Who said that? Voice Recognition Across Changes in Language and Emotional Affect

Melissa Paquette-Smith (paquettesmith@psych.ucla.edu)[1], Elizabeth K. Johnson[2], Melody Yuan[1] & Idan A. Blank[1]

[1] Department of Psychology, University of California, Los Angeles

[2] Department of Psychology, University of Toronto Mississauga

Abstract

Voice recognition depends on listeners' ability to retain and identify voices despite variability in the speech signal. Here, we examined the relative impact of changes in the speaker's language and changes in their emotional state on voice recognition. Participants were exposed to a speaker and later asked to identify that speaker from a lineup of voices. Across trials, exposure and test were either matched or mismatched in language and emotion, introducing variability in the speaker's phonological and acoustic characteristics. This design yielded four conditions: no-mismatch, language-mismatch, emotion-mismatch, and complete-mismatch. Overall, switching emotion impaired voice recognition more than switching languages, especially when the exposure took place in an unfamiliar language and a neutral tone of voice, demonstrating that the magnitude of these effects is context dependent. A tentative explanation of these results is proposed in the discussion.

Keywords: voice recognition; language familiarity; emotional cues; tone of voice; cue mismatch; speaker identity

Introduction

The ability to remember and recognize voices plays an essential role in social communication. When someone speaks, their voice carries cues about the speaker's identity, including relatively stable characteristics such as habitual patterns in the realization of particular phonemes, supra-segmental features, and words, as well as indexical cues, such as age and gender. Although some cues are relatively stable, voices can also vary dynamically in ways that alter the properties of the speech signal without constituting a stable marker of identity. For example, both a change in a speaker's language and a change in their emotional affect can alter a variety of acoustic and/or linguistic speech characteristics, including their production of phonemes, pitch, and voice quality. Successful voice recognition therefore requires listeners to extract speaker-specific information while tolerating variability introduced by more transient factors.

A number of studies have examined how specific sources of information (e.g., language familiarity, emotion), in isolation, influence people's ability to recognize voices. These studies have found, for example, that voice recognition is impaired when the listener lacks experience with the speaker's language or dialect (Perrachione et al., 2010, 2011; Yu et al., 2021) or when the speaker's emotional state changes (Winters et al., 2008). However, it is not clear how listeners contend with simultaneous variability along multiple dimensions, and how they integrate and weigh information from different sources. In particular, it is not well understood how changes in the language spoken and changes in the speaker's emotional affect interact to impact voice recognition. Addressing this question requires directly pitting transformations of the speech signal against one another, rather than examining them in isolation. The current study therefore exposed participants to voices speaking a specific language (familiar: English, or unfamiliar: Urdu) and conveying a specific emotion (angry or neutral), and then tested voice recognition under the same or a different language and the same or a different emotion.

The Impact of Language on Voice Recognition

A substantial body of research has demonstrated that listeners are better at recognizing voices when they speak in a familiar language (Fecher & Johnson, 2018; Goggin et al., 1991; Perrachione et al., 2010; Thompson, 1987; Yu et al., 2021). This effect, referred to as the "language familiarity effect", suggests that listeners are using language-specific knowledge to help them identify voices. This knowledge may include: 1) phonological knowledge, supported by evidence that individuals with dyslexia (who typically exhibit reduced phonological awareness) perform worse than neurotypical individuals on familiar-language voice recognition tasks (Perrachione et al., 2011); and 2) lexical knowledge, supported by evidence that voice recognition decreases when speakers produce nonwords instead of words (Abu et al., 2022).

The benefit of phonological and lexical information likely depends on listeners' access to these cues during both encoding and retrieval. Accordingly, the impact of language on voice recognition depends not only on familiarity, but also on the relationship between the exposure and test languages. Recognition seems to be hindered when these languages mismatch, at least when the exposure language is familiar to the listener. For example, when English-speaking participants were familiarized with English voices, they recognized those voices better when they were tested in English compared to when they were tested in German. In contrast, when listeners were familiarized with voices speaking German (an unfamiliar language), they showed similar recognition performance regardless of the test language (Winters et al., 2008). Therefore, speaker representations formed during familiarization are partly

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

language-dependent: when encoding occurs in a familiar language, listeners benefit from access to language-specific information during recognition, but this advantage is diminished when the test language differs (see also Wester, 2012).

The Impact of Emotion on Voice Recognition

Although voice recognition is more challenging when cues to phonemic and lexical realization are impoverished or uninformative, such as when talkers speak an unfamiliar language or when speech is reversed (van Lancker et al., 1985), listeners can still successfully identify voices under these conditions. This ability suggests that listeners can, in some circumstances, rely primarily on other cues to recognize speakers. These "indexical" cues include relatively stable acoustic and phonological (suprasegmental) characteristics of the voice, such as those correlated with age or gender; they also encompass aspects of the voice that happen to be stable between exposure and test but can, in principle, dynamically vary within a speaker across time and contexts. One particularly important dynamic indexical cue is emotional affect (or tone of voice), which can impact the physical characteristics of speech, including pitch, intensity, and prosody. Although many voice recognition studies do not test the effect of within-speaker stability vs. variability, because they only present voices speaking scripted sentences in a neutral affect, such variability can substantially impact recognition (Lavan et al., 2019; Orchard & Yarmey, 1995). For example, even though emotional expression can enhance the salience or memorability of a voice under some conditions (Kim et al., 2019), changes in emotion between exposure and test, much like changes in language, can disrupt voice recognition (Read & Craik, 1995; Saslove & Yarmey, 1980; Stevenage & Neil, 2014; Xu & Armony, 2021). Thus, mental representations of voices include such indexical cues, and show only limited tolerance to transformations of this information.

The Current Study

Taken together, listeners' ability to recognize voices relies on phonemic and lexical knowledge, as well as on indexical cues, such as acoustic features, that are available in the speech signal. However, how these factors interact to impact voice recognition, particularly in situations where multiple of these information sources show variability, remains unclear. How do changes in the languages spoken and emotional affect influence voice recognition? Are these changes additive or do they interact? And how does listeners' reliance on these cues vary depending on the language and emotion of the exposure language?

To address these questions, the current study used a voice recognition task to examine how changes in language and emotion between exposure and test affect voice recognition across familiar (English) and unfamiliar (Urdu) languages and across different emotions (angry vs. neutral). For each voice recognition trial, participants were exposed to a short passage spoken by one of four speakers. During the test phase, in a voice lineup, participants identified which of the four speakers they had just heard. Across trials, language and emotional affect either matched or mismatched between exposure and test, to create four trial types: no mismatch trials (language and emotional affect are the same in the exposure and test phases), emotion mismatch trials (the emotion changes from angry to neutral or vice versa), language mismatch trials (the language changes from Urdu to English or vice versa), or complete mismatch trials (both language and emotion change from exposure to test). Following previous work, we predicted that voice recognition would be hindered by changes in language (Winters et al., 2008) and emotion (Saslove & Yarmey, 1980; Xu & Armony, 2021) from exposure to test. We also anticipated possible directional effects, such that changes in emotion might be more detrimental in cases where participants cannot rely on linguistic cues (i.e., when the language is unfamiliar).

Methods

The sample size, exclusions and analyses for this study were pre-registered on AsPredicted: https://aspredicted.org/fx6t-mvq6.pdf

Participants

We recruited 156 monolingual English-speaking participants (age: $M = 20.14$; range = 18-40; self-reported gender identity: 121 female, 30 male, 5 prefer not to answer or prefer to describe) from the UCLA subject pool.

Seventy-three additional participants were tested but excluded from the final sample for the following reasons: 1) they did not meet the language inclusion criteria (e.g., did not learn English before the age of 6 in the USA or speak English at least 80% of the time; $n = 38$), 2) they indicated they had exposure to Hindi or Urdu ($n = 8$), 3) they reported a speech, language, vision, reading, or hearing issue that could interfere with their participation ($n = 4$), or 4) they indicated on the post-experiment survey that we should not use their data ($n = 23$).

Test trials in which response times were longer than 20s were removed from the analysis (we note that participants for whom more than 40% of trials were removed would have been excluded; however, no participants met this criterion).

Stimuli

The stimuli consisted of audio recordings of four English-Urdu bilingual speakers who learned both languages from birth and spoke them fluently without an identifiable "foreign" accent. Each speaker was recorded producing eight semantically neutral exposure passages (four per language; 32-36 syllables long), in both an angry and a neutral affect. In addition, each speaker produced 16 semantically neutral test sentences (eight per language), again in both affects. For each speaker, the clearest and most representative recordings were selected and edited, resulting in a final set of 16 exposure recordings and 32 test recordings per speaker. All

speech stimuli were RMS-equalized to control for differences in overall amplitude. For F0 and duration measurements by speaker, language and affect, see the online supplement: https://osf.io/ygpck/overview?view_only=b1c8f9e9f5a3455f8dc405aac824d8e5

Design

Each trial consisted of a brief exposure phase in which participants listened to one of four voices, followed by a 30s distractor task and then a test phase in which they selected the voice they had just heard from a voice lineup. Because only four voices were used in the experiment, the target voice for each trial was also presented in the lineup of the other trial(s), which may cause interference across trials. To reduce such interference, participants completed the experiments across four days: each day, they completed two blocks of four recognition trials, one of each type (i.e., no mismatch, language mismatch, emotion mismatch, complete mismatch). The exposure language (familiar vs. unfamiliar) varied across days (2 days in English, 2 days in Urdu). The exposure emotion varied within each day (2 neutral and 2 angry trials per block), such that each speaker was heard producing one passage with each emotion. Across days, participants completed a total of 32 trials (8 of each trial type). Each trial consisted of a distinct exposure passage (e.g., Speaker A's English, Angry, Passage #1) and a distinct set of test stimuli (e.g., Urdu, Angry, Sentence #2). Participants were randomly assigned to one of four lists (from a Latin Square design) that counterbalanced the pairings of speakers to exposure passages and trial types. The order of the trials within each block and the order of the four speakers in each voice lineup was computer randomized.

Procedure

Participants were given a link to the task on the online Platform Gorilla (Anwyl-Irvine et al., 2020; **Figure 1**). Each day began with a brief headphone screening task (Woods et al., 2017). Participants who passed the screening task were given onscreen instructions that described the learning and test phases of the experiment. The first day included a practice trial (that contained voices not used in the experiment). In each exposure phase, participants listened to one passage twice, followed by a 30s distractor task that asked them to count, and then report, the number of times a red dot appeared on the screen while instrumental music played in the background. Then, in the test phase, they were presented with a voice lineup: a given test sentence was produced sequentially by each of the four voices, and then participants were asked to select and drag the voice they heard in the exposure phase into a box in the center of the screen. On the last day, participants completed a short demographic and language questionnaire.

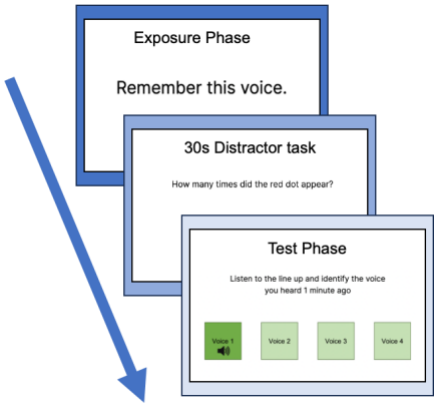

Figure 1. Diagram of the trial structure.

Analysis

Data were analyzed using a multilevel (i.e., mixed-effects) logistic regression with correctness as the dependent variable (0 incorrect; 1 correct), using the lme4 package (Bates et al., 2015) in R (R Core Team, 2025). We included fixed effects for *Exposure Language* (-0.5 Familiar, 0.5 Unfamiliar), *Exposure Emotion* (-0.5 Neutral, 0.5 Angry), *Language Change* (-0.5 No, 0.5 Yes), and *Emotion Change* (-0.5 No, 0.5 Yes),[1] as well as interactions between these variables. Because recognition could potentially differ across the four speakers[2], speaker was deviation-coded (sum coding) and included as a fixed effect to account for between-speaker variability (we had no a priori hypotheses about individual speakers). The complex random effect structure specified in our preregistration produced singular fits, even when only the slopes with the greatest variability were retained. Thus, we adopted a simplified model that included only random by-participant intercepts. Follow-up tests for three-way and two-way interactions were conducted using the emmeans package (Lenth, 2025). Differences between specific conditions were tested using holm-corrected comparisons.

Results

The results are presented in **Table 1** and **Figure 2**. Overall performance was above chance (25%) ($t_{(155)} = 10.10$, $p < .001$). We first discuss main effects, but those should be interpreted with caution given the interactions described later. Therefore, we accompany main effects with descriptions of marginal and simple effects as appropriate.

[1] Emotion and Language were coded in terms of Mismatches instead of Matches (i.e., we reversed the scoring from our AsPredicted), because it is more intuitive to have the "baseline category" be the matching condition.

[2] The speakers are a random sample from a larger population, which would typically render them a random effect. We treated them as a fixed effect because we only had four speakers, which was too few levels to reliably estimate the variance of a random effect.

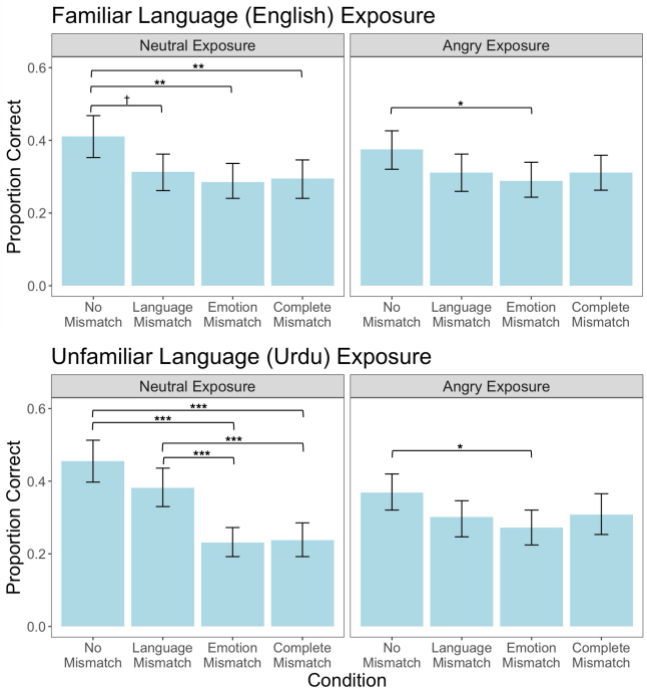

Figure 2. Voice recognition performance across the four conditions, by exposure language and emotion. For each pair of conditions, significance tests were Holm corrected across the four panels. † < .10, * p < .05, ** p < .01, ***p < .001

We found a robust main effect of emotion change: participants had greater difficulty recognizing voices when emotion changed from the familiarization phase to the test phase, regardless of whether language changed (one-tailed test, b = -0.39, p < .001). There was also an effect of language change (one-tailed test, b = -0.12, p = .028). This "language change" variable collapsed across trials with and without an accompanying emotion change; to characterize the effect of language change alone, we compared the marginal means in the No Mismatch condition to the Language Mismatch condition, and found that performance dropped when the language mismatched (b = 0.34, SE = 0.08, z = 3.97, p < .001). This effect was driven by the English-Neutral exposure (**Figure 2**, top left panel), which showed the greatest difference between these two conditions (b = 0.41, SE = 0.17, z = 2.44, p = .059); for the three other exposure types (the other panels) this difference was not significant (ps > .149).

Although language mismatch impaired performance, it was less disruptive to voice recognition than emotion mismatch. Collapsed across exposure language and emotion, the mean performance in the Emotion Mismatch condition was significantly lower than performance in the Language Mismatch condition (b = 0.28, SE = 0.09, z = 3.11, p = .006). As seen in **Figure 2**, this difference between the Emotion Mismatch and Language Mismatch conditions was most pronounced in the Urdu-Neutral exposure condition (bottom left panel). When exposed to a neutral passage in an unfamiliar language, a change to angry tone was much more disruptive than a change in language (b = 0.72, SE = 0.18, z = 4.02, p < .001). This difference between the Emotion Mismatch and Language Mismatch conditions after Urdu-Neutral exposure was significantly larger than after each of the other exposures (ps = .042); following these other exposures, emotion and language mismatches did not significantly differ (ps = 1).

These complex interactions between the exposure emotion, exposure language, and mismatch conditions are also captured in our multilevel model: here, effects of emotion change and language change varied as a function of exposure emotion and language.

Table 1. Fixed Effects for the Multilevel Logistic Regression Model Predicting Correctness

	b	*z*	*p*
Intercept	**-0.76**	**-22.82**	**<.001*****
ExpLang (*Familiar* vs. Unfamiliar)	-0.04	-0.61	.545
ExpEmo (*Neutral* vs. Angry)	-0.03	-0.45	.656
EmoChange (*No* vs. Yes)	**-0.39**	**-6.37**	**<.001*****
LangChange (*No* vs. Yes)	**-0.12**	**-1.92**	**.055**†
Speaker1	0.07	1.25	.212
Speaker2	-0.08	-1.55	.121
Speaker3	**0.14**	**2.68**	**.007****
ExpLang*ExpEmo	-0.02	-0.16	.876
ExpLang*EmoChange	**-0.28**	**-2.26**	**.024***
ExpEmo*EmoChange	**0.39**	**3.14**	**.002****
ExpLang*LangChange	0.04	0.29	.770
ExpEmo*LangChange	0.08	0.68	.496
EmoChange*LangChange	**0.43**	**3.51**	**<.001*****
ExpLang*ExpEmo*EmoChange	**0.49**	**1.99**	**.047***
ExpLang*ExpEmo*LangChange	-0.06	-0.23	.819
ExpLang*EmoChange*LangChange	0.00	0.02	.985
ExpEmo*EmoChange*LangChange	0.04	0.17	.869
ExpLang*ExpEmo*EmoChg*LangChg	0.11	0.22	.826

Note. Reference levels are in italics, all p-values are two-tailed. Final model: Correct ~ ExpLang * ExpEmo * EmoChange * LangChange + Speaker + (1| Participant)

First, the effect of Language Change depended on whether there was an Emotion Change (b = 0.43, p < .001). Follow-up tests for this two-way interaction revealed that, if the emotion matched between exposure and test, changing languages impaired recognition (Language Mismatch vs. No Mismatch: b = 0.34, SE = 0.08, z = 3.97, p < .001). In contrast, when emotion was mismatched, accuracy was low regardless of whether the language matched or mismatched (Emotion Mismatch vs. Complete Mismatch: b = -0.10, SE = 0.09, z = -1.09, p = .275). In other words, a language change did not further impair performance beyond the substantial disruption already introduced by the emotion change.

Second, the impact of Emotion Change varied depending on the Exposure Language (b = -0.28, p = .024). Even though switching emotion from exposure to test was associated with lower recognition accuracy for both exposure languages, follow-up tests indicated that this cost was larger when the

exposure language was unfamiliar (b = 0.53, SE = 0.09, z = 6.05, p < .001) compared to when it was familiar (b = 0.26, SE = 0.09, z = 2.94, p = .003). This pattern suggests that emotion switches are particularly disruptive in a foreign-language context: when exposed to a voice with an unfamiliar phonological system, encoding into memory may rely less heavily on information about the realization of phonemes or lexical items, and instead depend more on acoustic cues and/or prosody; when those aspects are changed during test due to a change in emotion, recognition is hampered.

Third, the effect of Emotion Change depended on the Exposure Emotion (b = 0.39, p = .002). Specifically, a change in emotion from neutral (exposure) to angry (test) appeared to be more detrimental to recognition (b = 0.59, SE = 0.09, z = 6.70, p <.001) than the opposite change, i.e., from angry to neutral (b = 0.20, SE = 0.09, z = 2.29, p = .022).

Importantly, the patterns reported above were qualified by a three-way interaction between Emotion Change, Exposure Emotion, and Exposure Language (b = 0.49, p = .047). Thus, we conducted follow-up tests to examine the interaction between Exposure Emotion and Emotion Change separately for familiar and unfamiliar language exposure. When exposure was in an unfamiliar language, there was a significant two-way interaction between Exposure Emotion and Emotion Change (b = 0.64, SE = 0.18, z = 3.60, p < .001). In these trials, the cost of switching emotion from exposure to test was greater following neutral exposure compared to angry exposure (**Figure 2**, compare the two bottom panels). In contrast, when the exposure was in a familiar language, the interaction between Exposure Emotion and Emotion Change was not significant (b = 0.14, SE = 0.17, z = 0.83, p = .410), indicating that the cost of switching emotions was less impacted by the exposure emotion (**Figure 2**, compare the two top panels).

Discussion

To recognize voices, listeners draw on multiple sources of information, including how speakers realize phonemes and words, as well as broader aspects of the speech signal such as the acoustics of the speaker's voice (e.g., pitch, voice quality). Previous research suggests that voice recognition suffers when the cues that listeners used to encode the speaker's voice are no longer available or have been modified at test (Read & Craik, 1995; Saslove & Yarmey, 1980; Winters et al., 2008; Xu & Armony, 2021). Consistent with those findings, we find that both language and emotion mismatches impair recognition; however, these sources of variation were not equally disruptive. In our study, changes in emotion (neutral↔angry) were more detrimental than changes in language (English↔Urdu), particularly when voices were learned in an unfamiliar language (Urdu), suggesting that listeners may rely heavily on acoustic information (e.g., pitch or voice quality) to differentiate speakers in circumstances where the speaker's specific realization of phonemes and words is less informative. Our results demonstrate that voice recognition depends on how listeners integrate linguistic and emotional cues, and that the relative importance of these cues shifts with exposure context.

Although changes in emotion from exposure to test were associated with larger and more consistent reductions in recognition accuracy than changes in language, this difference was driven primarily by a particular combination of exposure language and emotion: when exposed to an unfamiliar language in neutral tone, a change to angry tone (i.e., Emotion Mismatch) was more disruptive than a change in language (i.e., Language Mismatch). In addition, in most cases, changing both emotion and language was not more detrimental than changing either in isolation, which could suggest that the costs of these two changes are not additive (or supra-additive). However, this could also reflect a floor effect in a highly challenging task.

We propose the following, tentative (and post-hoc) account for the relationship between exposure language and emotion and the effects of language vs. emotion change (**Figure 2**). This account makes several assumptions about how listeners encode, and later compare, speaker information across exposure and test. First, when representing a speaker's voice, listeners weigh more heavily sources of information that allow them to better differentiate speakers. When one set of cues is less salient or is uninformative, listeners rely more heavily on the remaining cues, leading to a representation that may be dominated by a narrower subset of features. In contrast, when multiple sources of information are salient and/or informative, the representation is distributed across cue types, allowing for greater flexibility in response to changing conditions. Second, the precision of this representation depends, in part, on the stability of those cues during the exposure phase. Stable cues (e.g., stable pitch or speech rate) support more precise encoding, whereas variability in those cues (e.g., more variable pitch or speech rate) leads to coarser representations, which—although they are less precise—may be more tolerant to change. Finally, we propose that the cost of any change between exposure and test is some joint function of (a) the weight assigned to a given cue during encoding, (b) the precision of that cue's representation in memory, and (c) the magnitude of change in that cue. Changes are therefore most disruptive when they affect features that were both heavily weighed and precisely encoded (less tolerant to transformations).

We further assume that angry voices differ from neutral voices in specific ways that relate to the above assumptions. Angry speech may increase the salience of a small set of dimensions (e.g., pitch, loudness). This small set of dimensions, however, can vary substantially within the exposure phase (i.e., low stability), resulting in a coarser or less precise representation of the speaker's voice. On the one hand, this variability is shared across speakers: anger has systematic effects on voice (Paulmann et al., 2008), such that these dimensions behave in ways that are more diagnostic of anger itself than they are diagnostic of the speaker's identity. On the other hand, because the representation is coarse, it may be more tolerant to variations at test, particularly when

the test stimuli contain some overlap with the encoded features. For instance, if exposed to angry and tested on neutral, a speaker's acoustic characteristics (e.g., their pitch or voice quality) at test likely fall within the range experienced during exposure. Neutral emotion, in contrast, allows a wider range of dimensions to be weighed more heavily; and their encoding is more precise because neutral speech expresses acoustic properties more consistently (higher stability). However, the precision of the representation means that it is relatively intolerant to variation at test, particularly when the test stimuli contain features that do not fit into the narrowly exposed range.

As mentioned above, the type of exposure that resulted in the largest difference between the language mismatch and the emotion mismatch conditions was Urdu-neutral; in all other exposure types the effect of language and emotion mismatch was comparable. The assumptions above may begin to explain why emotion mismatches are particularly detrimental in this condition. When exposed to an unfamiliar language, the phonemic categories are unfamiliar and thus the listener cannot rely heavily on representations of the speaker's linguistic characteristics (e.g., their phonemic realizations) to differentiate them. Instead, they are forced to weigh acoustic cues heavily. During neutral speech, these acoustic cues are relatively stable (low variability), resulting in a precise representation of speaker identity that is based on a limited set of cues (due to language unfamiliarity). At test, an emotion change from neutral to angry produces a large change in acoustics that the precise memory representation is relatively intolerant to, resulting in a high cost. Other types of transformations, such as a language change from Urdu-neutral to English-neutral, may be less detrimental because, although they introduce new (and salient) phonological and lexical information at test, they preserve most of the encoded acoustic information. Therefore, recognition is disrupted to a lesser extent.

This account may also explain the three-way interaction we observed between Exposure Language, Exposure Emotion, and Emotion Change. Specifically, we found that the effect of emotion change depended on the exposure emotion for Urdu, but there was no such asymmetry in English. As described above, in an unfamiliar language, neutral exposure yields precise encoding, making a subsequent change to angry speech especially costly (due to specific but intolerant representations); however, angry exposure yields coarser representations that are more tolerant, such that a change to neutral speech is less disruptive in comparison. In a familiar language, by contrast, information about the speaker's linguistic characteristics (e.g., phonetic encoding) is available, allowing for a representation that is distributed across a wider variety of cues. This representation is more robust to transformations, which may make the cost of switching from neutral to angry closer to the cost of switching from angry to neutral. Although the weight of acoustic/prosodic emotional cues for English angry speech may still be somewhat higher than for English neutral speech (due to anger's higher salience), which would make an emotion change more disruptive, the relatively higher tolerance of the angry representation may counteract this difference in cost. (In an unfamiliar language, acoustic information is highly weighted even for neutral speech, because phonetic encoding is very limited, so there is little room to further increase the weight of these cues in angry speech).

We note that this account was developed post-hoc to explain the observed findings and should therefore be considered provisional. In addition, floor effects in this challenging task, and potentially underpowered analyses, may have obscured differences between conditions that are not fully consistent with the proposal. Therefore, future work should identify and test new predictions generated by this account with different language and emotion stimuli. The impact of switching emotion likely varies depending on the specific emotions selected. Moreover, it would be important to test how well these results generalize to different populations, such as young children, who may be less reliant on phonological cues to identify voices.

In sum, the present findings demonstrate that the impact of cue mismatch on voice recognition depends on both the linguistic and emotional context of exposure. By directly contrasting emotion and language mismatches within a single experimental framework, we provide evidence that emotion changes—at least between neutral and angry voices—strongly impact recognition, particularly when listeners are processing unfamiliar languages. These results underscore the importance of considering multiple sources of variability in voice recognition and suggest that models of talker memory must account for dynamic interactions between linguistic content, emotional expression, and listener experience. Understanding how these factors jointly shape recognition has implications for applied domains, including forensic voice identification and speech technologies, where listeners or systems must generalize voices across variable speaking conditions.

Acknowledgments

We would like to thank undergraduate research assistants in the UCLA Teaching Learning and Communication lab for assistance with setting up the study and collecting data. Artificial intelligence–based tools were used to assist with refining statistical analysis code and improving the clarity and organization of the manuscript text. The authors retain full responsibility for the study design, analyses, interpretations, and conclusions.

References

Abu, S., Adas, E., & Levi, S. V. (2022). Phonotactic and lexical factors in talker discrimination and identification. *Attention, Perception, & Psychophysics*, *84*, 1788–1804. https://doi.org/10.3758/s13414-022-02485-4

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research*

Methods, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Fecher, N., & Johnson, E. K. (2018). Effects of language experience and task demands on talker recognition by children and adults. *The Journal of the Acoustical Society of America*, *143*(4), 2409–2418. https://doi.org/10.1121/1.5032199

Goggin, J. P., Thompson, C., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition, 19*(5), 448–458. https://doi.org/10.3758/bf03199567

Kim, Y., Sidtis, J. J., & Van Lancker Sidtis, D. (2019). Emotionally expressed voices are retained in memory following a single exposure. *PLoS ONE*, *14*(10), e0223948. https://doi.org/10.1371/journal.pone.0223948

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review, 26*(1), 90–102. https://doi.org/10.3758/s13423-018-1497-7

Lenth, R. V. (2025). emmeans: Estimated marginal means, aka least-squares means. In *CRAN: Contributed Packages*. https://doi.org/10.32614/CRAN.package.emmeans

Orchard, T. L., & Yarmey, A. D. (1995). The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology*, (9), 249–260. https://doi.org/10.1002/acp.2350090306

Paulmann, S., Pell, M. D., & Kotz, Sonja. (2008). How aging affects the recognition of emotional speech. *Brain and Language, 104*(3), 262–269. https://doi.org/10.1016/j.bandl.2007.03.002

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*(1), 42–55. https://doi.org/10.1016/j.cognition.2009.08.012

Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. E. (2011). Human voice recognition depends on language ability. *Science, 333*(6042), 595. https://doi.org/10.1126/science.1207327

R Core Team. (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Read, D., & Craik, F. I. M. (1995). Earwitness identification: some influences on voice recognition. *Journal of Experimental Psychology: Applied*, *1*(1), 6–18. https://doi.org/10.1037/1076-898X.1.1.6

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: speaker identification. *Journal of Applied Psychology, 65*(1), 111–116. https://doi.org/10.1037/0021-9010.65.1.111

Stevenage, S. V., & Neil, G. J. (2014). Faces and seeing voices: The integration and interaction of face and voice processing. *Psychologica Belgica*, *54*(3), 266–281. https://doi.org/10.5334/pb.ar

Thompson, C. P. (1987). A language effect in voice identification. *Applied Cognitive Psychology*, *1*(2), 121–131. https://doi.org/10.1002/acp.2350010205

van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: patterns and parameters Part 1: Recognition of backward voices. *Journal of Phonetics, 13*(1), 19–38. https://doi.org/10.1016/S0095-4470(19)30723-5

Wester, M. (2012). Talker discrimination across languages. *Speech Communication*, *54*(6), 781–790. https://doi.org/10.1016/j.specom.2012.01.006

Winters, S. J., Levi, S. V., & Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages. *The Journal of the Acoustical Society of America*, *123*(6), 4524–4538. https://doi.org/10.1121/1.2913046

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

Xu, H., & Armony, J. L. (2021). Influence of emotional prosody, content, and repetition on memory recognition of speaker identity. *Quarterly Journal of Experimental Psychology*, *74*(7), 1185–1201. https://doi.org/10.1177/1747021821998557

Yu, M. E., Schertz, J., & Johnson, E. K. (2021). The other accent effect in talker recognition: Now you see it, now you don't. *Cognitive Science*, *45*(6), e12986. https://doi.org/10.1111/cogs.12986