Talker-specificity effects across and within social categories

William Clapp[1], Charlotte Vaughn[2], and Meghan Sumner[1]
[1]Stanford University (USA), [2]University of Maryland (USA)

Listeners are more likely to recognize a word upon second presentation if it is repeated in the same voice as compared to a different voice (e.g., [1, 3]). These now-classic *talker-specificity effects* have served as the foundation for theories central to laboratory phonology in which acoustically rich episodic memories are stored in the lexicon [2, 4]. While this finding has been replicated widely and is considered stable, nearly all talkers and listeners have been white, college-aged Midwesterners. As more recent advances have highlighted the asymmetrical encoding of spoken words based on social information [5], we might reconsider these classic studies through a new lens that centralizes demographic diversity. In newly published work, we investigated the effects of talker diversity on well-established encoding and recognition patterns (Exp. 1), and the effects of perceived standardness and experience on these patterns using two demographically homogeneous talker groups (identifiably Black and white males from California) and self-identified Black and white listener populations (Exp. 2). These experiments were designed to isolate an individual social contrast and measure its influence on memory encoding. In addition to reporting the main results, we supplement the analysis of encoding differences *across* demographic categories by highlighting differences *within* each category (Exp. 1, extended).

All participants were native speakers of American English recruited via Prolific (Exp. 1, N = 727; Exp. 2, N = 680). Replicating past work, we used the continuous recognition memory paradigm with individual words presented in isolation (respond OLD/NEW). Half of OLD words were repeated in the same voice (SAME) and half were repeated in the voice of a different talker (DIFF). In Exp. 1, stimuli were produced by 16 total talkers, and each participant heard 1, 2, 4, 6, or 8 talkers (Number of Voices, NOV). Talkers spanned 8 demographic groups including each combination of Black or white, male or female, and Southern or non-Southern. All voices were rated as identifiable as members of the target demographic categories. In Exp. 2, we used 16 total talkers: all were male, 8 were identifiably Black, and 8 were identifiably white. Each participant heard 8 talkers in one of 3 conditions: 8 Black talkers (B8), 8 white talkers (W8), or 4 talkers from each category (B4W4). Half of listeners self-identified as Black and half as white.

Hits, latency (RT), false alarms, and d' were analyzed, but for brevity we focus here on hits (proportion of correct responses on OLD words). In Exp. 1, listeners were more accurate on SAME than DIFF repetitions (β=0.58, SE=0.026, $p<0.001$). However, as NOV, and therefore diversity of the talker set increased, recognition decreased (β=–0.14, SE=0.033, $p<0.001$; Fig. 1). These findings replicate the classic talker-specificity effect but suggest a role of social information (cf. [3], where there was no effect of NOV). In Exp. 2, the classic effect also replicated robustly (SAME>DIFF; β=0.67, SE=0.021, $p<0.001$), but we additionally found that the perceived race of the talker influenced memory patterns even with decreased talker diversity as compared to Exp. 1 (Fig. 2). Participants were more likely to encode and recognize words produced by white talkers than by Black talkers (β=0.22, SE=0.026, $p<0.001$). This effect held both among the Black and white participant population.

Having demonstrated substantial encoding differences across categories, we report on an extended analysis of asymmetries within demographic categories for Exp. 1. Comparing each pair of talkers from the 8 groups, we found encoding differences in four categories, with a marginal difference in a fifth (Fig. 3). Asymmetries observed within categories, in addition to those observed between members of high-level demographic categories, indicate that listeners are finely attuned to talker and social information in the speech signal and allocate memory resources on an ad-hoc basis. We argue that this process is guided in part by social ideologies, experience, and cultural dynamics. The findings provide new ways forward in the study of episodic memory and illustrate both how cognitive processes are entangled with social realities, and how we might reliably investigate them in our research.

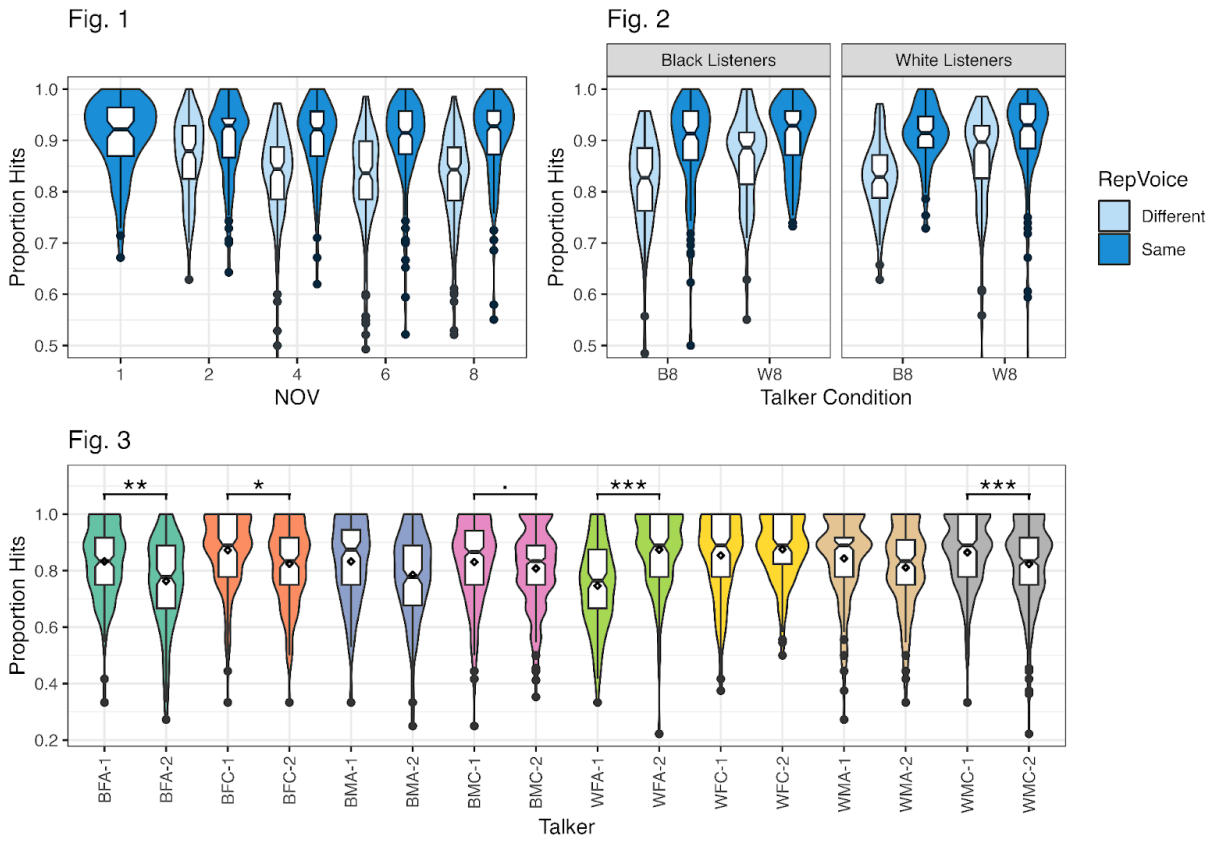

Fig. 1: Exp.1, Accuracy (OLD responses to repeated items) across NOV, where NOV represents the number of unique talkers each participant heard. SAME repetitions in dark blue and DIFF in light blue.
Fig. 2: Exp. 2, Accuracy by talker set, listener population, and RepVoice. Hits for participants who heard 8 Black talkers (B8) or 8 white talkers (W8) and self-identified as either Black or white. SAME repetitions in dark blue and DIFF in light blue.
Fig. 3: Exp. 1, Accuracy within each talker demographic category. Both members of each category are shown in the same color and share a three-letter code, where B = Black and W = white; F = female and M = male; A = Alabama (Southern) and C = California (non-Southern). Significance is shown for pairs where listeners' responses were more accurate for one member of the pair than for the other.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.