GLOSSA psycholinguistics

Zellou, G., Rácz, P., & Barreda, S. (2026). Social associations between voices and semantics affect word learning in voice-AI contexts. *Glossa Psycholinguistics*, 5(1): X, pp. 1–29. DOI: https://doi.org/10.5070/G6011.47021

eScholarship University of California

Social associations between voices and semantics affect word learning in voice-AI contexts

Georgia Zellou, Linguistics Department, UC Davis, gzellou@ucdavis.edu

Péter Rácz, Department of Cognitive Science, Faculty of Natural Sciences, Budapest University of Technology and Economics, racz.peter.marton@ttk.bme.hu

Santiago Barreda, Linguistics Department, UC Davis, sbarreda@ucdavis.edu

The present study investigates whether the relationship between voice gender (female and male voices) and gender-associated semantics (e.g., *flower* [female] vs. *football* [male]) affects the learning of an artificial lexicon after brief auditory exposure. Participants were trained on 16 novel words in either a gender-aligning condition (i.e., female voices produced novel words with female-associated meanings and male voices produced novel words with male-associated meanings) or a gender-mismatching condition (e.g., female voices produced words with male-associated meanings, etc.). Listeners either heard voices that contain stereotypical or non-stereotypical gross acoustic patterns. Overall, listeners showed worse word learning in gender-mismatching conditions. Voice stereotypicality modulated these effects: flipping gross gender cues reduced learning differences across conditions. These results demonstrate how gender associations modulate social attention when learning a new language. The findings also have implications for gender biases in human-computer interaction and speech technology, since the experiment was designed under the guise of a language learning application.

1. Introduction

Second language acquisition is difficult for adult learners. While much work has focused on explanations rooted in cognitive plasticity and age-related constraints, another potential contributing factor is the influence of statistical biases from one's first language and personal cultural experiences. This is an area that has received comparatively less attention and warrants further investigation. Statistical biases can arise from socio-indexical information. Linguistic patterns are socially structured, and this structure can shape how learners map sensory input onto linguistic forms. Prior work demonstrates how social associations affect first-language spoken word comprehension (e.g., Kim, 2016; McGowan, 2015) and can have societal consequences (e.g., accent discrimination, Baugh, 2016). The role of socio-indexical information in shaping adult language learning is also a growing area of research (cf. Feher et al., 2014; Needle & Pierrehumbert, 2018; Rácz et al., 2017; Roberts, 2017). Adults have expectations regarding what, for example, men and women of the same age sound like (e.g., Strand, 1999) and talk about (Eckert, 1989), based on statistical patterns in their experiences, and these expectations might influence how novel word forms are learned.

Among these socio-indexical cues, gendered expectations – rooted in socially constructed stereotypes – may play a particularly influential role in shaping how learners process and acquire new words. Our goal is to examine how these expectations function, not to endorse them. This is because gender is a highly salient social category that listeners consistently classify from voice and use to guide attention and interpretation during language processing. In the current study, we ask if social associations between semantic meaning and a speaker's voice acoustics affect the learning of novel words. Specifically, we investigate whether the relationship between perceived voice gender and stereotypically gender-associated meanings (e.g., *flower* = strongly female-associated vs. *football* = strongly male-associated) influences the learning of an artificial lexicon after brief auditory exposure. We also manipulate the voices producing the words such that they contain either stereotypical or non-stereotypical gross acoustic patterns across different versions of the experiment, thereby additionally exploring the effect of voice gender typicality on word learning. Furthermore, since device-based language learning applications are widely used (e.g., Duolingo), the experiment was designed to resemble a language learning app. Since learning is a general cognitive process, the current study can address longstanding issues at the intersection of language, society, and learning.

1.1 Socio-indexical information and language processing

Language contains socially structured variation: linguistic variants are not evenly distributed across speaker groups (e.g., Labov, 2006). Gender is a key variable in linguistic variation. Individuals across gender identities tend to produce systematically distinct acoustic, phonological, lexical, and discourse patterns within a given speech community (for reviews, see Cheshire, 2004; Coates, 2015; Eckert & McConnell-Ginet, 2013). Voices provide a listener with multiple layers of

indexical meaning about talkers and are used by listeners to perform social categorizations. Based on voice alone, listeners show consistent classification of a speaker's gender (e.g., Poon & Ng, 2015), among other attributes such as age (e.g., Barreda & Assmann, 2021) and nativeness (e.g., Girard et al., 2008). Such social categorizations subsequently affect how language and speech are processed by the listener (e.g., Drager & Kirtley, 2016; Hay et al., 2006; Niedzielski, 1999; Strand, 1999). For instance, listeners' belief that a gender-ambiguous voice is either a woman or a man shifts their categorizations of the fricatives /s/ and /ʃ/ (Strand, 1999). The proposed explanation for this is that listeners make a gender classification of the speaker, and then identify the speech sound based on gender-dependent acoustic priors for linguistic categories.

Socially-based expectations about how language variation patterns across different types of speakers impacts language comprehension: McGowan (2015) found that comprehension of foreign-accented speech was enhanced when viewing an Asian vs. a white face (an extension of earlier work by Rubin (1992) that showed lower comprehension of non-accented speech when viewing an image of an Asian face, compared to an image of a white face). For speaker age, Kim (2016) found that listeners respond faster and more accurately to Korean words rated as being stereotypically "older" and "younger" when produced by an adult voice matching in the age category; meanwhile, processing was slower and less accurate when word and age category mismatched. Most, Sorber, and Cunningham (2007) had participants perform an "auditory Stroop task" where they categorized the gender of voices producing names and words stereotypically associated with female and male gender roles (e.g., *Rachel*, *football*) and found slower reaction times in trials where voice and word gender-associations were incongruent. Such effects can be explained using existing frameworks of language processing which generally assume that contextual cues are stored alongside linguistic representations, which is why context (like semantic content or social attributes, like speaker gender) can cue representations (such as how acoustic patterns map onto phonetic categories based on experienced priors) and vice versa (e.g., Johnson & Sjerps, 2021, discussed further below). There is evidence that these statistical patterns arise very early during language learning: For instance, caretakers produce higher rates of standard phonological forms when talking to girls than boys (Foulkes et al., 2005; Smith et al., 2009). Consequently, even very young children display gendered use of linguistic variants (Munson et al., 2022; Prystawski et al., 2020; Robertson & Murachver, 2003; Stennes et al., 2005).

There is some prior work looking at how socio-indexical information affects novel word learning. Rácz, Hay, and Pierrehumbert (2017) found that participants can learn an artificial language containing speaker-gender conditioned morpho-phonological variation; thus, listeners are sensitive to socially structured variation in the input during language learning. They also compared learning in the gender-correlated variation condition to a condition where variant use correlated with different spatial orientations of the talker (e.g., a forward-facing speaker used one alternant and a side-facing speaker used another). Participants learned gender-correlated variation (a socially-structured way that language variation occurs in the real world) better

than spatially-oriented variation (not a socially-relevant cue in the real world), suggesting that perceivers are attending more strongly to socially interpretable cues when learning language. More evidence comes from Needle and Pierrehumbert (2018) who found that participants' gender identifications of real words matched those words' statistical distributions across a corpus of women's and men's writing (e.g., words more likely to appear in women's writing were more likely identified as produced by a female author). Moreover, Needle and Pierrehumbert also found an effect for pseudowords that contained English-like derivational affixes that had gender-biased statistical distributions: for instance, the *-case* suffix (e.g., *pillowcase, suitcase*) is more likely to occur in women's writing, and participants were more likely to identify the author as female when evaluating pseudowords like *clumcase*.

Theoretical models of processing which posit that attention guides the encoding and storage of information can account for, and make predictions about, how social and linguistic details in the input might influence the strength of learning. For example, predictive-processing theories propose that comprehenders continuously generate expectations about upcoming information and allocate greater attention to unexpected inputs; the resulting attention boost when stimuli contain expectations violations are proposed to facilitate encoding and generalization. This has been observed in artificial language studies, where items containing expectation violations are demonstrated to strengthen abstraction and generalization of newly learned structures by learners (e.g., Bovolenta & Marsden, 2021). One might predict, then, that learners exposed to items containing social and semantic cues that violate expectations will display stronger learning than those exposed to matching cues, since they attract greater attentional engagement and, thus, lead to stronger encoding and generalization.

However, not all violations are beneficial to learning. Experiments that have manipulated social cues in discourse observe that incongruent turn-taking or prosody hinder the success of word learning (Lee & Lew-Williams, 2022). More support comes from a series of large-scale experiments on cross-situational word learning which found that adults relied heavily on social cues (e.g., gaze) to reduce referential uncertainty (MacDonald et al., 2017). When these cues were less reliable, learners reverted to tracking multiple hypotheses instead of committing to a single mapping, which slowed learning. This suggests that violations of cue reliability expectations hinder efficient encoding. Thus, finding that mismatched social and semantic information leads to lower word learning in the present study would be consistent with models of processing where learners are assessing the reliability of social cues and use that to guide attention during learning.

1.2 Social factors in human-computer interaction

Humans are now in a new digital era where many people of all ages regularly talk to voice-activated artificially intelligent (AI) personal assistants, such as Siri, Alexa, and Google Assistant, that spontaneously produce interactive speech (Ammari et al., 2019). While language-based

interaction with computers has long existed – primarily through typing – the advent of spoken language technology now enables users to operate devices using speech alone. The ease and efficiency in using speech to interact with machines to complete everyday tasks explains why people find speech-enabled functions natural and effortless, which means that voice-AI is being integrated into a larger proportion of our spoken interactions (De Renesse, 2017). It also suggests that many more humans will be spending a lot of time talking to machines in the future.

Social expectations are so robust that they even apply in contexts where listeners know they are not interacting with a real human. Indeed, there is much research showing that people anthropomorphize machines by attributing human-like traits and characteristics to them (Waytz et al., 2010). Moreover, prior research shows that people apply human-based behavioral norms to interactions with machines and computers, including norms shaped by gender stereotypes. For instance, Kuchenbrandt and colleagues (2014) tested whether the gender-typicality of a task with a robot would affect users' performance. They had participants complete either a stereotypically-female task (sorting a sewing kit) or a stereotypically-male task (sorting a toolbox) while being instructed by a robot with either a female or a male guise. They found participants performed worse with the robot in the context of a stereotypically-female task, for both the female and male guises. Kuchenbrandt and colleagues remarked that many of the activities we perform with robots are inherently related to societal gender roles, and this will influence how users interact with robots.

For voice-enabled devices in particular, just by the nature of their use of speech, it has been shown that people apply voice bias to machines. For instance, Nass, Moon, and Green (1997) had participants complete a tutoring session with a computer that produced either a female voice or a male voice and were presented with either stereotypically masculine topics (i.e., "technology") or stereotypically feminine topics (i.e., "relationships and love"). They found that congruence between the apparent gender of the voice and the stereotypicality of the topic led to higher ratings of the computer's competence than if there was a voice-topic mismatch. Ernst and Herm-Stapelberg (2020) investigated whether gender stereotyping influences the perceived likability of virtual voice assistants. They had participants interact with a virtual assistant, assigned either to a female-voiced condition or a male-voiced condition. Post-interaction surveys revealed that participants who interacted with the female-voiced assistant rated the system as more likable than the male-voiced assistant. These studies show that people are particularly prone to applying human-based gender expectations to machines when they express apparent gender through their voice.

This shift toward voice-based interaction raises important questions about how social expectations, particularly gender stereotypes, influence user experience and learning outcomes. Language learning apps, such as Duolingo, and computer programs, such as Rosetta Stone, are digital systems that users can use as a didactic tool for vocabulary building in a second

language. Language learning apps are most commonly used for home learning (Godwin-Jones, 2011), but they are also being explored for use in the classroom (Guaqueta & Castro-Garces, 2018). And there are projections that device-based language learning applications will increase in future years (Godwin-Jones, 2017). There is a growing body of work exploring whether people exhibit social behavior towards machines in the same ways that they do with other humans (e.g., Gambino & Liu, 2022; Nass et al., 1997). For instance, in extending the classic Asch study of social conformity to peers, Brandstetter and colleagues (2014) found that while participants show conformity to inaccurate verbal judgements made by human peers, they do not show similar influence from robot "peers". With respect to speech and language patterns, the literature is mixed. In some prior work, participants show perceptual adaptation of a vowel shift produced by an apparent voice-AI device to the same degree as they do for an apparent human speaker (Zellou et al., 2023). Yet other work has shown that college-age participants show little to no phonetic imitation of device voices (Zellou, Cohn, & Ferenc Segedin, 2021; Zellou, Cohn, & Kline, 2021). Thus, it remains an open question as to whether people learn, and how human-based social patterns will affect learning, when interacting with a voice-AI device. What role does voice gender play in learning via computer-based applications? The Nass et al. (1997) study found that incongruence between the gender of the voice and the semantic content of a computer-based tutoring session affects listeners' learning experience. Thus, for the present study, we predict that, even under the guise of a word learning computer app, gender biases in word learning will be found.

There are important societal implications of the growing ubiquity of voice-AI and an observation of gender associations impacting language learning when interacting with machines. For one, female-voiced digital assistants as the default option can lead to reinforcing and exacerbating gender asymmetries that exist in society. Digital devices are often used to assist, are subservient toward, and rarely disagree with or are unpleasant to, the user. Female-voiced AI systems send the message to users that women and girls should express themselves similarly (West et al., 2019, UNESCO report). Moreover, as mentioned earlier, the current state of speech technology is that most commercially available systems in use provide text-to-speech voices that reflect a gender binary. These choices can impact behavior during interactions with speech and language technology. More generally, such biases could reinforce and amplify the notion that gender is a binary, as well as societal biases around gender-nonconformity.

Examining the role of social factors on lexical acquisition in the context of human-computer interaction can also speak to theoretical accounts of attention and learning. Apparent gender in synthetic voices might activate gendered expectations about content and competence, which, in turn can shape attention during learning tasks. Device voices can act as sources of prior expectations, which can lead to differential allocations of attention. Within a predictive-processing account, incongruence of voice and semantic cues could trigger helpful surprise,

boosting learning. In contrast, cue-reliability accounts of learning predict that cue mismatches impose processing costs that hinder learning; when voice-AI-based social cues are congruent with content, perceptual fluency is higher and learning is facilitated.

1.3 Current study

The current study investigates whether statistical associations between social properties of voices and semantic meaning together influence word learning. We tested this using a mock word learning app called WordBot. In general use, text-to-speech voices typically align stereotypical gender cues (f0, vocal tract length, prosody, source characteristics), which means that non-stereotypical combinations remain understudied in human-computer interaction.

In the training phase, listeners hear sixteen novel words in an invented language produced by two talkers. We manipulate the relationship between voice and semantic content across two conditions: gender-aligning (female voice produces female-associated words like *lipstick*, *flower*, male voice produces male-associated words like *hammer*, *football*) and gender-mismatching (voice-meaning pairings are reversed). In the test phase, listeners hear the same voices as in the training phase, producing the same words. Listeners need to identify associated objects.

Following the test phase, listeners enter a second test phase in which they hear the same words in two additional voices. The relationship between voice and semantic content is the same as with the training voices. This phase examines whether generalization differs across alignment conditions. Prior work on perceptual learning shows that generalization to new talkers is a hallmark of robust learning, yet it is often weaker than same-talker performance, because cross-talker variability introduces processing costs (Eisner & McQueen, 2005; Xie & Myers, 2017). In the present study, we predict that participants will generalize their learning to novel voices, but with reduced accuracy compared to the same-voice test block. This expectation is based on theories of attention and cue reliability: when social cues (voice–semantic associations) align with prior expectations, perceptual fluency should facilitate generalization; when mismatched, reduced cue reliability may hinder it. Thus, we anticipate successful generalization overall, but also expect that alignment versus mismatch will modulate the strength of generalization effects.

We test voice stereotypicality across two conditions, original text-to-speech voices (stereotypical associations between gendered speech patterns and gross acoustic cues) or flipped voices (gross acoustic patterns and fine-grained phonetic patterns incongruent, creating non-stereotypical male and female voices). For example, in the flipped condition, male voices would receive the average f0 and spectral cues typical of female voices, while maintaining the more subtle gendered cues associated with the male voices (e.g., intonation, phonological variation).

Following the word learning study, participants provide social evaluations of the voices: communicative competence ratings and assessment of whether they would use a language learning app with each voice.

We had the following hypotheses:

- H1. Participants will show better word learning performance in the gender association-aligning conditions, compared to gender-mismatching conditions.
- H2. Participants will successfully generalize learning to novel voices, but accuracy will be lower than in the same-voice test block, and this reduction may interact with alignment conditions (gender-aligning vs. gender-mismatching).
- H3. The mismatch effect will be weaker with the flipped AI voices than with the original voices.
- H4. If explicit judgments of voices affect word learning, we predict that if participants rate the voices as less communicatively competent and less preferred, the participants will have lower learning outcomes.

2. Methods

2.1 Stimuli

In the current study, all participants learned 16 novel word-image pairings. The novel words were taken from a previous artificial language study (Kaushanskaya & Marian, 2009). We took 16 of the "phonologically familiar" items, designed to be similar in structure to many English real words. They were constructed from 8 English phonemes: four vowels (/ɑ/, /ɛ/, /i/, and /u/) and four consonants (/f/, /n/, /t/, and /g/).

In order to identify what the meaning associations would be, we used a publicly available corpus of psycholinguistic ratings of over 5,500 English words (Scott et al., 2019). In that study, they had participants rate words on a number of scales, including gender. Participants were told "A word's gender is how strongly its meaning is associated with male or female behavior. A word can be considered MASCULINE if it is linked to male behavior. Alternatively, a word can be considered FEMININE if it is linked to female behavior. Please indicate the gender associated with each word on a scale of VERY FEMININE (1) to VERY MASCULINE (7), with the midpoint being neuter (neither feminine nor masculine)."

From the Scott et al. (2019) ratings data, we selected only meanings that were highly imageable (rated 5.4 or higher on a 7 point scale of imageability). Then, we selected the top eight most female-associated words (gender-associated rating mean of 1.5 for the eight words) and the top eight most male-associated words (gender-associated rating mean of 6.06 for the eight words). Words were replaced with the next appropriate word if they referred to drugs/addictive substances (*cigar*), weapons (*grenade, gun*), a physical person (*lady*), or a body part (*vagina*, *penis*). The lexical items and meanings (with imageability and gender ratings) used in this study are provided in the Appendix.

2.2 Voices

Each of the 16 words was generated in 8 text-to-speech US-English voices (4 female: Danielle, Kendra, Kimberly, Ruth; 4 male: Gregory, Matthew, Joey, Stephen) using AWS (Amazon Web Services) Polly in neural text-to-speech. The items were generated using SSML tags with the IPA form of the word (e.g., <speak> <phoneme alphabet="ipa" ph="fɑnɛt"> </phoneme> </speak>). Words were downloaded from the AWS console and amplitude normalized to 60 dB.

All stimuli were resynthesized using the 'Change gender' function in Praat. This function manipulates the sampling frequency to achieve a linear rescaling of the sound spectral envelope, thereby mimicking differences in speaker vocal-tract length (Barreda, 2020). Differences in f0 are implemented using overlap-add synthesis. In order to maintain uniformity in their processing, all stimuli were resynthesized using this function in two conditions: unmodified and flipped. For the unmodified condition, spectral envelopes were not manipulated, but male voices were modified to have a median f0 of 110 Hz, and female voices were modified to have a median f0 of 220 Hz. This allowed each speaker to maintain their unique pitch contour, but placed speakers of each gender in the same general range. In the flipped condition, male voices were given a median f0 of 220 Hz, and the spectral envelope was scaled up by a factor of 1.2; female voices were given a median f0 of 110 Hz, and the spectral envelope was scaled down by a factor of 1/1.2 (0.83). The result of these manipulations is a potential mismatch between gross gender cues (overall f0 and spectral cues to vocal tract length) and fine-grained gender cues (everything else: prosody, coarticulation, etc.) in the flipped condition. This mismatch will reduce the gender stereotypicality of flipped voices, and may affect the voices' perceived masculinity/femininity, even for voices all classified as female or male.

2.3 Participants

231 native English speakers took part in the study (self-reported: 174 female, 1 trans woman, 5 non-binary/gender non-conforming, 51 male; mean age = 19.7 years old). They completed the experiment online via a Qualtrics survey. Participants were recruited from the UC Davis subjects' pool and given partial course credit for their participation. This study was approved by the UC Davis Institutional Review Board and all participants completed informed consent. Participants were instructed to complete the experiment in a quiet room without distractions or noise, to silence their phones, and to wear headphones. None of the listeners reported having a hearing or language impairment.

2.4 Procedure

The study began with a pre-test of participant audio: participants heard one sentence presented auditorily ("She asked about the host") and were asked to identify the sentence from three multiple choice options, each containing a phonologically close target word (*host*, *toast*, *coast*). All participants passed this audio check.

Participants then completed the experiment, which consisted of four phases: training phase, talker-specific testing block, generalization testing block, then, finally, a voice ratings block. **Figure 1** provides a flowchart demonstrating the blocks and their ordering as completed by participants.

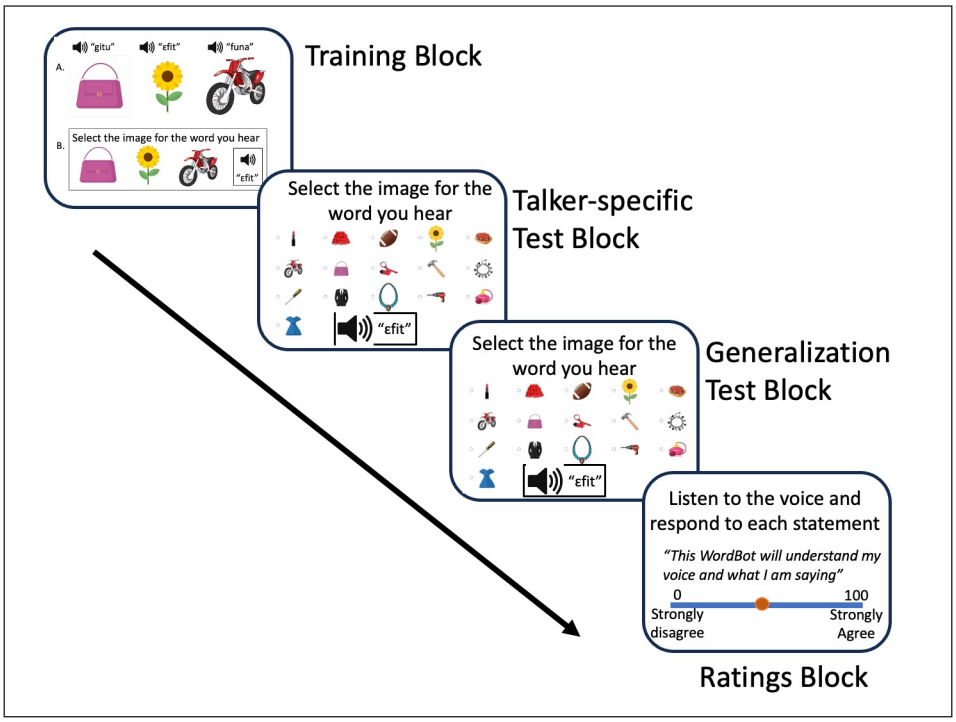

Figure 1: Flowchart of the experimental blocks.

The training phase consisted of a paradigm adapted from Wong and Perrachione (2007). In this phase, participants were trained to identify word meanings as depicted by drawings of concrete nouns. Each participant was trained on a vocabulary of 16 items. Every participant was trained on the same 16 item-meaning pairings (see the Supplementary Materials), with 8 of the items referring to a "female-associated" object and 8 of the items referring to a "male-associated" object.

The format of the training session is illustrated in **Figure 2**. To facilitate learning, participants were trained on a group of 3 words at a time (i.e., a trial group of 3 items). Each of the 16 items was presented a total of 3 times, each item randomly assigned to one of 16 trial groups. The trial groups varied in having gender-mixed words and gender-consistent words (the example in **Figure 1** happens to show a gender-mixed trial group). In a learning session, participants heard a word production and saw an image which they were told corresponded to the meaning of the word (**Figure 2.A**). After three trials, participants were then given a mini-test on the

three words they just learned. One of the three words was played and participants selected the correct corresponding image from among three images (**Figure 2.B**). After making a selection, feedback was provided to facilitate recognition of the items and to correct if a mistake was made. Participants heard 48 items (16 words * 3 repetitions across different trials groups).

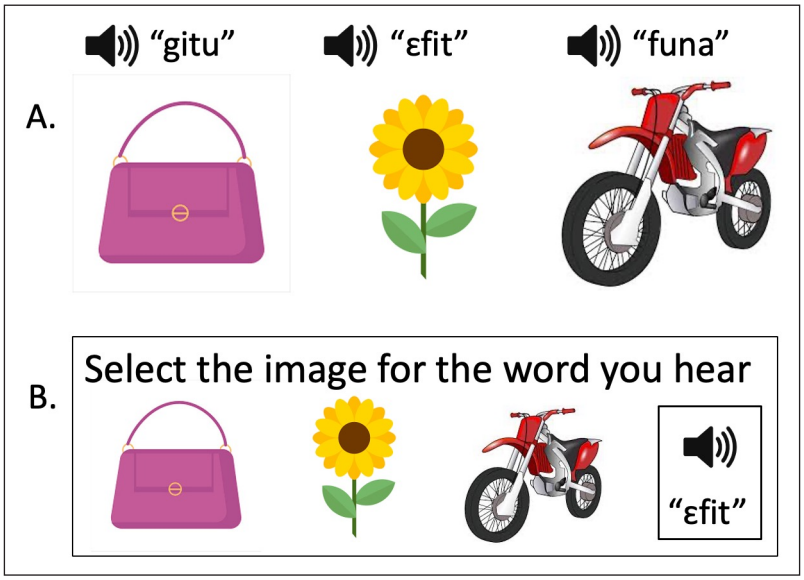

Figure 2: Example of a training session trial group. Panel A illustrates three separate trials from the same group, shown together for compactness; in the actual experiment, each word–image pair was presented one at a time. After hearing all three words with their corresponding images, participants were quizzed on those items (Panel B). During training, feedback was provided after each quiz response. Across the training phase, participants learned 16 word–image associations.

In the talker-specific testing block, which directly followed the training phase, participants were presented with each of the 16 images (see **Figure 1**). On each trial, they heard one of the 16 learned words presented in the voice that they had heard for that item in the training phase. They were asked to select the correct image. Words were presented randomly, twice each. No feedback was provided. Participants completed 32 total testing trials (16 words * 2 repetitions).

Directly following the talker-specific (Same Talker) testing block, participants completed the generalization (Novel Talker) testing block. The trial design was identical to the previous block, except participants now heard two novel voices produce the items. The gender-meaning pairing of the novel voices was the same as that for exposure for each participant. For instance, if they had heard the female voice produce female-associated words in training and testing block 1, in the novel-talker testing block, they heard female-associated words in a novel female voice. Words were presented randomly, twice each. No feedback was provided. Participants completed 32 total testing trials (16 words * 2 repetitions).

Participants were randomly assigned to one of the four experimental conditions: 57 were assigned to gender-aligning conditions with original voices; 57 were assigned to gender-mismatching conditions with original voices; 57 were assigned to gender-aligning conditions with flipped voices; 60 were assigned to gender-mismatching conditions with flipped voices.

Finally, participants completed several ratings for each of the four voices they had heard in the ratings block. First, we asked participants to provide a gender categorization of each of the voices ("What do you think the gender of this voice is?"; binary choice: Female or Male). Next, we asked them to rate the voices on gender stereotypicality (two questions: "How feminine/masculine does this voice sound?", Likert scale (1 = Totally feminine, 5 = Totally masculine); "How typical does this voice sound for its gender?", Likert scale (1 = Very unusual, 5 = Very typical). Then, we asked them to provide ratings of two statements to assess the voice's communicative competence. One statement assessed comprehension competence: "This WordBot will understand my voice and what I am saying". One statement assessed production competence: "This WordBot speaks clearly so that I understand everything it says". Next, we asked them to respond to two statements to assess the voice's naturalness, which prior work has shown to vary in perception of robots that have congruent vs. conflicting cues (Moore, 2012): Naturalness: "This WordBot sounds naturalistic/human-like" and Creepiness: "This WordBot sounds creepy". Finally, we asked them to provide an overall rating of the voice: "Overall, I would use this WordBot if it were available". Responses were provided on a Likert scale (1 = Strongly disagree, 5 = Strongly agree).

2.5 Data analysis

We collected 14,784 responses to items in the test and generalization phases of the task from 231 participants. We also collected a set of ratings from each participant on each of the voices that participant encountered.

We used R, ggplot, and sjplot for data analysis and visualization (Lüdecke & Lüdecke, 2015; R Core Team 2024; Wickham, 2016). We fit linear models using brms and rstan (Bürkner, 2017; Stan Development Team, 2024). The models were fit using 5000 iterations on 4 chains and weakly informative priors. We used leave-one-out cross-validation to select the best model. Data analysis was pre-registered (https://aspredicted.org/dpy9-6ffy.pdf) for the learning data. We performed post hoc tests on the rating data.

We fit a model on the rating data and one set of models on the learning data to assess participant accuracy across conditions. For the rating data, in order to see whether "natural" ratings correlate with our modulation of the voices, we fit a Bayesian ordinal regression model predicting the Likert scale ratings for naturalness from voice type (original/flipped), using weakly informative priors. For participant accuracy, we fit a logistic regression model on the learning data predicting whether a response to an item was accurate (= the participant gave

the appropriate label for an object). The predictors included: alignment between meaning and voice (aligned = a female voice read the items with female associations, and a male voice read the items with male associations, mismatched = a male voice read the items with female associations, a female voice read the items with male associations), voice type (original AI voice, flipped voice modulated to have mismatched gender cues), block (same voice or novel voice, see 2.2, **Figure 1**) and the participants' rating on whether the specific voice "sounds natural". That is, we used the "natural" rating as a proxy for other ratings in our model of participant accuracy. The model had a grouping factor for participants and items. The best model has a three-way interaction between alignment, voice type, and block.

We deviated from the pre-registration in three ways. First, we did not remove participants or trials based on the exclusion criteria, since we deemed task accuracy sufficient to reliably assess participant attention. Also, since we did not put enough emphasis on response speed in participant instructions, some participants might have been fully paying attention and still have been extra slow. Test and generalization trials required the participant to select one out of 16 options, and the relative difficulty of this trial type likely increased variability in trial duration. Second, we ended up with more participants than the sample size put forward in the preregistration (231 participants vs. 200 pre-registered). We did not stop recruitment based on observing the data, and since the final sample size was only 115% of the preregistered size, we decided not to discard extra participants. Third, we added participant ratings of the voices' typicality as a predictor in the learning analysis to account for the participants' subjective response to the voices.

3. Results

Because the interpretation of the results hinges on the perception of gender, we will begin by outlining gender perception across conditions. The original voices were identified as their intended gender in 100% of cases for female speakers, and in 96% of cases for male speakers. The voice flip condition succeeded in reversing these judgments: in this condition, female speakers were identified as female in only 5% of cases, while male speakers were identified as male in only 15% of cases. We were also interested in the graded measure of gender typicality. We fit an ordinal model predicting the Likert scale response on gender typicality, using the voice flip condition as the predictor. The ordinal model revealed a meaningful effect for the voice flip condition (est = –0.81, se = 0.22, 95% CI: [–1.26; –0.4]), where original voices were rated as more typical for their gender than flipped voices. To present this in terms of the response scale (1 = Very unusual, 5 = Very typical), for flipped voices, the predicted probabilities were highest for a rating of 4 (0.53) and 3 (0.29), with very few predicted responses of 5 (0.03). In contrast, for original voices, the model predicted a higher probability of a rating of 4 (0.69), a lower probability of a rating of 3 (0.14), and a much higher probability of 5 (0.14), indicating a notable shift toward less typical gender ratings in the flipped condition. The gender identity of participants is notably absent from this analysis – we return to this at the end of this section.

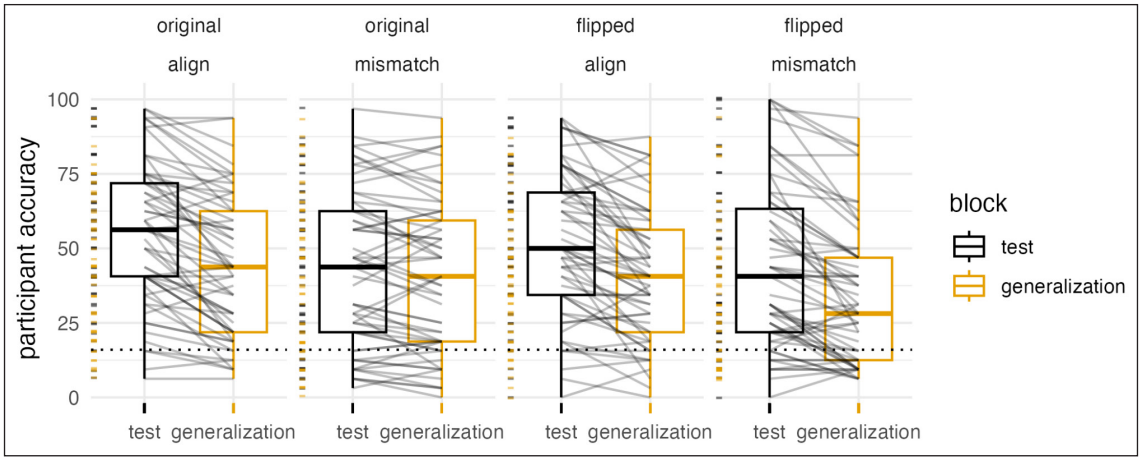

Figure 3: Participant means across experiment block (test or generalization), voice-item alignment (aligned or mismatched) and AI voice type (original or flipped). Each participant responded to both test and generalization trials (within-participant condition). Each participant encountered either aligned or mismatched and either original or flipped voices (across-participant conditions). The line segments mark individual participant accuracy in test and generalization. The dotted horizontal line shows chance level.

Raw participant means across conditions can be seen in **Figure 3**. Participants were broadly more accurate in the aligned condition than in the mismatched condition, and this was particularly visible for test trials (as opposed to generalization trials) and participants training and testing with the original AI voices (as opposed to the flipped voices). However, aggregating over the raw data is a poor proxy for the signal in the experiment, because it masks two important facets of the data, namely, that (a) participants responded to individual items, which is not shown by the means, and (b) the subjective experience of participants was likely more relevant for their accuracy than our labels for the AI voices.

Participants rated voices using several Likert scales: sounds feminine or masculine, creepy sounding, natural sounding, speaks clearly, sounds typical for their gender, easy to understand, and would use in the future. **Figure 4** shows the Pearson correlations between ratings across participants. The “stereotypical” rating is included for the reader’s convenience. It is based on the feminine/masculine rating applied to female or male voices. If the voice is originally female, stereotypicality refers to the female end of the rating scale; if the voice is originally male, the other way around. For simplicity’s sake, this includes both the original and flipped voices. Across rating scales, responses are moderately correlated, indicating that (a) participants evaluated the voices holistically, and (b) voice gender perception was related to the other dimensions of the rating scales: if a participant rates a voice as less gender-confirming, it also sounds less natural and harder to understand.

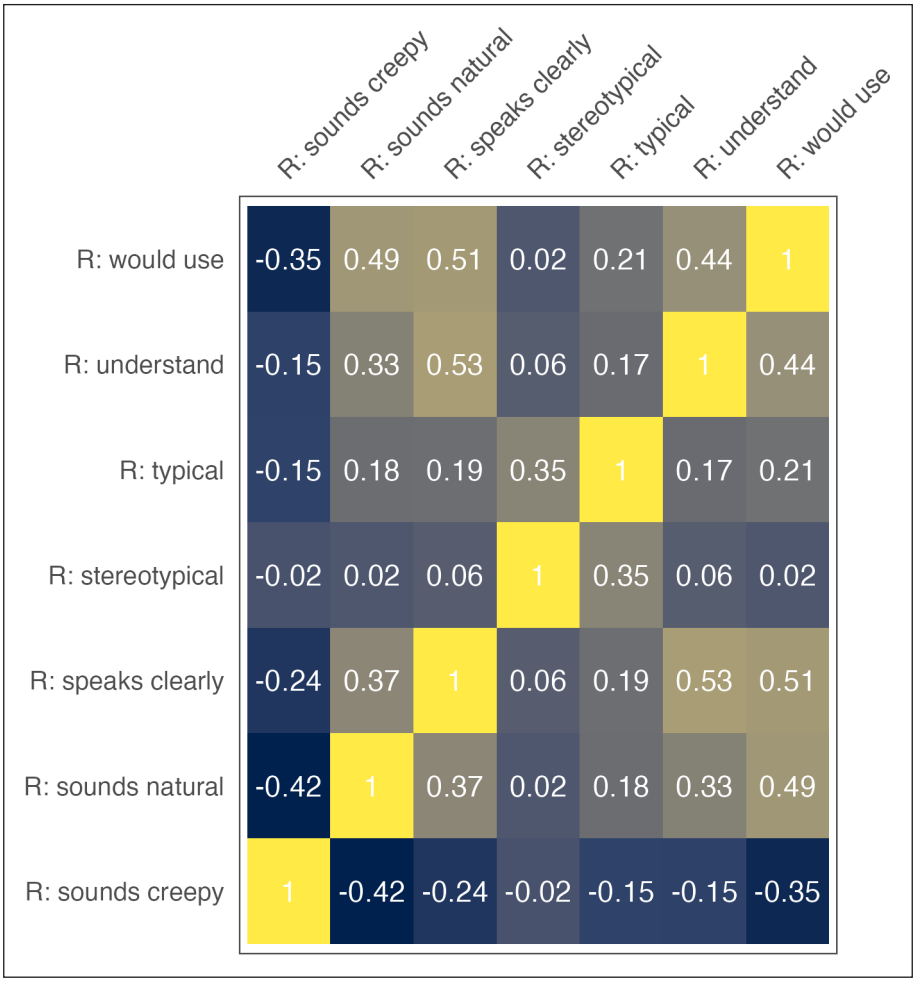

Figure 4: Pearson correlations between participant ratings of the voices used in the experiment.

Principal components analysis reveals a stronger natural/clear/non-creepy dimension and a weaker typical/gender-stereotypical dimension. These are visible in the correlations in **Figure 4**, as well. We used the "sounds natural" rating as our benchmark (as opposed to the "creepy" or "would use" rating, etc.) because it shows the strongest correlation with the other ratings (though not "typical") and it is relatively straightforward to interpret.

Once we establish that "sounds natural" works as a good benchmark of how participants perceive the voices, we can go on to look at a model of participant accuracy that incorporates this subjective dimension. The estimates of our best model of participant accuracy across conditions can be seen in **Figure 5**. Treatment coding was used, so the intercept captures the model's baseline, a response to an object matching the voice gender, with an original voice in the same-talker test block. The model uses the logit scale. Response accuracy in this group is between

–0.18 and 0.49; this translates to an accuracy of 45 – 62%, which is much higher than the chance accuracy of 6.25% (since each trial had 16 possible response choices). The rest of the estimates show relative accuracy compared to the intercept. If the model's estimate largely excludes zero (the dotted vertical line) and has a magnitude that is large enough to noticeably affect outcomes, the effect is likely meaningful. For example, if a participant rated a voice as more natural, the participant was also more accurate in the experiment. The rest of the estimates are harder to interpret in isolation, because they take part in a three-way interaction between object-voice alignment, experiment block, and AI voice type.

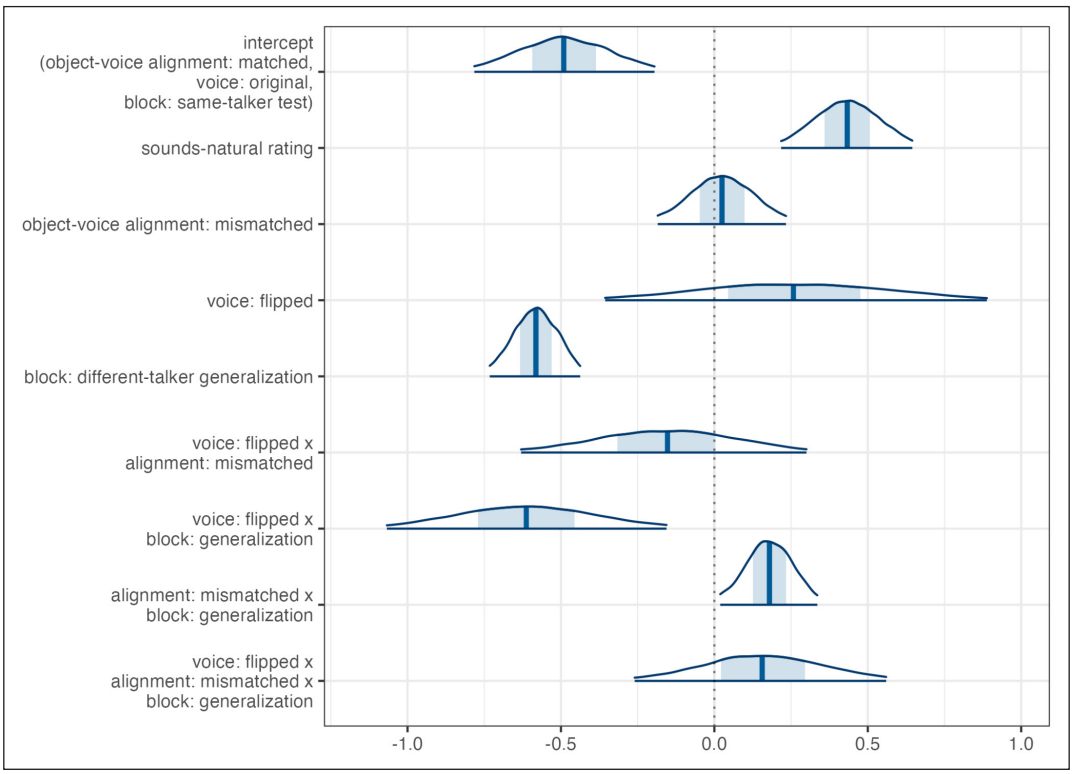

Figure 5: Posterior distributions from the best model fit on the learning data. The vertical lines show the median, the shaded areas, and the 50%, the outlines, the 95% credible intervals.

We can better understand this interaction if we look at **Figure 6**, which shows model predictions. We see predicted values for participants training and testing with original (left panel) and flipped AI voices (right panel). The predictions are similar to the raw data in **Figure 3**. We see what appears to be an interaction between alignment and mismatch in the original condition, but not in the flipped condition (right panel). In the flipped condition, we see additive effects for generalization and alignment, with no interaction between the two, such that the lines in the plot are nearly parallel. One way to interpret this is that word-voice alignment is easier than word-voice mismatch, but this is mostly seen in the test phase and for the original voices. The

aligned/mismatched difference is much smaller for flipped voices. Generalization is harder than test across the board, though still above chance level (1/16).

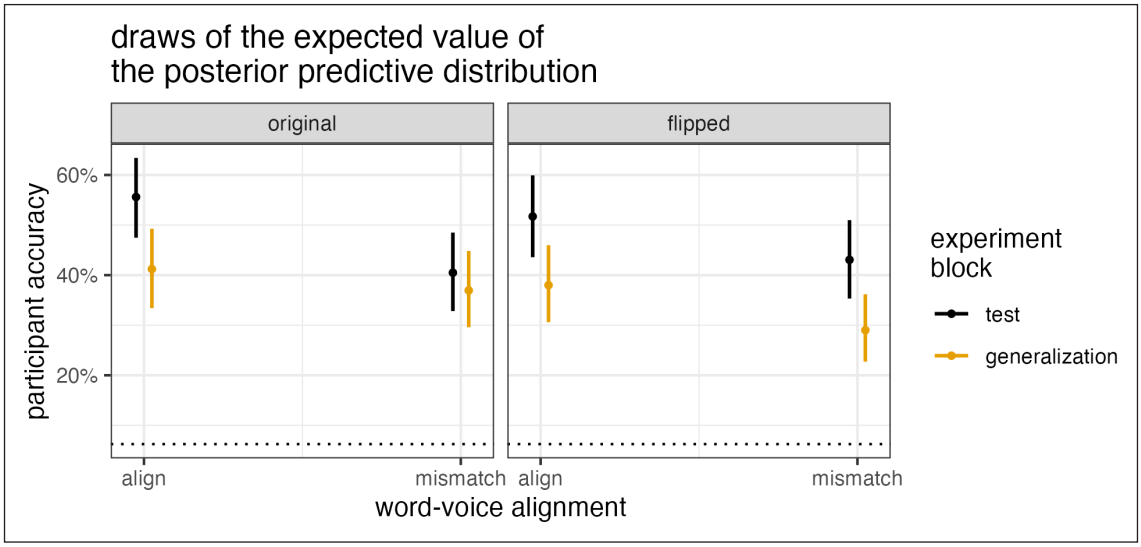

Figure 6: Samples (draws) of expected values of the posterior predictive distributions from the model fit on the learning data. Points show medians. Line ranges show 95% credible intervals. Predictions transformed from log odds to percentages. The dotted horizontal line shows chance level.

We can explicitly tie our results to our hypotheses using this model, detailed above, and Bayesian hypothesis testing.

- H1. Participants were less accurate with word-voice mismatch than with word-voice alignment (est = –0.61, se = 0.23, 95% CI: [–0.99; –0.22]).
- H2. Participants were less accurate in the generalization block than in the test block (est = –0.58, se = 0.08, 95% CI: [–0.71; –0.46]). However, participants were still above chance (p = 1/16 or lo = -2.71) in the generalization block (est = 2.28, se = 0.02, 95% CI: [1.87; 2.69]).
- H3. The mismatch effect was weaker with flipped AI voices. If participants were trained with a flipped AI voice, there was a much smaller difference between the aligned and mismatched conditions (est = 0.42, se = 0.51, 95% CI: [–0.4;1.26]).
- H4. If a participant found a voice more natural, their responses were more accurate (est = 0.18, se = 0.08, 95% CI: [0.04; 0.31]). Voice naturalness best captures the behaviour of other ratings and robustly correlates with voice type (original or flipped), so we used this as the proxy for the participant preference and competence judgements. At the same time, naturalness ratings explain additional variation beyond voice type. That is, the subjective experience of participants does not reduce to our labels for the voices.

Any discussion of how participants reacted to voice gender directly ties in with participant gender identity. However, 175 of our 231 participants are women, and 174 are cis women. As a result, we cannot provide a comprehensive account of the effects of listener gender identity across our conditions. However, participants were randomly assigned to conditions. When we look at participant random intercepts in our best model, which express an overall rate of accuracy for participants, above and beyond treatment effects, we do not see that ciswomen and the other groups in the sample cluster or pattern differently. This can be seen in **Figure 7**. This strongly suggests that, at least in the sample, participant gender identity did not have a major, discernible effect on participant accuracy.

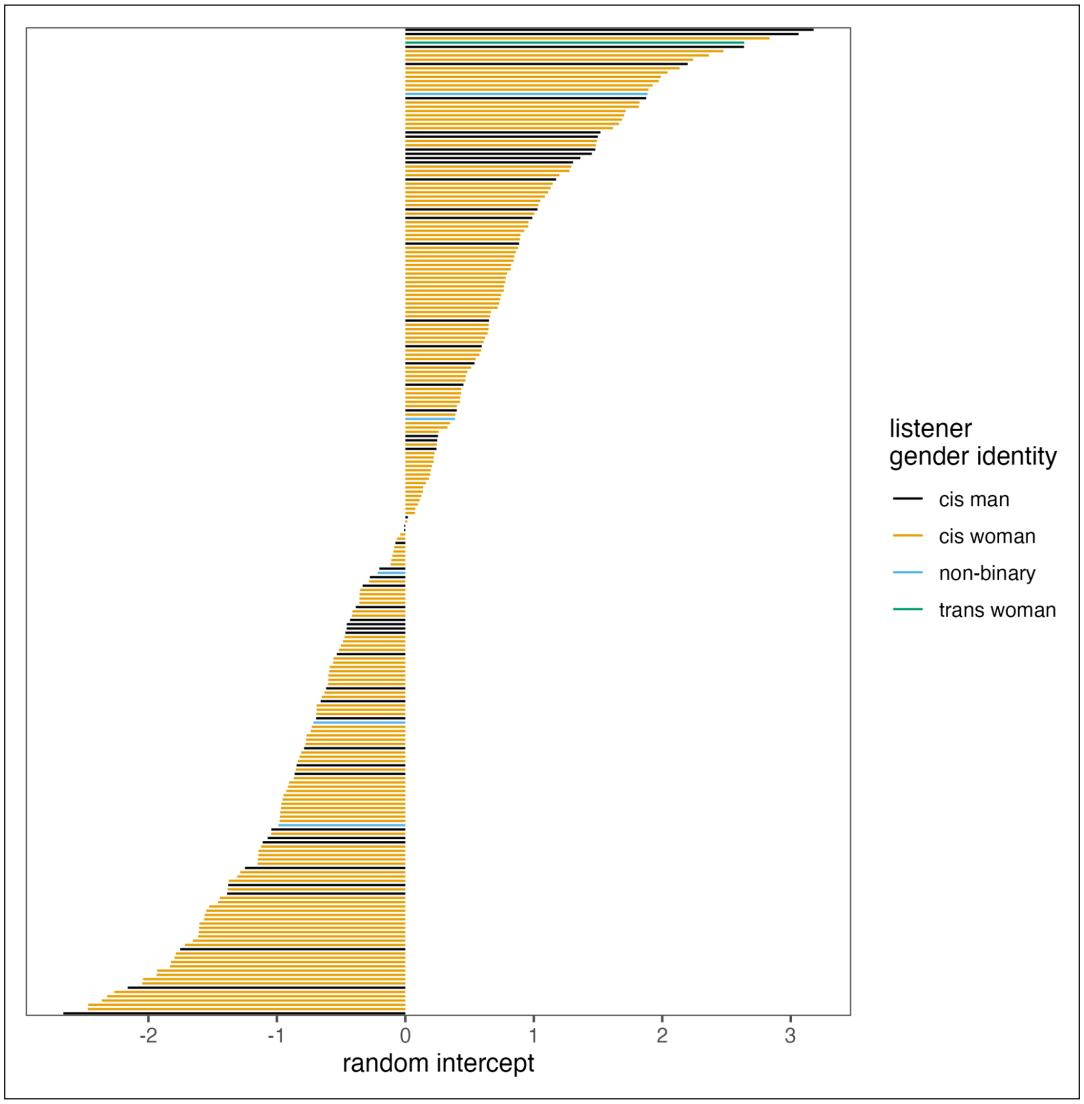

Figure 7: Listener random intercepts and listener gender identity in our best model.

4. General discussion

The current study investigated the impact of social associations between a speaker's voice acoustics and lexical meaning on the learning of novel words, using a paradigm that simulates a language learning app. In the subsections below, we discuss our findings in detail, and interpret their implications for learning at the intersection of speech variation, society, and technology.

4.1 Role of mismatching voice and semantic gender cues

Our key finding is that participants learned words less well when the gender associations of their meanings mismatched with the gross gender acoustics of the voices (H1). We observed this across two separate experiments in which voices contained stereotypical and non-stereotypical fine-grained acoustic features.

These results can contribute to our understanding of the role of socio-indexical information on language processing and learning. Prior work shows that the statistical distribution of real words across social groups affects the usage of novel pseudo-words (Needle & Pierrehumbert, 2018) and that associations between language variation and speaker gender in an artificial language are learned (Rácz et al., 2017). The present study extends this to the auditory domain and novel word learning, by finding that learning performance is lower when gendered associations of the meaning and the voice are mismatching. Why is this the case? Since prior work shows that the social distribution in which people experience words affects the processing and comprehension of their first language (Kim, 2016; McGowan, 2015; Niedzielski, 1999; Rubin, 1992; Strand, 1999; Zellou et al., 2023), such statistical patterns could be transferred when learning and processing new words. For instance, the real world statistical and episodic experience people have with words in their native language is proposed to form the representations from which listeners process language (Johnson & Sjerps, 2021; Pierrehumbert, 2016). Such patterns might include voices that have female-presenting attributes (e.g., higher f0) producing words that evoke concepts such as those used in the present study (e.g., dresses).

Thus, the learning of new lexical forms can be supported when the gendered associations between the voice and meaning align with past experiences learners have from their first language. More broadly, this finding supports models of cognitive processing in which sensory inputs containing features that align with prior experience are easier to process and, therefore, easier to learn and remember (e.g., Westerman et al., 2002), as well as models of memory in which stimuli containing expected information are encoded and recognized better (e.g., Bein et al., 2015). This is consistent with prior work examining how the reliability of social cues in the input guides attention and learning: when social cues are incongruent, learners might find the cues less reliable and divert their attention to tracking multiple hypotheses about the sources of the uncertainty. This distributed attention can hinder learning performance. The results in

the present study are also consistent with work showing that mismatches in expectations about social associations lead to reduced language comprehension (i.e., seeing an Asian face and hearing native-English accented speech in Rubin, 1992; cf. increased comprehension when there is alignment between an Asian face and hearing Mandarin-accented speech in McGowan, 2015). Thus, when multiple social cues align with listeners' prior expectation, this allows them to devote attentional resources more robustly to the encoding and storage of new word forms, as well.

Of course, explanations that hinge on statistical relations derived from previous experience rely on the existence of a veridical empirical relationship between, for example, "female" semantic content (e.g., *flowers*) and expected female speech acoustics. Another possibility is that listener expectations, even when these are only tenuously connected to the empirical facts, can affect perception. For example, Barreda and Predeck (2024) found that perceived gender affects the apparent size of speakers independently of speech acoustics. In a whispered speech condition, listeners rated the same tokens as produced by taller speakers when they identified the speakers as male, and shorter when they identified them as female. In fact, listeners identified all male speakers as taller than all female speakers, even when vocal-tract length cues were equivalent, and even when the female speakers were, in fact, taller than male speakers in the experiment. In this case, the general expectation that men are taller than women overrides the empirical facts of the matter, which indicate that speakers who produce approximately the same formant frequencies tend to be approximately the same size (Turner et al., 2009).

4.2 Generalization of learning

Our experiments also tested the extent to which participants generalized learning to new voices – a hallmark indicator of successful, productive language learning. Participant learning performance was above chance across all groups. Thus, a second key finding is that learning did generalize to different voices. Yet, performance was reduced in the generalization condition, which is consistent with prior work indicating that cross-talker learning can be weaker than same-talker learning (e.g., Eisner & McQueen, 2005; Xie et al., 2021; Xie & Myers, 2017).

With respect to theories about how attention shapes learning and generalization, we propose that adult word learning in voice-AI contexts depends on (a) prior expectations about voice-semantic content associations, (b) the reliability of social cues, and (c) processing fluency. Violations of expectations that increase informativity under reliable cueing (e.g., social cues in voices that align with prior expectations about distributions between social and semantic information) can boost encoding, leading to stronger learning and generalization, whereas violations that undermine cue reliability (e.g., incongruent voice–content pairs) can hinder learning.

4.3 How voice typicality modulates these effects

Another key finding is how voice typicality modulated learning across conditions. We find that these differences in learning performance are reduced if the acoustic gender cues of the text-to-speech voices are flipped (H2). More specifically, overall, the difference in accuracy was actually diminished in the flipped voices condition, contrary to our expectation that presenting non-stereotypical voices would make learning in the mismatched condition harder (H3, H4). This is likely because overall accuracy was lower in the flipped voices condition, so that smaller differences did not have the space to shine, and, possibly, because the alignment-mismatch distance was simply smaller for the gender-flipped voices, such that training with a female item and a male voice was less taxing if the voice sounded less stereotypically male. An additional, noteworthy aspect of our data is that flipped voices are hard to work with: participants find it difficult to generalize to new items when training and testing with a flipped voice. This suggests that voices that defy gendered expectations regarding gross acoustics (average f0 or formant frequencies) and more subtle aspects of gender in voice incur additional processing costs that generally make learning more difficult.

Thus, the degree to which a voice conforms to expectations regarding gross speech acoustics (e.g., average f0) and more subtle variation related to gender (e.g., specific intonation patterns) can influence the learning of novel words. Because of the arbitrariness of most gender-related speech variation, it is important to note that what is typical or atypical in the opinion of listeners is socially constructed and may be largely inconsistent across languages and cultures. As a result, the atypicality of some voices, and the associated processing costs, stem from the listener's expectations regarding what male and female speakers "should" sound like, rather than from any inherent property held by these so-called "atypical" speakers. Thus, exploring how linguistic stereotyping, word learning, and human-computer interaction changes with a range of gender-diverse voices is important to understanding how listener expectations regarding gender shape language learning and processing.

4.4 The effect of the learning exposure condition on voice ratings

In general, our ratings study yielded mostly unsurprising results: participants used gross gender acoustics to categorize female voices as female and male voices as male. Participants made a clear difference between flipped and unmodified voices. If the gender cues were flipped, participants were more likely to classify female voices as male and give female voices higher masculinity ratings (and the other way around). Participants also found flipped voices less natural, which, overall, accounted for most of the ratings differences across voices.

An unexpected finding from our ratings was the effect that the *alignment* condition in the learning experiment had on participants' later ratings of the voices. When participants were trained on flipped voices, they provided lower naturalness ratings of the voices.

4.5 Implications for language learning via voice-AI

There are important implications for the role of gendered associations in language learning using devices. In the current study, the effect of voice gender on learning occurred in a context where participants were told they were using a word learning computer application, not actual human language teachers. Computers and applications do not have bodies or genders – they are machines. Yet, as demonstrated by Nass et al. (1997) and Ernst and Herm-Stapelberg (2020), people still apply human-based social patterns of behavior to computers when they display cues of humanness, such as using voices with apparent genders. There are societal implications of the observation that race and gender biases impact behavior during interactions with speech and language technology. More generally, such biases could reinforce and amplify societal stereotypes. For instance, for gender, having female voice assistant voices as the default could amplify the stereotype that women are the helpers, rather than the ones in charge (West et al., 2019). This can explicitly and implicitly send users the message that women and girls should communicate in the way voice assistants do – by being subservient, pleasant, and not questioning.

Moreover, our stereotypicality manipulation is also relevant for understanding speech communication with devices. Our unmodified experiment utilized widely-available text-to-speech voices. The current state of speech technology is skewed toward providing voices to users that reflect a binary. Unlike in the real world, where people experience human talkers who express a range of gender identities, the overwhelming majority of text-to-speech voices and voice-assistant personae are presented to reflect stereotypical expressions of binary gender identity: non-binary, gender non-conforming, and gender-neutral voices are not available in commercially-available systems (Sutton, 2020; though genderless voices have been developed by various researchers: e.g., Project Q; Assink, 2021; Carpenter, 2019; Chaves, 2021).

Our findings reveal that people extend social expectations from human speakers to synthetic voices; thus, design choices about apparent gender and cue reliability can affect word learning and evaluation in speech and language technologies. Systems that allow gender-diverse and less stereotyped voice options, and that align social cues with instructional content, may improve perceptual fluency without reinforcing harmful stereotypes.

5. Conclusion

Language acquisition is challenging for adult learners. One reason for this is that the linguistic and cultural associations adults have, based on the statistical patterns they experience in their native language, make some words even harder to learn than others. Gendered associations between speaker and word frequency, for instance, create an expectation that women and men have distinct language usage patterns through which incoming experiences are filtered. Under that premise, the current study tested whether adults show differences in word learning performance

of an artificial language in which lexical items are presented in either a male or a female voice. Participants in gender-aligning training conditions learned the novel words at higher rates – for both the same and other voice – than participants in gender-mismatching conditions. Ratings of the voices across conditions were not different, indicating that the gendered associations have an implicit effect on language learning. These results have significant practical and societal implications for the role of social cognition and biases in second language learning, speech and language technology, and human-computer interaction. Finally, our results suggest that "atypical" voices may incur increased processing costs, which can hinder learning by incurring additional processing costs. Though some may interpret this as motivation to avoid the use of such "atypical" voices, we suggest the opposite is the case. By increasing listeners' exposure to a wide range of voices, the range considered atypical may be narrowed, and the potential processing costs associated with the voices of such speakers may be minimized.

Appendix

	Novel Word (IPA)	Image	Mean Imageability Rating	Mean Gender-association Rating
Female-associated items	fɑnɛt	lipstick	6.7	1.3
	ɛgun	skirt	6.5	1.4
	gitu	handbag	6.8	1.4
	utɑf	dress	6.8	1.5
	tɑfun	necklace	6.5	1.5
	ɛfit	flower	6.8	1.7
	tugi	bracelet	6.5	1.7
	itun	perfume	5.4	1.8
Male-associated items	funɑ	motorbike	6.8	6.1
	fɛti	screwdriver	6.9	6.1
	gɑfun	suit	6.5	6.1
	ɛtug	football	6.8	6.2
	nigɛf	tie	6.6	6.2
	nɑfit	steak	6.6	5.9
	iguf	hammer	6.6	6
	nutig	drill	6.5	6

Data accessibility statement

All data and code for this article are available at: https://doi.org/10.5281/zenodo.18367222.

Ethics and consent

Research was performed in accordance with the UCD IRB (IRB number 1328085).

Competing interests

The authors have no competing interests to declare.

Authors' contributions

GZ conceptualized the study. GZ, PR, and SB designed the experiment, and wrote and edited the paper. PR performed the statistical analysis. All authors contributed to the article and approved of the final version.

ORCiD IDs

Georgia Zellou: https://orcid.org/0000-0001-9167-0744

Péter Rácz: https://orcid.org/0000-0001-7896-4801

Santiago Barreda: https://orcid.org/0000-0002-1552-083X

References

Ammari, T., Kaye, J., Tsai, J. Y., & Bentley, F. (2019). Music, search, and IoT: How people (really) use voice assistants. *ACM Transactions on Computer-Human Interaction (TOCHI), 26*(3), 1–28. https://doi.org/10.1145/3311956

Assink, L. M. (2021). *Making the invisible visible: Exploring gender bias in AI voice assistants*. [Master's thesis]. University of Twente.

Barreda, S. (2020). Vowel normalization as perceptual constancy. *Language, 96*(2), 224–254. https://doi.org/10.1353/lan.2020.0018

Barreda, S., & Assmann, P. F. (2021). Perception of gender in children's voices. *The Journal of the Acoustical Society of America, 150*(5), 3949–3963. https://doi.org/10.1121/10.0006785

Barreda, S., & Predeck, K. (2024). Inaccurate but predictable: Vocal-tract length estimation and gender stereotypes in height perception. *Journal of Phonetics*, *102*, 101290. https://doi.org/10.1016/j.wocn.2023.101290

Baugh, J. (2016). Linguistic profiling and discrimination. In O García, N Flores & M Spotti (Eds.), *The Oxford handbook of language and society* (pp. 349–368). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190212896.013.13

Bein, O., Livneh, N., Reggev, N., Gilead, M., Goshen-Gottstein, Y., & Maril, A. (2015). Delineating the effect of semantic congruency on episodic memory: The role of integration and relatedness. *PLoS One, 10*(2), e0115624. https://doi.org/10.1371/journal.pone.0115624

Bovolenta, G., & Marsden, E. (2021). Expectation violation enhances the development of new abstract syntactic representations: Evidence from an artificial language learning study. *Language Development Research, 1*(1), 193–243. https://doi.org/10.31219/osf.io/zyegf

Brandstetter, J., Rácz, P., Beckner, C., Sandoval, E. B., Hay, J., & Bartneck, C. (2014, September). A peer pressure experiment: Recreation of the Asch conformity experiment with robots. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 1335–1340). IEEE. https://doi.org/10.1109/IROS.2014.6942730

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*, 1–28. https://doi.org/10.18637/jss.v080.i01

Carpenter, J. (2019). Why project Q is more than the world's first nonbinary voice for technology. *Interactions, 26*(6), 56–59. https://doi.org/10.1145/3358912

Chaves, C. (2021). *Voice as identity: Creating a genderless voice assistant.* [Doctoral dissertation]. San Francisco State University. https://doi.org/10.46569/20.500.12680/1g05fh80z

Cheshire, J. (2004). Sex and gender in variationist research. In J. K. Chambers, P. Trudgill, & N. Schilling-Estes (Eds.), *The handbook of language variation and change* (pp. 423–443). Wiley. https://doi.org/10.1002/9780470756591.ch17

Coates, J. (2015). *Women, men and language: A sociolinguistic account of gender differences in language.* Routledge. https://doi.org/10.4324/9781315835778

De Renesse, R. 2017. Virtual digital assistants to overtake world population by 2021. *Ovum*, May, 17. Retrieved November 26, 2020, from https://ovum.informa.com/resources/product-content/virtual-digital-assistants-to-overtakeworld-population-by-2021

Drager, K., & Kirtley, M. J. (2016). Awareness, salience, and stereotypes in exemplar-based models of speech production and perception. In K. Campbell-Kibler & A. Babel (Eds.), *Awareness and control in sociolinguistic research* (pp. 1–24). Cambridge University Press. https://doi.org/10.1017/CBO9781139680448.003

Eckert, P. (1989). The whole woman: Sex and gender differences in variation. *Language Variation and Change, 1*(3), 245–267. https://doi.org/10.1017/S095439450000017X

Eckert, P., & McConnell-Ginet, S. (2013). *Language and gender*. Cambridge University Press. https://doi.org/10.1017/CBO9781139245883

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67*(2), 224–238. https://doi.org/10.3758/BF03206487

Ernst, C. P. H., & Herm-Stapelberg, N. (2020). The impact of gender stereotyping on the perceived likability of virtual assistants. In *AMCIS Proceedings 4*. https://aisel.aisnet.org/amcis2020/cognitive_in_is/cognitive_in_is/4

Feher, O., Kirby, S., & Smith, K. (2014). Social influences on the regularization of unpredictable linguistic variation. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 36, No. 36).

Gambino, A., & Liu, B. (2022). Considering the context to build theory in HCI, HRI, and HMC: Explicating differences in processes of communication and socialization with social technologies. *Human-Machine Communication, 4*, 111–130. https://doi.org/10.30658/hmc.4.6

Girard, F., Floccia, C., & Goslin, J. (2008). Perception and awareness of accents in young children. *British Journal of Developmental Psychology, 26*(3), 409–433. https://doi.org/10.1348/026151007X251712

Godwin-Jones, R. (2011). Mobile apps for language learning. *Language Learning & Technology, 15*(2), 2–11. http://llt.msu.edu/issues/june2011/emerging.pdf

Godwin-Jones, R. (2017). Smartphones and language learning. *Language Learning & Technology, 21*(2), 3–17. http://llt.msu.edu/issues/june2017/emerging.pdf

Guaqueta, C. A., & Castro-Garces, A. Y. (2018). The use of language learning apps as a didactic tool for EFL vocabulary building. *English Language Teaching, 11*(2), 61–71. https://doi.org/10.5539/elt.v11n2p61

Hay, J., Nolan, A., & Drager, K. (2006). From fush to feesh: Exemplar priming in speech perception. *Linguistic Review, 23*(3), 351–379. https://doi.org/10.1515/TLR.2006.014

Johnson, K., & Sjerps, M. J. (2021). Speaker normalization in speech perception. In J. S. Pardo, L. C. Nygaard, R. E. Remez, & D. B. Pisoni (Eds.), *The handbook of speech perception* (pp. 145–176). Wiley. https://doi.org/10.1002/9781119184096.ch6

Kaushanskaya, M., & Marian, V. (2009). Bilingualism reduces native-language interference during novel-word learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(3), 829–835. https://doi.org/10.1037/a0015275

Kim, J. (2016). Perceptual associations between words and speaker age. *Laboratory Phonology*, *7*(1). https://doi.org/10.5334/labphon.33

Kuchenbrandt, D., Häring, M., Eichberg, J., Eyssel, F., & André, E. (2014). Keep an eye on the task! How gender typicality of tasks influence human–robot interactions. *International Journal of Social Robotics, 6,* 417–427. https://doi.org/10.1007/s12369-014-0244-0

Labov, W. (2006). *The social stratification of English in New York City*. Cambridge University Press. https://doi.org/10.1017/CBO9780511618208

Lee, C., & Lew-Williams, C. (2022). Speech and social cues combine at discourse boundaries to promote word learning. *Cognitive Development, 64*, 101254. https://doi.org/10.1016/j.cogdev.2022.101254

Lüdecke, D., & Lüdecke, M. D. (2015). Package 'sjplot'. *R package Version*, *1*(9), 1–106.

MacDonald, K., Yurovsky, D., & Frank, M. C. (2017). Social cues modulate the representations underlying cross-situational learning. *Cognitive Psychology, 94*, 67–84. https://doi.org/10.1016/j.cogpsych.2017.02.003

McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Language and Speech, 58*(4), 502–521. https://doi.org/10.1177/0023830914565191

Moore, R. K. (2012). A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. *Scientific Reports, 2*(1), 864. https://doi.org/10.1038/srep00864

Most, S. B., Sorber, A. V., & Cunningham, J. G. (2007). Auditory Stroop reveals implicit gender associations in adults and children. *Journal of Experimental Social Psychology, 43*(2), 287–294. https://doi.org/10.1016/j.jesp.2006.02.002

Munson, B., Lackas, N., & Koeppe, K. (2022). Individual differences in the development of gendered speech in preschool children: Evidence from a longitudinal study. *Journal of Speech, Language, and Hearing Research, 65*(4), 1311–1330. https://doi.org/10.1044/2021_JSLHR-21-00465

Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology, 27*(10), 864–876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x

Needle, J., & Pierrehumbert, J. (2018). Gendered associations of English morphology. *Laboratory Phonology, 9*(1). https://doi.org/10.5334/labphon.134

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology, 18*(1), 62–85. https://doi.org/10.1177/0261927X99018001005

Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics*, *2*(1), 33–52. https://doi.org/10.1146/annurev-linguistics-030514-125050

Poon, M. S., & Ng, M. L. (2015). The role of fundamental frequency and formants in voice gender identification. *Speech, Language and Hearing, 18*(3), 161–165. https://doi.org/10.1179/2050572814Y.0000000058

Prystawski, B., Grant, E., Nematzadeh, A., Lee, S. W., Stevenson, S., & Xu, Y. (2020). Tracing the emergence of gendered language in childhood. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 42).

R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rácz, P., Hay, J. B., & Pierrehumbert, J. B. (2017). Social salience discriminates learnability of contextual cues in an artificial language. *Frontiers in Psychology, 8*, 207450. https://doi.org/10.3389/fpsyg.2017.00051

Roberts, G. (2017). The linguist's Drosophila: Experiments in language change. *Linguistics Vanguard*, *3*(1), 20160086. https://doi.org/10.1515/lingvan-2016-0086

Robertson, K., & Murachver, T. (2003). Children's speech accommodation to gendered language styles. *Journal of Language and Social Psychology, 22*(3), 321–333. https://doi.org/10.1177/0261927X03255304

Rubin, D. L. (1992). Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education, 33*, 511–531. https://doi.org/10.1007/BF00973770

Scott, G. G., Keitel, A., Becirspahic, M., Yao, B., & Sereno, S. C. (2019). The Glasgow Norms: Ratings of 5,500 words on nine scales. *Behavior Research Methods, 51*(3), 1258–1270. https://doi.org/10.3758/s13428-018-1099-3

Smith, J., Durham, M., & Fortune, L. (2009). Universal and dialect-specific pathways of acquisition: Caregivers, children, and t/d deletion. *Language Variation and Change, 21*(1), 69–95. https://doi.org/10.1017/S0954394509000039

Stan Development Team. (2024). RStan: the R interface to Stan. R package version. (https://mc-stan.org/rstan/)

Stennes, L. M., Burch, M. M., Sen, M. G., & Bauer, P. J. (2005). A longitudinal study of gendered vocabulary and communicative action in young children. *Developmental Psychology, 41*(1), 75–88. https://doi.org/10.1037/0012-1649.41.1.75

Strand, E. A. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of Language and Social Psychology, 18*(1), 86–100. https://doi.org/10.1177/0261927X99018001006

Sutton, S. J. (2020, July). Gender ambiguous, not genderless: Designing gender in voice user interfaces (VUIs) with sensitivity. In *Proceedings of the 2nd conference on conversational user interfaces* (pp. 1–8). https://doi.org/10.1145/3405755.3406123

Turner, R. E., Walters, T. C., Monaghan, J. J., & Patterson, R. D. (2009). A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data. *The Journal of the Acoustical Society of America, 125*(4), 2374–2386. https://doi.org/10.1121/1.3079772

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*(3), 219–232. https://doi.org/10.1177/1745691610369336

West, M., Kraut, R., & Chew, H. E. (2019). I'd blush if I could: closing gender divides in digital skills through education. https://unesdoc.unesco.org/ark:/48223/pf0000367416, (EQUALS/UNESCO).

Westerman, D. L., Lloyd, M. E., & Miller, J. K. (2002). The attribution of perceptual fluency in recognition memory: The role of expectation. *Journal of Memory and Language, 47*(4), 607–617. https://doi.org/10.1016/S0749-596X(02)00022-0

Wickham, M. H. (2016). Package 'ggplot2'. Create elegant data visualisations using the grammar of graphics. Version 2.1, 1–189.

Wong, P. C., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics, 28*(4), 565–585. https://doi.org/10.1017/S0142716407070312

Xie, X., Liu, L., & Jaeger, T. F. (2021). Cross-talker generalization in the perception of nonnative speech: A large-scale replication. *Journal of Experimental Psychology: General, 150*(11), e22. https://doi.org/10.1037/xge0001039

Xie, X., & Myers, E. B. (2017). Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *Journal of Memory and Language, 97*, 30–46. https://doi.org/10.1016/j.jml.2017.07.005

Zellou, G., Cohn, M., & Ferenc Segedin, B. (2021). Age-and gender-related differences in speech alignment toward humans and voice-AI. *Frontiers in Communication, 5*, 600361. https://doi.org/10.3389/fcomm.2020.600361

Zellou, G., Cohn, M., & Kline, T. (2021). The influence of conversational role on phonetic alignment toward voice-AI and human interlocutors. *Language, Cognition and Neuroscience, 36*(10), 1298–1312. https://doi.org/10.1080/23273798.2021.1931372

Zellou, G, Cohn, M., & Pycha, A. (2023). Listener beliefs and perceptual learning: Differences between device and human guises. *Language, 99*(4), 692–725. https://doi.org/10.1353/lan.2023.a914191