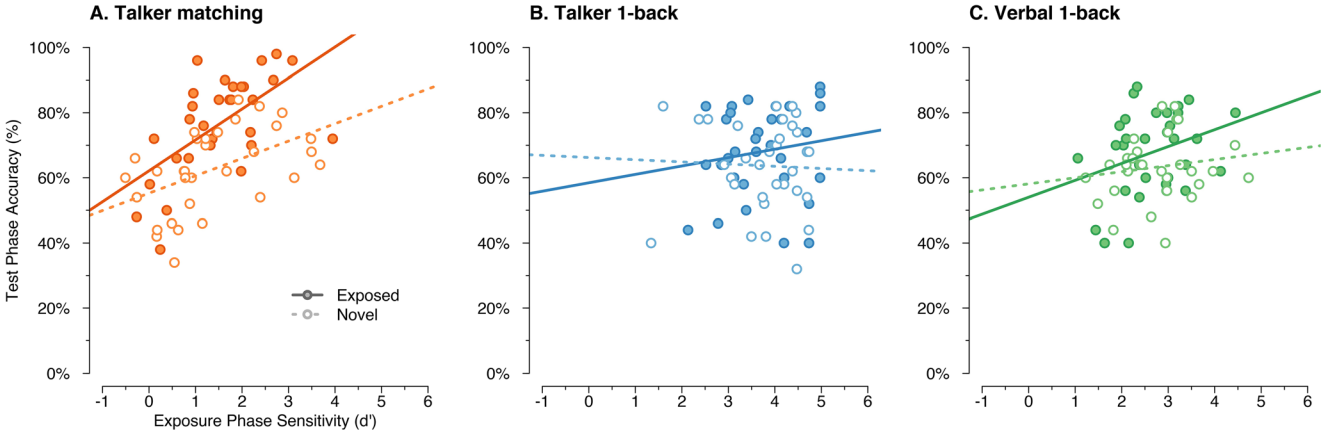

Fig. 4 Correlations between talker identification accuracy during the test phase and *d′* score from the exposure phase for the three exposure tasks

differences (*exposure task*) or deviation (*familiarity condition*).

The ANOVA on this model revealed a significant effect of *d′ score* [$\chi^2(1) = 14.72$, $p < .0002$], such that better performance during the exposure phase was associated with better performance on the talker identification test. There were no significant effects of *familiarity* [$\chi^2(1) = 0.15$, $p = .70$] or *exposure task* [$\chi^2(2) = 0.38$, $p = .83$]. The *d′ score* × *familiarity condition* interaction was significant [$\chi^2(1) = 4.54$, $p = .03$], indicating that the relationship between exposure task performance and talker identification task performance differed depending on whether the talkers at test were familiar or novel. As illustrated by the greater slope for the solid vs. dashed lines in Fig. 4, talker identification performance during the test phase was more positively related to performance during the exposure phase when the same talkers were heard during both phases. Likewise, the *d′ score* × *exposure task* interaction was significant [$\chi^2(2) = 7.63$, $p = .02$]. As reflected in the slopes of the trend lines across panels in Fig. 4, the relationship between exposure task performance (*d′*) and talker identification (test phase accuracy) was greatest when listeners' attention was directed to talkers' identities during exposure (Fig. 4a), less positive when their attention was directed to the verbal content (Fig. 4c), and weakest when they were discriminating talkers' voices (Fig. 4b). The *familiarity condition* × *exposure task* [$\chi^2(2) = 1.39$, $p = .50$] and three-way interactions [$\chi^2(2) = 0.45$, $p = .80$] were not significant.

Discussion

In real-world settings, listeners appear to ascertain talkers' vocal identities through quotidian communicative interactions that prioritize extracting linguistic information, rather than explicitly practicing to identify who is speaking. In this study, we used a carefully controlled laboratory setting to examine how attention to talkers' voices versus the verbal content of their utterances affected listeners' subsequent ability to explicitly identify those talkers. Listeners were familiarized with the sound of talkers' voices through tasks that required varying degrees of explicit attention to talkers' vocal characteristics: an explicit talker matching task, a 1-back voice working memory task, or a 1-back verbal working memory task.

The major finding of this study is that exposure to talkers' voices improves the subsequent ability to learn to identify those voices, even when listeners' initial exposure to the talkers did not explicitly focus their attention on talkers' vocal identities. Perhaps unsurprisingly, this effect is strongest when listeners' initial exposure to the talkers directed their attention to their unique vocal identity (as in the talker matching task). However, while comparatively weaker, this effect was nonetheless still evident when listeners' attention during the exposure phase was directed to the verbal content of the talkers' speech and not their vocal identity (as in the verbal 1-back task).

The finding of a familiarization effect in the verbal 1-back task is important, as this exposure task most closely approximated real-world listening demands by requiring listeners to attend to the verbal content of the speech in each trial. Here, listeners were significantly better at the subsequent talker identification test when those talkers were the same as the ones that they had heard during the exposure task, even though they had not previously needed to explicitly attend to these talkers' vocal identities. This demonstrates that listening to talkers' speech, despite not intentionally attempting to learn what they sound like, nonetheless leads listeners to learn and retain information about those voices for future talker identification. It is possible that these talker-specific representations emerge as a result of talker normalization processes that are engaged automatically during speech perception when listeners deploy additional computational resources to accommodate the potential for ambiguity in the acoustic-phonetic mappings across talkers (e.g., Choi et al., 2018; Kleinschmidt & Jaeger, 2015; Magnuson et al., 2021). In the case of the present experiment, most trials of the exposure phase involved a change in talker from the previous trial,

obligatorily triggering the cognitive processes involved in talker normalization (Choi et al., 2022). As listeners adapted their mapping between the talker-specific stimulus acoustics on each trial and the linguistic representations those acoustics encode, it is likely that they also begin to develop an emergent, generalizable representation of the vocal characteristics of each individual talker, which could then be accessed for the purpose of talker identification.

The idea that internal representations of talker identity are, perhaps in part, derived from normalizing processes during speech perception may also help explain a related phenomenon—the language-familiarity effect in talker identification (Goggin et al., 1991). Listeners may be less able to learn to identify speakers of a foreign-language in part because they are unable to build the talker-specific acoustic-phonetic mappings during speech perception that help contribute to native language talker identification (McLaughlin et al., 2019; Perrachione et al., 2011; Perrachione & Wong, 2007). Moreover, this idea—that when listeners develop talker-specific representations in speech perception they simultaneously form nascent representations of talker identity—also provides new insight into how voice learning can be incorporated into prominent existing models of voice processing (Belin et al., 2004). Such models have primarily attempted to explain the computations and representations involved in a mature, intact voice processing faculty, but without considering how these levels of representation emerge, nor how they can incorporate new information over time. Taking the present results in the context of this model, we note that, during "vocal speech analysis", feedback to lower levels of "voice structure analysis" (i.e., the talker normalization process in speech) may induce simultaneous, feedforward percolation of this information up to a parallel level where the listeners' "voice recognition units" are being tuned to both the vocal and phonetic idiosyncrasies of a particular talker. An open question remains, though, whether this learning requires feedback through a lower level of representation, or whether there may be concurrent connections between higher-order speech and voice processing levels (Perrachione et al., 2010; Perrachione & Wong, 2007).

It is important to note that these representations of talker identity cannot be strictly episodic memories for the words spoken by each talker (cf. Goldinger, 1998), as the stimuli used in the talker identification test were wholly different from those used during the exposure tasks. Instead, it is likely that the perceptual experience of talker identity emerges as an abstraction over multiple levels of representation (Goldinger & Azuma, 2003; Samuel, 2020), leading to effective generalization of voice perception across diverse utterances (Kreiman & Sidtis, 2011). This parallels the phenomenon of talker-specific learning (i.e., the "familiar talker advantage") in speech perception, wherein listeners are effective at generalizing talker-specific representations to novel phonemes and words (Case et al., 2018; Nygaard et al., 1994). Ultimately, this condition provides a laboratory demonstration validating our intuitions about how talkers' vocal identities are learned in real life: as an abstraction over the collective utterances we have heard a person say when listening to their speech (McLaughlin et al., 2015; Perrachione et al., 2011; Scott, 2019).

In contrast, it is perhaps surprising to note that the faciliatory effect of prior exposure to voices on talker identification accuracy appears to have been the smallest when that exposure came from the talker 1-back task (Fig. 3b). One might initially have expected this exposure task to produce an effect intermediate between the talker matching and verbal 1-back tasks because, while it does not ask listeners to learn to individuate the talkers, it does ask them to explicitly attend to—and even briefly remember—the talkers' vocal characteristics. Furthermore, listeners' performance during the talker 1-back task was the best across all exposure tasks (Fig. 2b), suggesting they were particularly sensitive to the acoustic differences among talkers. Why, then, did this task fail to generate a robust familiarization effect? First, the talker 1-back task was effectively a voice discrimination task, in which listeners had to decide whether the talker on the present trial was the same as (or different from) the preceding trial. When applied to voice perception, discrimination tasks may prioritize perceptual decisions based on low-level acoustic features that are stimulus-specific, and which may not generalize well to more abstract representations of talker identity that are generalizable across multiple utterances (Fecher & Johnson, 2018; Lavan et al., 2019; Levi, 2019; Perrachione et al., 2014, 2019; Van Lancker & Kreiman, 1987). Thus, while requiring listeners to attend explicitly to the talkers' voices, this task may have led them to do so in a way that did not lead to generalizable learning of the talkers as distinct auditory objects. Second, because the talker 1-back task was the easiest (i.e., had the highest performance) across the familiarization tasks, it may have been less attentionally demanding, leading listeners to learn less about the talkers, and thus subsequently having the smallest benefit of prior exposure. (Although it is worth noting that lower performance on the talker matching task cannot be disentangled from either its more complex response demands relative to the other two tasks [i.e., 5AFC vs. 2AFC] or from its unique requirement for learning paired associations between visual and auditory stimuli that were consistent across trials, whereas no explicit learning was required for the other two tasks, which were performed strictly in the auditory domain.) It is possible that using more similar-sounding voices in this condition could have, counterintuitively, led to better implicit learning of talker identity, as listeners would have needed to pay attention to more generalizable, as opposed to superficial, vocal features of the talker's utterances. Overall, the relatively weak talker-specific learning induced by this

task further underscores the importance of avoiding tasks that prioritize superficial processing, such as talker discrimination, when making theoretical inferences about how listeners learn to recognize individuals by the sound of their voice in ecological communicative interactions.

While real-world interactions may prioritize attention to the linguistic content of speech, they do not do so exclusively, and the particular behavioral demands on a listener to attend to the content of an utterance vs. who said it may unfold dynamically during conversational interactions. Listeners may, intermittently, find themselves needing to pay attention to talkers' vocal characteristics, whether explicitly or not: For instance, in conversations involving multiple partners, listeners may need to attend to talkers' indexical characteristics in order to attribute the content of speech to the person who said it. Similarly, when trying to listen to one talker in the presence of competing speech, listeners may need to attend to talker-specific indexical features in order to successfully isolate the relevant acoustic information from the background "noise" (Best, Swaminathan, et al., 2018b; Bressler et al., 2014; Kreitewolf et al., 2018). Therefore, an important methodological consideration in this study was to prevent listeners from expecting the future need to identify the voices, so that they would not adopt listening strategies during the exposure phase that divided their attention between talker identity and the target task. We made special effort to ensure that participants could not anticipate the connection between the voices heard during the familiarization tasks and those heard during the explicit identification test in two ways: First, we counterbalanced whether participants heard novel vs. exposed voices during the test phase. We made the change in voices highly salient by using talkers of different sexes during the exposure and test phases for the *novel* condition. Thus, if listeners were in the *novel* condition first, the disconnect between the exposure and test talkers during their first run would not have led them to expect that they needed to pay attention to the talkers' voices during their second run (in the *exposed* condition). Similarly, if listeners were in the *exposed* condition first, any expectations they had about needing to pay attention to the talkers' identities from the first run would be violated when the talkers changed at test. In this way, changing their listening strategy during their second pass through the exposure phase could not have affected their test performance, making the test phase of the *novel* exposure condition a reliable, unbiased baseline across all participants. Statistical analysis of listeners' performance during the exposure phase also indicated no difference between the exposed vs. novel runs (Fig. 2), suggesting that listeners' strategies during this task were not affected by anticipation of needing to know the talkers' identities for the upcoming test. As a second control, we changed the talker-specific avatars at test. This ensured that listeners in the exposed condition of the talker matching task would still have to learn the correspondence between the voices and the avatars at test, increasing the similarity of the task demands between this condition and the other five familiarity-by-task conditions.

Further evidence that exposure to talkers' voices results in implicit learning of talker identity comes from analyzing the correlation between exposure task performance and performance on the explicit talker identification test. Across all exposure conditions, explicit talker identification performance was more positively correlated with performance on the exposure task when the talkers were the same in both phases of the experiment. That is, when listeners had to perform two different tasks with the same voice stimuli, good performance on the exposure task was associated with better performance on the talker identification task, even when participants' performance on the exposure task did not differ across conditions. The lack of an interaction effect in this model suggested that this effect of familiarity holds similarly across all exposure tasks (difference in the slopes of the solid vs. dashed lines in Fig. 4). However, the relationship between exposure performance and test performance clearly differed across tasks, with the strongest relationship observed, intuitively, between performance on talker matching and talker identification. Notably, the strength of the exposure/test performance correlations follow the same pattern across tasks as the magnitude of the effect of voice familiarity on talker identification test performance: greatest for talker matching, less for verbal 1-back, and least for talker 1-back. This result, in which talker discrimination performance is so poorly correlated with the ability to subsequently learn to identify the same talkers, further suggests that performance on a talker discrimination task may be a poor index of a listener's talker identification abilities.

The results of this study show that general findings from the perceptual learning literature for basic stimulus features (e.g., Ahissar & Hochstein, 1993; Banai et al., 2010; Watanabe et al., 2001) also extend to the domain of voice perception, where the convolution of acoustic features related to talker identity and speech articulation is highly complex. Consistent with task-irrelevant learning of various perceptual skills in both vision and audition (e.g., Seitz & Watanabe, 2005; Wright et al., 2010b), we found that implicit learning of the features that are important for talker identity occurs even while listeners are focusing their attention on an orthogonal task. This demonstrates that fundamental psychological principles derived from the perceptual learning literature are likely meaningful in explaining complex, real-world phenomena, such as implicit learning of talker identity during typical communicative interactions that focus on exchanging linguistic messages. Indeed, there is great utility in understanding how general-purpose learning mechanisms can support simultaneous learning of the various, parallel linguistic and paralinguistic demands of speech communication, where generalizing across the complexity and diversity of speech signals may

require a multitude of complementary learning processes acting in parallel (Batterink et al., 2016; Choi et al., 2022; McMurray, 2007; Pierrehumbert, 2003; Romberg & Saffran, 2010; Samuel & Kraljic, 2009).

The ultimate goal of this line of work is to understand how listeners learn and remember talkers' vocal identities as a result of ecological communicative interactions. The present report makes an incremental but important step towards this goal by demonstrating, in a series of carefully controlled laboratory tasks that parametrically manipulated listeners' explicit attention to talkers' vocal characteristics, that exposure to talkers' voices improves listeners' subsequent ability to identify those voices. In particular, listeners were more accurate at subsequent talker identification even if they had previously only been paying attention to the verbal content of the talkers' speech during a cognitively-demanding speech working memory task. By showing that learning of talker identity occurs implicitly and in parallel with speech perception, this study further demonstrates the bidirectional integration of linguistic and indexical information during speech and voice processing (Scott, 2019; Sidtis & Kreiman, 2012).

Acknowledgments We thank Yaminah Carter, Kristina Furbeck, Michelle Njoroge, Jessica Tin, Lauren Gustainis, Cheng Cheng, Ja Young Choi, Alexandra Kapadia, and Deirdre McLaughlin for their assistance with this work and Frank Guenther and Terri Scott for their helpful discussion. This work was supported by the National Institutes of Health under grants R03 DC014045 (to T.P.), R01 DC004545 (to Gerald Kidd) and T32 DC013017 (to Christopher Moore).

References

Ahissar, M., & Hochstein, S. (1993). Attentional control of early perceptual learning. *Proceedings of the National Academy of Sciences of the United States of America, 90*, 5718–5722.

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature, 387*(6631), 401–406.

Ahissar, M., Nahum, M., Nelken, I., & Hochstein, S. (2009). Reverse hierarchies and sensory learning. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 364*, 285–299.

Banai, K., Ortiz, J. A., Oppenheimer, J. D., & Wright, B. A. (2010). Learning two things at once: Differential constraints on the acquisition and consolidation of perceptual learning. *Neuroscience, 165*, 436–444.

Batterink, L. J., Cheng, L. Y., & Paller, K. A. (2016). Neural measures reveal implicit learning during language processing. *Journal of Cognitive Neuroscience, 28*(10), 1636–1649.

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences, 8*, 129–135.

Bent, T., Buchwald, A., & Pisoni, D. B. (2009). Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech. *Journal of the Acoustical Society of America, 126*(5), 2660–2669.

Best, V., Ahlstrom, J. B., Mason, C. R., Roverud, E., Perrachione, T. K., Kidd, G., & Dubno, J. R. (2018a). Talker identification: Effects of masking, hearing loss and age. *Journal of the Acoustical Society of America, 143*, 1085–1092.

Best, V., Swaminathan, J., Kopčo, N., Roverud, E., & Shinn-Cunningham, B. (2018b). A "buildup" of speech intelligibility in listeners with normal hearing and hearing loss. *Trends in Hearing, 22*, 1–11.

Boersma, P. (2001). Praat: A system for doing phonetics by computer. *Glot International, 5*, 341–345.

Bolia, R. S., Nelson, T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitalker communications research. *Journal of the Acoustical Society of America, 124*(6), 1065–1066.

Bregman, M. R., & Creel, S. C. (2014). Gradient language dominance affects talker learning. *Cognition, 130*, 85–95.

Bressler, S., Masud, S., Bharadwaj, H., & Shinn-Cunningham, B. (2014). Bottom-up influences of voice continuity in focusing selective auditory attention. *Psychological Research, 78*(3), 349–360.

Case, J., Seyfarth, S., & Levi, S. V. (2018). Short-term implicit voice-learning leads to a familiar talker advantage: The role of encoding specificity. *Journal of the Acoustical Society of America, 144*(6), 497–502.

Choi, J. Y., Hu, E. R., & Perrachione, T. K. (2018). Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing. *Attention, Perception, & Psychophysics, 80*, 784–797.

Choi, J. Y., Kou, R. S. N., & Perrachione, T. K. (2022). Distinct mechanisms for talker adaptation operate in parallel on different timescales. *Psychonomic Bulletin & Review, 29*, 627–634.

Cook, S., & Wilding, J. (1997). Earwitness testimony: Never mind the variety, hear the length. *Applied Cognitive Psychology, 11*(2), 95–111.

Desimone, R. (1996). Neural mechanisms for visual memory and their role in attention. *Proceedings of the National Academy of Sciences of the United States of America, 93*(24), 13494–13499.

Fecher, N., & Johnson, E. K. (2018). Effects of language experience and task demands on talker recognition by children and adults. *Journal of the Acoustical Society of America, 143*(4), 2409–2418.

Fleming, D., Giordano, B. L., Caldara, R., & Belin, P. (2014). A language-familiarity effect for speaker discrimination without comprehension. *Proceedings of the National Academy of Sciences, 111*(38), 13795–13798.

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J. R., Schweinberger, S. R., Warren, J. D., & Duchaine, B. (2009). Developmental phonagnosia: A selective deficit of vocal identity recognition. *Neuropsychologia, 47*(1), 123–131.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition, 19*(5), 448–458.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251–279.

Goldinger, S. D., & Azuma, T. (2003). Puzzle-solving science: the quixotic quest for units in speech perception. *Journal of Phonetics, 31*(3/4), 305–320.

Goldstein, A. G., Knight, P., Bailis, K., & Conover, J. (1981). Recognition memory for accented and unaccented voices. *Bulletin of the Psychonomic Society, 17*(5), 217–220.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology, 49*, 585–612.

Hollien, H., Majewski, W., & Doherty, E. T. (1982). Perceptual identification of voices under normal, stress and disguise speaking conditions. *Journal of Phonetics, 10*, 139–148.

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.

Karni, A. (1996). The acquisition of perceptual and motor skills: a memory system in the adult human cortex. *Brain Research: Cognitive Brain Research, 5*, 39–48.

Karni, A., & Sagi, D. (1993). The time course of learning a visual skill. *Nature, 365*, 250–252.

Kidd, G., Best, V., & Mason, C. R. (2008). Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *Journal of the Acoustical Society of America, 124*(6), 3793–3802.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*, 148–203.

Kreiman, K., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. Wiley.

Kreitewolf, J., Mathias, S. R., & von Kreigstein, K. (2017). Implicit talker training improves comprehension of auditory speech in noise. *Frontiers in Psychology, 8*(1584), 1–8.

Kreitewolf, J., Mathias, S. R., Trapeau, R., Obleser, J., & Schönwiesner, M. (2018). Perceptual grouping in the cocktail party: Contributions of voice-feature continuity. *Journal of the Acoustical Society of America, 144*(4), 2178–2188.

Latinus, M., & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology, 2*(175), 1–12.

Latinus, M., McAleer, P., Bestelmeyer, P. E. G., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology, 23*, 1075–1080.

Lavan, N., Burston, L. F. K., & Garrido, L. (2018). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology, 110*, 576–593.

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable voice signals. *Psychonomic Bulletin & Review, 26*, 90–102.

Law, C. T., & Gold, J. I. (2008). Neural correlates of perceptual learning in a sensory-motor, but not a sensory, cortical area. *Nature Neuroscience, 11*, 505–513.

Levi, S. V. (2015). Talker familiarity and spoken word recognition in school-age children. *Journal of Child Language, 42*(4), 843–872.

Levi, S. V. (2019). Methodological considerations for interpreting the language familiarity effect in talker processing. *WIREs Cognitive Science, 10*, e1483.

Levi, S. V., Winters, S. J., & Pisoni, D. B. (2011). Effects of cross-language voice training on speech perception: Whose familiar voices are more intelligible? *Journal of the Acoustical Society of America, 130*(6), 4053–4062.

Lim, S. J., Shinn-Cunningham, B. G., & Perrachione, T. K. (2019). Effects of talker continuity and speech rate on auditory working memory. *Attention, Perception, & Psychophysics, 81*, 1167–1177.

Lim, S. J., Carter, Y. D., Njoroge, J. M., Shinn-Cunningham, B. G., & Perrachione, T. K. (2021). Talker discontinuity disrupts attention to speech: Evidence from EEG and pupillometry. *Brain and Language, 221*, 104996.

Luthra, S., Fuhrmeister, P., Molfese, P. J., Guediche, S., Blumstein, S. E., & Myers, E. B. (2019). Brain-behavior relationships in incidental learning of non-native phonetic categories. *Brain and Language, 198*, 1–27.

Magnuson, J. S., Nusbaum, H. C., Akahane-Yamada, R., & Saltzman, D. (2021). *Attention. Perception, & Psychophysics, 83*(4), 1842–1860.

McLaughlin, D. E., Dougherty, S. C., Lember, R. A., & Perrachione, T. K. (2015, August). *Episodic memory for words enhances the language familiarity effect in talker identification.* 18th International Congress of Phonetic Sciences, Glasgow, UK.

McLaughlin, D. E., Carter, Y. D., Cheng, C. C., & Perrachione, T. K. (2019). Hierarchical contributions of linguistic knowledge to talker identification: Phonological versus lexical familiarity. *Attention, Perception, & Psychophysics, 81*, 1088–1107.

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science, 317*(5838), 631.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Psychological Science, 5*(1), 42–46.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Perception & Psychophysics, 60*(3), 355–376.

Orena, A. J., Theodore, R. M., & Polka, L. (2015). Language exposure facilitates talker learning prior to language comprehension, even in adults. *Cognition, 143*, 36–40.

Ortiz, J. A., & Wright, B. A. (2010). Differential rates of consolidation of conceptual and stimulus learning following training on an auditory skill. *Experimental Brain Research, 201*(3), 441–451.

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), 309–328.

Perea, M., Jimenez, M., Suarez-Coalla, P., Fernandex, N., Vina, C., & Cuetos, F. (2014). Ability for voice recognition is a marker for dyslexia in children. *Experimental Psychology, 61*(6), 480–487.

Perrachione, T. K. (2019). Speaker recognition across languages. In S. Frühholz & P. Belin (Eds.), *The Oxford handbook of voice perception*. Oxford University Press https://hdl.handle.net/2144/23877

Perrachione, T. K., & Wong, P. C. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia, 45*(8), 1899–1910.

Perrachione, T. K., Chiao, J. Y., & Wong, P. C. M. (2010). Asymmetric cultural effects on perceptual expertise underlie an own-race bias for voices. *Cognition, 114*, 42–55.

Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. (2011). Human voice recognition depends on language ability. *Science, 333*(6042), 595.

Perrachione, T. K., Stepp, C. E., Hillman, R. E., & Wong, P. C. (2014). Talker identification across source mechanisms: experiments with laryngeal and electrolarynx speech. *Journal of Speech, Language, and Hearing, 57*(5), 1651–1665.

Perrachione, T. K., Furbeck, K. T., & Thurston, E. J. (2019). Acoustic and linguistic factors affecting perceptual dissimilarity judgments of voices. *Journal of the Acoustical Society of America, 146*(5), 3384–3399.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech, 46*(2/3), 115–154.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(6), 906–914.

Samuel, A. G. (2020). Psycholinguists should resist the allure of linguistic units as perceptual units. *Journal of Memory and Language, 111*, 104070.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics, 71*(6), 1207–1218.

Schweinberger, S. R., Herholz, A., & Sommer, W. (1997). Recognizing famous voices: Influence of stimulus duration and different types of retrieval cues. *Journal of Speech, Language, and Hearing Research, 40*(2), 453–463.

Schweinberger, S. R., Kawahara, H., Simpson, A. P., Skuk, V. G., & Zaske, R. (2014). Speaker perception. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(1), 15–25.

Scott, S. K. (2019). From speech and talkers to the social world: The neural processing of human spoken language. *Science, 366*, 58–62.

Seitz, A., & Dinse, H. R. (2007). A common framework for perceptual learning. *Current Opinion in Neurobiology, 17*(2), 148–153.

Seitz, A., & Watanabe, T. (2005). A unified model for perceptual learning. *Trends in Cognitive Sciences, 9*(7), 329–334.

Seitz, A., & Watanabe, T. (2009). The phenomenon of task-irrelevant perceptual learning. *Vision Research, 49*(21), 2604–2610.

Seitz, A., Protopapas, A., Tsushima, Y., Vlahou, E., Gori, S., Grossberg, S., & Watanabe, T. (2010). Unattended exposure to components of speech sounds yields same benefits as explicit auditory training. *Cognition, 115*, 435–443.

Senior, B., Hui, J., & Babel, M. (2018). Liu vs. Liu vs. Luke? Name influence on voice recall. *Applied Psycholinguistics, 39*(6), 1117–1146.

Sidtis, D., & Kreiman, J. (2012). In the beginning was the familiar voice: Personally familiar voices in the evolutionary and contemporary biology of communication. *Integrative Psychological and Behavioral Science, 46*(2), 146–159.

Souza, P., Gehani, N., Wright, R., & McCloy, D. (2013). The advantage of knowing the talker. *Journal of the American Academy of Audiology, 24*(8), 689–700.

Studebaker, G. A. (1985). A "rationalized" arcsine transform. *Journal of Speech, Language, and Hearing Research, 28*(3), 455–462.

Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2014). The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology, 4*(1015), 1–13.

Szpiro, S., Wright, B. A., & Carrasco, M. (2014). Learning one task by interleaving practice with another task. *Vision Research, 101*, 118–124.

Thompson, C. P. (1987). A language effect in voice identification. *Applied Cognitive Psychology, 1*, 121–131.

Tzeng, C. Y., Alexander, J. E., Sidaras, S. K., & Nygaard, L. C. (2016). The role of training structure in perceptual learning of accented speech. *Journal of Experimental Psychology: Human Perception and Performance, 42*(11), 1793–1805.

Van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia, 25*(5), 829–834.

Vlahou, E. L., Protopapas, A., & Seitz, A. R. (2012). Implicit training of nonnative speech stimuli. *Journal of Experimental Psychology: General, 141*(2), 363–381.

Watanabe, T., & Sasaki, Y. (2015). Perceptual learning: Toward a comprehensive theory. *Annual Review of Psychology, 66*, 197–221.

Watanabe, T., Nanez, J. E., & Sasaki, Y. (2001). Perceptual learning without perception. *Nature, 413*, 844–848.

Wester, M. (2012). Talker discrimination across languages. *Speech Communication, 54*, 781–790.

Winters, S. J., Levi, S. V., & Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages. *Journal of the Acoustical Society of America, 126*(6), 4524–4538.

Wright, B. A., & Zhang, Y. (2009). A review of the generalization of auditory learning. *Philosophical Transactions of the Royal Society: B, 364*(1515), 301–311.

Wright, B. A., Wilson, R. M., & Sabin, A. T. (2010a). Generalization lags behind learning on an auditory perceptual task. *Journal of Neuroscience, 30*, 11635–11639.

Wright, B. A., Sabin, A. T., Zhang, Y., Marrone, N., & Fitzgerald, M. B. (2010b). Enabling perceptual learning by alternating practice with sensory stimulation alone. *Journal of Neuroscience, 30*, 12868–12877.

Wright, B. A., Baese-Berk, M., Marrone, N., & Bradlow, A. R. (2015). Enhancing speech learning by combining task practice with periods of stimulus exposure without practice. *Journal of the Acoustical Society of America, 138*, 928–937.

Xie, X., & Myers, E. B. (2015). General language ability predicts talker identification. In: D. C. Noell, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society.

Yarmey, A. D. (1995). Earwitness speaker identification. *Psychology, Public Policy, and Law, 1*(4), 792–816.

Zarate, J. M., Tian, X., Woods, K. J. P., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific Reports, 5*, 11475.