INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Attention to indexical information improves voice recall

Grant L. McGuire[1], Molly Babel[2]

[1]Department of Linguistics, University of California, Santa Cruz, USA
[2]Department of Linguistics, University of British Columbia, Canada

gmcguir1@ucsc.edu, molly.babel@ubc.ca

Abstract

In an exposure phase, two groups of listeners were exposed to a set of 10 voices. These groups differed in terms of the task assigned during exposure: one group was asked to make a decision about the regional affiliation of the voices (Indexical Condition), while the other group orthographically transcribed the words presented (Lexical Condition). Both groups were given an identical test phase where they were presented with 20 voices (10 old, 10 new) and asked to make old/new decisions on the voices. While both groups of listeners performed at above chance accuracy levels in recognizing voices at test as old/new, listeners in the Indexical Condition performed more accurately. These results suggest that the nature of attention during exposure has consequences for subsequent performance, suggesting encoding differences as a result of task demands.

Index Terms: Speech Perception, Voice Recognition, Attention, Encoding Strategies

1. Introduction

Convergent evidence suggests that the processing of spoken language involves both detailed and abstracted phonetic memories [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 13, 14]. Indexical information about the talker is often what is claimed to substantiate the detailed phonetic memories. Indexical information appears to serve at least two clear functions in spoken language processing. Listeners need indexical information to calibrate the thresholds for phoneme-level decisions through some kind of normalization process, essentially contextualizing phoneme-boundary thresholds based on talker size, which is related to gender and age (e.g., [15, 16]). Listeners also need indexical information for its social relevance in speech processing, contextualizing phoneme boundary thresholds based on purely social dimensions like dialect (e.g., [17]. A substantive body of the literature supporting the existence of detailed phonetic representations comes from listeners' use of indexical and social knowledge in speech processing tasks [18, 19, 20, 21, 22]. In order to account for these effects, researchers have proposed variations on instance-based or exemplar models [23, 24, 25, 1, 26, 27]. This broad class of models all have in common a mechanism where the processing system has the potential to store a great deal of phonetic information that is tagged or associated with various linguistic and para-linguistic information (e.g. age, class, gender, etc.) The cognitive system then makes higher order generalizations based on this vast store of knowledge. Crucially, such models are not only able to handle linguistic information, but also the associated indexical information.

Exemplar-based models provide a mechanism for attention to and retention of phonetic detail, and while models appeal to "rich phonetic detail", no reasonable model posits a veridical memorization of all incoming acoustic-phonetic information. What then determines the amount of encoded phonetic detail? Currently, the mechanisms that modulate the amount of encoded detail are unclear, though some have argued that talker-specificity effects arise in speech when retrieval processing is slowed by task difficulty [28]. More recently, Theodore and colleagues re-examined this claim using a paradigm that asks listeners to determine whether a word presented in the test phase had been previously presented in the exposure phase. Theodore and colleagues offer the hypothesis that talker-specificity effects in word recognition memory could stem from differences in the process of *encoding* items, rather than retrieving items [29]. Using two voices, they demonstrated specificity effects when listeners encoded single words while categorizing talker gender, but not when told to "listen carefully to each word" [29]. The results of [29] indicate that attending to indexical information by virtue of categorizing talker gender, listeners encode word-level information in a talker-specific fashion than when passively listening. When asked to categorize the syntactic category of the word, listeners show higher levels of recognition memory for words than in a talker-focused condition. As what differed across these experiments was not the test phase, but what listeners were tasked with in the exposure phase, these results suggests that the nature of the listening affects the substance of phonetic memories. Thus, differences in encoding strategy may contribute to how the items are stored and what kind of phonetic information is subsequently available when listeners are accessing word-level memories.

The work by Theodore and colleagues' focuses on word recognition, while we shift our focus to voice recognition. How do different attentional demands at exposure affect subsequent voice recognition decisions? The focus on voice recognition will eventually allow us to test the directionality of indexical and linguistic decision processes. As shown in [29], attention in exposure must be directed towards indexical information for talker-specific effects to emerge in subsequent word recognition. Does attention to word information attenuate sensitivity to talker-specific indexical information in subsequent voice recognition?

Crucially, just as words are not all equal in speech processing (see review in [30]), not all voices are equally memorable and a wide-range of talker- and listener-based factors affect voice recognition performance. For example, listeners are more accurate at identifying and recognizing voices of native or familiar languages (e.g., [31, 32, 33, 34, 35]. Within an accent or dialect, distinctiveness is predictive of voice recognition, such that more distinctive voices are recalled with higher accuracy [36, 37, 38, 39, 40].

We provide an initial inquiry into the role of encoding strategy on voice recognition by presenting listeners with exposure tasks that elicited different encoding strategies in a between-subjects design: in an Indexical Condition that focuses attention in exposure on detecting regional accents and in Lexical Condition that focuses attention on linguistic content. For the former,

http://dx.doi.org/10.21437/Interspeech.2020-3042

we asked listeners to rate the regional affiliation of the voices – specifically whether each voice is from California or not. In this way the listeners, all natives of California, would be making an in-group / out-group judgment based, presumably, on a suite of acoustic-phonetic on the features of the voice including, for example, vowel quality and voice quality. In the latter condition listeners performed a lexical identification task. The Lexical Condition promotes an attentional approach that downgrades the importance of talker-specific information in favor of generalized linguistic information related to lexical contrasts.

2. Methods

The design is a voice recall task where listeners are first presented with a subset of the voices in an *exposure phase* and later presented with the full set of voices in the *test phase*. The exposure phase varies by condition: an Indexical Condition is designed to direct attention to talker-specific indexical features and the Lexical Condition focuses attention on words. We analyze these data as a between-subjects design.

2.1. Participants

Subject ($n = 54$; 36 female, 18 male) were divided between the two conditions (23 Lexical, 31 Indexical). All reported dominance in English and have lived in California and spoken English since the age of five. The participants were recruited from the undergraduate population at the University of California, Santa Cruz and were compensated with partial course credit.

2.2. Stimuli

The stimuli are a subset of those used in [41]. Briefly, the full set consists of 60 male and female voices each producing the following monosyllabic words: *boot, cot, deed, hoop, key, peel, pod, sock, teal, toot, tot*, and *zoo*. These voices have been rated for several measures, notably gender stereotypicality [41]. This was a subjective measure where listeners used a Likert scale to evaluate how stereotypical each voice was. Given the role of distinctiveness and typicality in voice memory, the measure of stereotypicality is used to narrow the voice set and divide them for counterbalancing purposes as follows. Because voice distinctiveness has been shown to factor in voice recall tasks [37, 40], we separately ranked the male and female voices by stereotypicality using the data provided by [41]. The 10 most and 10 least stereotypical male and female voices were removed leaving 10 male and 10 female voices. For counterbalancing purposes, each of these sets was split in half such that the odd numbered ranked voices were assigned to one set and the even numbered ones in the other. The word list was similarly divided into two. All combinations of voice list and word list were used and counterbalanced across listener.

2.3. Procedure Overview

All experiments consisted of an exposure task where 10 voices were presented in randomized order, followed immediately by a voice recall (new/old voice decision) task with all 20 voices, also randomized. Subjects were given instructions for completing the exposure task and only told that there would be a second task and instructions would be given after the first task. Participants were seated at a computer workstation having both a keyboard and 5-button response box located in a sound-attenuated booth. Up to three participants were run at one time. The entire procedure lasted no more than 30 minutes.

2.3.1. Exposure Tasks

In the Indexical task, listeners were told they would hear 10 voices and that they were to rate each voice on a scale from 1-9 based on how likely it was that the talker was from California, where 1 = "very unlikely to be from California" and 9 = "definitely from California". They were further instructed to use the whole scale. The voices were presented in randomized order and six productions from each voice were used, randomized within voice. Listeners were prompted to rate each voice after all six words were presented. Words were presented with 500 ms of silence between each production and subjects were not time limited in making their rating.

In the Lexical Condition, listeners were presented with the identical set of voices and words as the Indexical Condition, but listeners were instructed to listen to each word produced by the talker, typing each word as it was presented. The words were blocked by talker and randomized therein. The talker list was fully randomized as well.

2.3.2. Testing

The task at test was the same for listeners assigned to the Indexical and Lexical Conditions. At test, all 20 voices were presented in a randomized order and for each voice the same six words (physically identical to those in the exposure phase) were presented in randomized order. The repetition of items across exposure and test was to prevent floor effects in this challenging task. After all six words were presented, the subject was prompted to respond using the response box as to whether the voice was one they had heard previously during the exposure task (Old) or one they had not heard before (New). There was no time limit for this decision and subjects could not advance until they responded.

3. Results

While performance at test is of primary interest in this experiment, we first describe participant behaviour in the exposure phases.

3.1. Exposure Performance

3.1.1. California Rating

The number of voices from California in the talker set is not relevant to this assessment. Here, we seek to confirm that listeners, who were all Californian, behaved in an acceptably uniform fashion based on inter-rater reliability.

Listeners assigned to the Indexical Condition were asked to rate the perceived Californianness of each voice (where 1 = "very unlikely to be from California" and 9 = "definitely from California".) Due to a coding error, listeners could respond with integers above 9; all such responses were removed from the data ($< 3\%$ of the total.) The mean rating for all talkers was 5.28 (sd = 2.33). Mean ratings for talkers spanned from a maximum of 7.73 (sd = 1.85) to a minimum of 3.22 (sd = 2.15). As a measure of inter-rater reliability, Cronbach's alpha was calculated for each of the two counter-balanced talker exposure groups, the "odd" and "even" lists as described in the Stimuli section. They showed a moderate [odd list; $\alpha = 0.72$] and high [even list; $\alpha = 0.81$] degree of inter-rater reliability.

3.1.2. Transcription Accuracy

Listeners assigned to the Lexical Condition were asked to orthographically transcribe the items in the exposure condition. Transcriptions were scored as correct only when they exactly matched the exposure item with the following exceptions: 'taught' and 'taut' were scored as accurate for *tot*, 'caught' for cot, and 'peal' for *peel*.

Overall, subjects performed well; the mean proportion correct was 0.88 (s.d. = 0.28). Proportions correct were similar across items with the lowest mean values for *cot* and *deed*, at 0.85. By-talker values were similar with all above 0.82. No subject had an overall proportion correct below 0.88. The overall high accuracy suggest that the subjects attended closely to the exposure task and that no talkers or items were especially difficult to perceive.

The fact that listeners were not at ceiling in performance on this task may be influenced by the conservative scoring of spelling errors as incorrect.

3.2. Test Phase: Voice Recognition

Figure 1 shows the proportion correct results for both conditions. The overall proportion correct was above chance for both groups, with mean by-listener performance being quantitatively higher in the Indexical Condition [Lexical Condition $M = 0.61, t(20) = 4.33, p < 0.001$; Indexical Condition: $M = 0.71, t(32) = 12.1, p < 0.001$] . We do not analyze response times in the test phase, as the collected response times were unreliable due to high variance resulting from the experiment design: listeners heard a string of six items and could log their response at any point and were not under any time pressure. The mean response time was 1.4 seconds (s.d. = 1.76 s). Response times are not discussed further.

3.2.1. Proportion Correct

To model listeners' decisions regarding the voices, a mixed effects logistic regression model was used to analyze the probability that listeners could correctly identify the voices as Old or New. Exposure Condition (Indexical, Lexical) and Voice Status (New, Old) were both contrast coded and entered as a fixed effects with random slopes for the fixed effects and Subject, and random intercepts for Subject. The intercept was significant [$B = 0.70, SE = 0.07, z = 10.09, p < 0.001$]. There was an effect of Exposure Condition [$B = 0.48, SE = 0.14, z = 3.43, p < 0.001$], thus confirming the difference presented in Figure 1. Neither Voice Status nor its interaction with Exposure Condition were significant.

3.2.2. Sensitivity and Bias

Because the model above does not fully account for listeners' response biases, two more models were constructed using the signal detection theoretic measures *d'* and bias (criterion). For these analyses correct identification of a voice as Old was coded as a Hit and incorrect identification of a New voice as Old was coded as a False Alarm. This coding results in bias results where a negative value indicates a bias to respond Old and a positive value indicates a bias to label a voice as New. These analyses necessitated aggregated data by listener, which were fit to a simple linear regression model with Exposure Condition as the predictor variable. For the sensitivity (*d'*) analysis both the intercept [$B = 0.60, SE = 0.13, t = 4.62, p < 0.001$] and Exposure Condition (Indexical) [$B = 0.56, SE = 0.17, t = 3.34, p < 0.01$] were significant. The *d'* results are shown in the left panel of Figure 2.

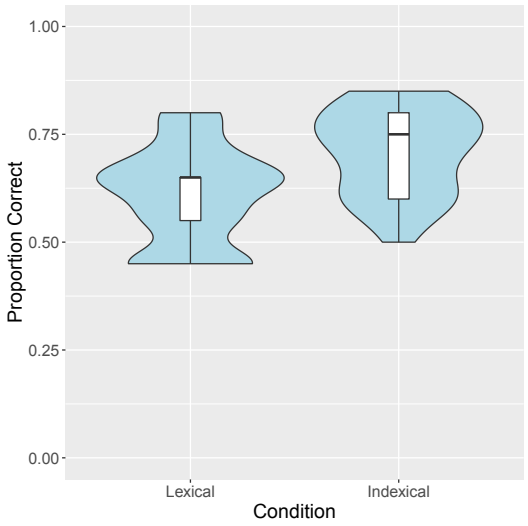

Figure 1: *Proportion correct identification of Old and New voices at test for the Lexical and Indexical Conditions.*

A model similar to that for the *d'* values was fit for the bias values. Both the model intercept [$B = -0.17, SE = 0.07, t = -2.61, p < 0.05$] and Exposure Condition (Indexical) [$B = 0.24, SE = 0.09, t = 2.84, p < 0.01$] were significant. The effect of Exposure Condition for bias is presented in the right panel of Figure 2. Two t-tests were run to determine whether the bias distributions differed significantly from zero, which indicates an unbiased response pattern. Only the test for the Lexical group was significant $t(20) = -2.16, p < 0.05$, indicating that participants in the Lexical condition were biased to respond Old, while the Indexical condition subjects showed no significant bias.

4. Discussion

The results of the accuracy and *d'* data indicate that listeners are better at recognizing voices as either Old or New when instructed to rate the regional affiliation of the voice than when asked to attend to the lexical content. These conditions presented listeners with a different task at exposure, but gave the same exercise at test. As such, these data suggest that the nature of the attention given to spoken language affects what is encoded, supporting the results of [29].

The fact that listeners in the Lexical Condition were above chance in the voice recognition memory task demonstrates that talker-specific information is accessible, even when not explicitly attended to, for subsequent decisions, supporting classic work like [12]. [12] argue that indexical voice-specific information is encoded automatically and not strategically. While our data cannot adjudicate the automaticity claim, these data and those of [29] do suggest that while *strategy* may be an overly agentive term, sensitivity to talker-specific information is affected by the task undertaken during exposure. As noted in the Introduction, indexical information is inherently fundamental to spoken language processing as it serves (at least) two crucial roles in linguistic decision-making. Listeners need indexical information to calibrate decisions related to inherent and acquired social dimensions like size, gender, age, and regional accent. It

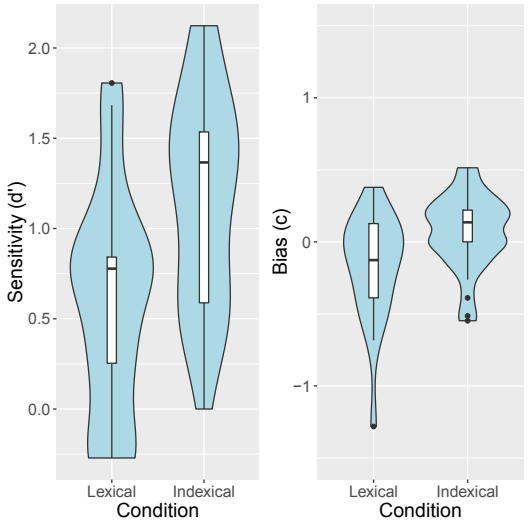

Figure 2: *Sensitivity (d') in the left panel and bias (c) in the right panel violin plots for each exposure condition. For the bias calculation, a negative value indicates a bias to respond Old, and a positive value indicates a bias to respond New.*

may be worth noting that the exposure task we had pushed on social relevance, given the nature of the assigned task.

These results demonstrate that auditory exposure to items does not guarantee equivalent performance at test. Listeners attending to indexical information performed better at the voice recall test. Two explanations present themselves. More phonetic detail may be encoded when attending to an indexical signal than when listeners are tasked with attending to predictable linguistic content. This increase in encoded detail means that at test listeners who attended to the indexical information have more phonetic detail to tap when accessing information relevant to recognizing a voice as New or Old. A second possibility is that by virtue of attending to indexical information, it is not that listeners encoded *more* phonetic detail, but rather they encoded different information that is subsequently more useful in the later memory voice recall task. This makes the distinction at encoding one of quality and not one of quantity of the memory trace.

The notion of encoding strategies is not unique. A parallel may be found in considering the attentional set adopted by a listener. Attentional sets in speech perception have been claimed to come in two broad types: comprehension- and perception-oriented [42]. Comprehension-oriented or diffuse attentional sets are in use when listeners are intent on the larger linguistic or communicative message. Focused or perception-oriented attentional sets are in use when listeners are focused on signal properties. Different kinds of paradigms naturally elicit these attentional sets. Tasks focused on recognition and lexical levels (e.g., transcription tasks, lexical decision tasks, etc.) elicit comprehension-oriented attentional sets, whereas paradigms that draw attention to the speech signal (e.g., phoneme monitoring, talker gender identification, etc.) engage perception-oriented attentional sets [42, 43, 44, 45, 46]. Without explicitly using these terms and concepts, scholars testing episodic memory in speech have often exploited attentional sets, demonstrating that directing listeners' attention to signal-properties (e.g., talker gender) leads to improved performance on subsequent talker-related tasks.

In interpreting the signal detection theoretic measures, listeners are more sensitive to the difference between Old and New voices when tasked with an indexical decision during the exposure phase, having adjusted their threshold for the decision. Listeners in the Lexical Condition were more biased to respond Old. This suggests that listeners who engaged in an indexical encoding strategy may have improved their ability to tell voices apart, while those exposed to the voices in a lexically-biased task were confusing voices together (e.g., assuming "sameness" by assuming all voices are old and previously heard).

Research in the domain of voice memory has demonstrated that talker-specific vocal traits – more specifically, the interaction of talker and listener traits like accent familiarity – affect voice recognition performance (e.g., [32, 47, 48]. That variation by task affects the kind of encoding during exposure indicates that listeners may have some amount of control in the process of voice recall. In other words, these results indicate that listeners' listening strategies influence the nature of phonetic memories.

5. Conclusion

Ultimately, what these results demonstrate is that by virtue of simply being exposed to the same speech samples, listeners do not have equivalent performance at test. In line with the explanatory mechanisms of [29], we suggest that the nature of the attention at exposure determines the encoding strategy deployed by the listener.

6. Acknowledgments

This project draws on research supported by the Social Sciences and Humanities Research Council of Canada (SSHRC) and the Humanities Division of UC Santa Cruz. We are grateful to Amanda Choe for help with subject running.

7. References

[1] J. B. Pierrehumbert, "2002. word-specific phonetics," in *Laboratory phonology VII, 101–140. Berlin: Mouton de Gruyter*. Citeseer.

[2] J. S. German, K. Carlson, and J. B. Pierrehumbert, "Reassignment of consonant allophones in rapid dialect acquisition," *Journal of Phonetics*, vol. 41, no. 3-4, pp. 228–248, 2013.

[3] M. Sumner, S. K. Kim, E. King, and K. B. McGowan, "The socially weighted encoding of spoken words: A dual-route approach to speech perception," *Frontiers in psychology*, vol. 4, p. 1015, 2014.

[4] J. S. Allen and J. L. Miller, "Listener sensitivity to individual talker differences in voice-onset-time," *The Journal of the Acoustical Society of America*, vol. 115, no. 6, pp. 3171–3183, 2004.

[5] D. Dahan, S. J. Drucker, and R. A. Scarborough, "Talker adaptation in speech perception: Adjusting the signal or the representations?" *Cognition*, vol. 108, no. 3, pp. 710–718, 2008.

[6] F. Eisner and J. M. McQueen, "The specificity of perceptual learning in speech processing," *Perception psychophysics*, vol. 67, no. 2, pp. 224–238, 2005.

[7] S. D. Goldinger, D. B. Pisoni, and J. S. Logan, "On the nature of talker variability effects on recall of spoken word lists." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 17, no. 1, p. 152, 1991.

[8] K. Johnson, "The role of perceived speaker identity in f 0 normalization of vowels," *The Journal of the Acoustical Society of America*, vol. 88, no. 2, pp. 642–654, 1990.

[9] T. Kraljic, S. E. Brennan, and A. G. Samuel, "Accommodating variation: Dialects, idiolects, and speech processing," *Cognition*, vol. 107, no. 1, pp. 54–81, 2008.

[10] T. Kraljic and A. G. Samuel, "Perceptual adjustments to multiple speakers," *Journal of Memory and Language*, vol. 56, no. 1, pp. 1–15, 2007.

[11] J. M. McQueen, A. Cutler, and D. Norris, "Phonological abstraction in the mental lexicon," *Cognitive science*, vol. 30, no. 6, pp. 1113–1126, 2006.

[12] T. J. Palmeri, S. D. Goldinger, and D. B. Pisoni, "Episodic encoding of voice attributes and recognition memory for spoken words." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 19, no. 2, p. 309, 1993.

[13] R. M. Theodore and J. L. Miller, "Characteristics of listener sensitivity to talker-specific phonetic detail," *The Journal of the Acoustical Society of America*, vol. 128, no. 4, pp. 2090–2099, 2010.

[14] D. L. Schacter and B. A. Church, "Auditory priming: Implicit and explicit memory for words and voices." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 18, no. 5, p. 915, 1992.

[15] B. Munson and M. Babel, "The phonetics of sex and gender," in *The Routledge Handbook of Phonetics*. Taylor and Francis, 2019, pp. 499–525.

[16] S. Barreda and T. M. Nearey, "The direct and indirect roles of fundamental frequency in vowel perception," *The Journal of the Acoustical Society of America*, vol. 131, no. 1, pp. 466–477, 2012.

[17] T. Kendall and V. Fridland, "Variation in perception and production of mid front vowels in the us southern vowel shift," *Journal of Phonetics*, vol. 40, no. 2, pp. 289–306, 2012.

[18] P. Ladefoged and D. E. Broadbent, "Information conveyed by vowels," *The Journal of the acoustical society of America*, vol. 29, no. 1, pp. 98–104, 1957.

[19] E. A. Strand and K. Johnson, "Gradient and visual speaker normalization in the perception of fricatives." in *KONVENS*, 1996, pp. 14–26.

[20] K. Johnson, E. A. Strand, and M. D'Imperio, "Auditory–visual integration of talker gender in vowel perception," *Journal of phonetics*, vol. 27, no. 4, pp. 359–384, 1999.

[21] J. Hay, P. Warren, and K. Drager, "Factors influencing speech perception in the context of a merger-in-progress," *Journal of phonetics*, vol. 34, no. 4, pp. 458–484, 2006.

[22] B. Munson, S. V. Jefferson, and E. C. McDonald, "The influence of perceived sexual orientation on fricative identification," *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2427–2437, 2006.

[23] K. Johnson and J. W. Mullennix, *Talker variability in speech processing*. Morgan Kaufmann Publishers Inc., 1997.

[24] S. D. Goldinger, "Echoes of echoes? an episodic theory of lexical access." *Psychological review*, vol. 105, no. 2, p. 251, 1998.

[25] J. B. Pierrehumbert, "Stochastic phonology," *Glot international*, vol. 5, no. 6, pp. 195–207, 2001.

[26] K. Johnson, "Resonance in an exemplar-based lexicon: The emergence of social identity and phonology," *Journal of phonetics*, vol. 34, no. 4, pp. 485–499, 2006.

[27] A. Walker and J. Hay, "Congruence between '91word age'92and '91voice age'92facilitates lexical access," *Laboratory Phonology*, vol. 2, no. 1, pp. 219–237, 2011.

[28] C. T. McLennan and P. A. Luce, "Examining the time course of indexical specificity effects in spoken word recognition." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 31, no. 2, p. 306, 2005.

[29] R. M. Theodore, S. E. Blumstein, and S. Luthra, "Attention modulates specificity effects in spoken word recognition: Challenges to the time-course hypothesis," *Attention, Perception, Psychophysics*, vol. 77, no. 5, pp. 1674–1684, 2015.

[30] M. S. Vitevitch and P. A. Luce, "Phonological neighborhood effects in spoken word perception and production," *Annual Review of Linguistics*, vol. 2, pp. 75–94, 2016.

[31] C. P. Thompson, "A language effect in voice identification," *Applied Cognitive Psychology*, vol. 1, no. 2, pp. 121–131, 1987.

[32] J. P. Goggin, C. P. Thompson, G. Strube, and L. R. Simental, "The role of language familiarity in voice identification," *Memory cognition*, vol. 19, no. 5, pp. 448–458, 1991.

[33] S. J. Winters, S. V. Levi, and D. B. Pisoni, "Identification and discrimination of bilingual talkers across languages," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4524–4538, 2008.

[34] T. K. Perrachione and P. C. Wong, "Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex," *Neuropsychologia*, vol. 45, no. 8, pp. 1899–1910, 2007.

[35] B. Senior, J. Hui, and M. Babel, "Liu vs. liu vs. luke? name influence on voice recall," *Applied Psycholinguistics*, vol. 39, no. 6, pp. 1117–1146, 2018.

[36] G. Papcun, J. Kreiman, and A. Davis, "Long-term memory for unfamiliar voices," *The Journal of the Acoustical Society of America*, vol. 85, no. 2, pp. 913–925, 1989.

[37] J. Kreiman and G. Papcun, "Comparing discrimination and recognition of unfamiliar voices," *Speech Communication*, vol. 10, no. 3, pp. 265–275, 1991.

[38] A. D. Yarmey, "Descriptions of distinctive and non-distinctive voices over time," *Journal of the Forensic Science Society*, vol. 31, no. 4, pp. 421–428, 1991.

[39] T. L. Orchard and A. D. Yarmey, "The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification," *Applied Cognitive Psychology*, vol. 9, no. 3, pp. 249–260, 1995.

[40] J. W. Mullennix, A. Ross, C. Smith, K. Kuykendall, J. Conard, and S. Barb, "Typicality effects on memory for voice: Implications for earwitness testimony," *Applied Cognitive Psychology*, vol. 25, no. 1, pp. 29–34, 2011.

[41] M. Babel and G. McGuire, "Perceptual fluency and judgments of vocal aesthetics and stereotypicality," *Cognitive science*, vol. 39, no. 4, pp. 766–787, 2015.

[42] A. Cutler, J. Mehler, D. Norris, and J. Segui, "Phoneme identification and the lexicon," *Cognitive Psychology*, vol. 19, no. 2, pp. 141–177, 1987.

[43] D. Norris and A. Cutler, "The relative accessibility of phonemes and syllables," *Perception Psychophysics*, vol. 43, no. 6, pp. 541–550, 1988.

[44] M. A. Pitt and A. G. Samuel, "Attentional allocation during speech perception: How fine is the focus?" *Journal of Memory and Language*, vol. 29, no. 5, pp. 611–632, 1990.

[45] M. A. Pitt and C. M. Szostak, "A lexically biased attentional set compensates for variable speech quality caused by pronunciation variation," *Language and Cognitive Processes*, vol. 27, no. 7-8, pp. 1225–1239, 2012.

[46] M. McAuliffe and M. Babel, "Stimulus-directed attention attenuates lexically-guided perceptual learning," *The Journal of the Acoustical Society of America*, vol. 140, no. 3, pp. 1727–1738, 2016.

[47] S. V. Levi, S. J. Winters, and D. B. Pisoni, "Effects of cross-language voice training on speech perception: Whose familiar voices are more intelligible?" *The Journal of the Acoustical Society of America*, vol. 130, no. 6, pp. 4053–4062, 2011.

[48] T. K. Perrachione, J. Y. Chiao, and P. C. Wong, "Asymmetric cultural effects on perceptual expertise underlie an own-race bias for voices," *Cognition*, vol. 114, no. 1, pp. 42–55, 2010.