Listener beliefs and perceptual learning: Differences between device and human guises

Georgia Zellou, Michelle Cohn, Anne Pycha

Language, Ahead of Print, (Article)

Published by Linguistic Society of America
DOI: *https://doi.org/10.1353/lan.0.a913403*

LISTENER BELIEFS AND PERCEPTUAL LEARNING: DIFFERENCES BETWEEN DEVICE AND HUMAN GUISES

Georgia Zellou
University of California, Davis

Michelle Cohn
University of California, Davis

Anne Pycha
University of Wisconsin-Milwaukee

Listeners have a remarkable ability to adapt to novel speech patterns, such as a new accent or an idiosyncratic pronunciation. In almost all of the previous studies examining this phenomenon, the participating listeners had reason to believe that the speech signal was produced by a human being. However, people are increasingly interacting with voice-activated artificially intelligent (voice-AI) devices that produce speech using text-to-speech (TTS) synthesis. Will listeners also adapt to novel speech input when they believe it is produced by a device? Across three experiments, we investigate this question by exposing American English listeners to shifted pronunciations accompanied by either a 'human' or a 'device' guise and testing how this exposure affects their subsequent categorization of vowels. Our results show that listeners exhibit perceptual learning even when they believe the speaker is a device. Furthermore, listeners generalize these adjustments to new talkers, and do so particularly strongly when they believe that both old and new talkers are devices. These results have implications for models of speech perception, theories of human-computer interaction, and the interface between social cognition and linguistic theory.*

Keywords: speech perception, accent adaptation, phonetic learning, listener beliefs, human-computer interaction

1. Introduction. Speech perception is a fundamentally social act. As early as 1957, Ladefoged and Broadbent conducted an elegant experiment to demonstrate that listeners did not judge a speech sound, such as a vowel, in isolation. Instead, listener judgments crucially relied upon information about the speaker—in that case, how the speaker pronounced other vowels in the surrounding sentence context. In the decades since, dozens of studies have found support for a basic tenet that a listener's perception of a speech signal depends crucially on their beliefs about the speaker (for review, see Drager 2010). These beliefs can be shaped by an impressive variety of influences. To take just one example, Johnson et al. (1999) showed that listeners shifted their category boundaries for vowels when they listened to a female versus a male voice. Social influence can also occur beyond the auditory domain. In the same study, listeners also shifted boundaries when they saw a visual image of a female versus a male speaker, and even when they simply imagined a female versus a male speaker. Other studies have demonstrated similar findings for a wide range of additional influences, including differences in perception based on speakers' apparent ages (e.g. Hay, Warren, & Drager 2006, Drager 2011), socioeconomic statuses, race/ethnicity (e.g. Staum Casasanto 2008), regional origins (e.g. Niedzielski 1999), and sexual orientations (e.g. Munson et al. 2006). Even seeing stuffed animals associated with different speaker groups (e.g. a kiwi cueing a New Zealand dialect) can shape speech perception (Hay & Drager 2010). The takeaway is that, depending on what the listener believes about the speaker, the same acoustic value—and in some cases, the exact same stimulus—can yield different linguistic interpretations.

To model the social nature of speech perception, many researchers have appealed to exemplar theories (Goldinger 1996, Johnson 1997, 2006, Pierrehumbert 2002, 2016,

* We thank the editorial team, Meredith Tamminga, Andries Coetzee, Shelome Gooden, and anonymous referees, for their detailed and very constructive comments. This research was supported by the National Science Foundation SBE Postdoctoral Research Fellowship to MC under Grant No. 1911855 and an Amazon research grant to GZ. Thanks to Bruno Ferenc Segedin for assistance with the stimuli.

Sumner et al. 2014). Within this framework, every instance of a heard word is stored in memory, along with its phonetic details. Exemplars that are acoustically similar can cluster together to form higher-level categories (e.g. the word *cat*; Johnson 2006). When listeners hear a new input, it activates similar-sounding exemplars. If activation is sufficient, the higher-level category is also activated and word recognition occurs ('I heard *cat*'). Crucially, higher-level categories are not limited to words; because all phonetic details are stored, exemplars can cluster together based on other dimensions of acoustic similarity, such as gender (Johnson 2006). Thus, when listeners hear a new input, it not only activates words (*cat*), but also activates categories such as 'female speaker' or 'male speaker'. This basic mechanism can account for social influences in the auditory domain. To account for influences in other domains, such as visual images, researchers have suggested that exemplars are not limited to acoustic information, but also contain rich contextual details from other modalities (Pierrehumbert 2016). Thus, an exemplar for *cat* may include, for example, information gleaned about the speaker from their physical characteristics, as well as other information from the particular context in which the word was uttered. Arguably, Ladefoged and Broadbent (1957:102) anticipated such a framework when they made the simple statement that 'the response evoked by a stimulus is influenced by the stimuli with which it is closely associated'.

By modeling the role of social information in this manner, exemplar theories also make an important claim about listeners: they are always adapting to new situations. Since every instance of speech input gets stored, every new exemplar can potentially alter the shape of clusters and update their associated categories. Within this perspective, then, learning the pronunciation patterns of an idiosyncratic talker or a new accent—a process known as PERCEPTUAL LEARNING, ADAPTATION, or PHONETIC RECALIBRATION—is not a problem to be solved, but a basic expectation about how speech perception works. (Note that we use the term *perceptual learning* in the current study, and also consider this operational term to be interchangeable with those other terms, following prior work (e.g. Norris et al. 2003).) This expectation is borne out in the experimental literature. For example, after only brief exposure to speech input that has been shaped by an unfamiliar accent, listeners can typically adapt to the shift, to ultimately comprehend the speaker's intended message (for review, see Baese-Berk 2018). Perceptual learning can be targeted, such as when listeners adapt their representations for specific vowels and consonants. For instance, following brief exposure to an idiosyncratic speaker, listeners will accept *wetch* [wɛʧ] as an instance of *witch* [wɪʧ] (Maye et al. 2008) or adjust the boundary between the consonant /s/ and the consonant /f/ (Norris et al. 2003). As we review below, the literature contains many additional examples of rapid perceptual learning, which attest to the dynamic nature of speech perception.

Like exemplar theories, the IDEAL ADAPTER FRAMEWORK (Kleinschmidt & Jaeger 2015, Kleinschmidt 2019) also predicts rapid adaptation by listeners, but with constraints on what information is represented. In this framework, individual episodes are not stored in memory in their entirety. Instead, listeners use both their past and present experiences of speech to select a generative model that, given a particular speech cue, estimates the most likely category. To take an example from American English, if the speech cue contains relatively high frication frequency, listeners may estimate that the category is /s/, rather than /ʃ/, since past experience has shown that /s/ tends to have higher frequencies. Sensitivity to social information is formalized by structured representations, which are talker- and group-specific beliefs about generative models. If the speaker is a male, for instance, the listener may adopt the belief that the generative model should map somewhat lower frication frequencies to the category /s/, since past

experience has shown that male speakers have lower frication frequencies than female speakers. In forming these structured representations, listeners assess the UTILITY of grouping variables. For example, vowel quality often provides a useful way to group speakers according to their gender or dialect and is more likely to be included in representations than are other acoustic features that are not strongly correlated with socioindexical characteristics (Kleinschmidt 2019). By proposing that listeners draw on a set of generative model parameters that crucially are compact, rather than fully elaborated, the ideal adapter charts a middle ground between abstractionist theories, in which no phonetic details are stored (e.g. McClelland & Elman 1986), and exemplar theories (e.g. Johnson 2006), in which every detail is stored. Nevertheless, the broad predictions are quite similar to those of exemplar theory: speech perception is social, and speech perception is adaptive.

Within this landscape, the twenty-first century has brought a remarkable communication innovation: people now regularly speak and listen to interlocutors that are not human. In the United States, for example, millions of people regularly interact with voice-activated artificially intelligent (voice-AI) devices, such as Apple's Siri, Amazon's Alexa, and Google Assistant (de Renesse 2017, Ammari et al. 2019). For the notion that 'speech perception is social', this development raises fundamental questions. After thousands of years of listening to exclusively human interlocutors, for example, we might have become accustomed to treating accents as distinctly human traits. Similarly, we might have considered quirky pronunciations of vowels or consonants to be characteristic of the particular human that produced them. If so, perhaps one reason that listeners adapt to speech so rapidly is because they understand that humans are inherently idiosyncratic. Does this belief extend to devices? Will listeners rapidly adjust to novel speech input they believe has been uttered not by a person, but by a machine? The current study pursues these questions by measuring perceptual learning of a novel shift while explicitly manipulating listener beliefs about whether the speaker is a person or a machine.

These questions are important for linguistic theory for several reasons. First, studying how people interact with voice-AI devices can inform theories of human-computer interaction and, more broadly, social language use. Do people interact with devices in the same way that they interact with humans? Some authors have argued that people treat computers as interlocutors that are less communicatively competent than humans (e.g. Burnham et al. 2010, Cowan et al. 2015, Cohn et al. 2022), but others have argued that people behave toward computers as if they are true social actors (Nass et al. 1994, Nass et al. 1999). By examining people's linguistic behavior toward talking devices, we can help provide a fuller, more accurate characterization of our increasingly technological world and its impact on human language.

Second, examining adaptation for humans and devices can shed light on the sociocognitive mechanisms that underpin perceptual learning. The phenomenon is robustly attested across a range of different experimental set-ups (e.g. Samuel & Kraljic 2009), but the underlying motivation for listener adaptation remains underspecified. In order to recalibrate, do listeners need to hold a certain set of beliefs about the speaker (for example, that their idiosyncratic way of speaking originates from a particular source)? One way to approach this question is to present listeners with different speaker guises: that is, visual or auditory information that indicates certain characteristics of the speaker's identity. To date, however, only a handful of studies in this area have employed different guises accompanying altered speech patterns—such as a visual cue of a speaker

with a pen in her mouth (Kraljic, Samuel, & Brennan 2008) or an auditory cue of rapid speech rate (Liu & Jaeger 2019)—and these have produced mixed results (described in more detail in §1.2). By asking whether people exhibit perceptual learning for speech presented in a 'device' guise, as we do in the current study, we can better understand the role of listener beliefs in speech perception.

Finally, exploring adaptation under different guise conditions can inform theories of speech representation. Following the broad hypothesis of the ideal adapter framework, listeners do not store every detail of every utterance (Kleinschmidt & Jaeger 2015, Kleinschmidt 2019). Instead, they store only those variables that create useful groupings of speakers. Previous work has explored how grouping variables might be selected in the auditory domain and has shown that, for example, American English listeners use vowel quality, but not voice onset time, to form structured social representations (Kleinschmidt 2019). In the current study, we extend some of these ideas to the visual domain and use visual images of human versus device speakers to influence listener beliefs about the speaker. For the purposes of building a structured representation, is information from one modality more useful than information from the other? By asking how auditory versus visual guises affect perceptual learning, we take another step toward the goal of building a compact model for the representation of speech.

In the current article, we present three experiments designed to investigate whether listeners undergo perceptual learning to the same extent when they believe that the speech signal has been generated by a voice-AI device, compared to a human talker. We use a traditional perceptual-learning paradigm, in which participants are trained with exposure words containing a phonetic shift (e.g. *bib* /bɪb/ produced as *beb* [bɛb]) (Maye et al. 2008), paired with guises indicating whether the speech originated from a device versus a human. Our key manipulation is the guise of device versus human, which we implement using visual cues in experiment 1 and auditory cues in experiments 2 and 3. In addition, we also manipulate the exposure task, asking participants to perform lexical identification in experiments 1 and 2, versus giving socioindexical ratings in experiment 3. In all three experiments, we test whether adaptation is restricted to the exposure voice and guise, or whether it also extends to novel voices and guises. Our ultimate goal is to determine whether, and under what circumstances, listeners adapt their perceptual categories for speech when they believe that the speech has originated from a device.

In sum, this article tests the boundaries of speech perception as a social phenomenon. To our knowledge, only a handful of studies have previously explored how listeners adapt to synthesized speech, and no prior work has addressed the contribution of visual versus acoustic guises in perceptual learning. As we outline in the following sections, the sociocognitive influences do not yet present a clear or definitive portrait of perceptual learning (§1.1) or generalization of the learning (§1.2), highlighting the ongoing need for additional research in order to more comprehensively understand how listeners adapt to different talkers' ways of speaking.

1.1. Perceptual learning. Two basic approaches have been used to induce perceptual learning: an audiovisual approach, in which listeners' interpretation of a target sound is influenced by videos of a speaker's mouth (McGurk & MacDonald 1976, Bertelson et al. 2003), and a lexical-retuning approach, in which interpretation of a sound is influenced by hearing it in the context of a word (for a review, see Samuel & Kraljic 2009). Here, we focus on lexical retuning. In the paradigm introduced by Norris et al. (2003), Dutch listeners were exposed to an ambiguous fricative, designated here as 'X', that was acoustically modified to be between an /s/ and an /f/.

One group heard /f/-final words with the last sound replaced with X (e.g. *witloX*, from *witlof* 'chicory'), in addition to nonmanipulated /s/-final word productions. The other group heard /s/-final words with the last sound replaced by X (e.g. *naaldboX*, from *naaldbos* 'pine forest') and nonmanipulated /f/-final productions. At test, listeners categorized individual sounds on an [f]–[s] continuum. Results showed that listeners in the /f/-final group shifted their interpretation of the ambiguous sound toward /f/, while listeners in the /s/-final group shifted their interpretation toward /s/. This demonstrates that lexical information can induce rapid adaptation of speech sound categories. Note that there is a parallel line of research that uses the term *perceptual adaptation* to refer to improved comprehension of globally accented speech, such as that produced by a nonnative talker (e.g. Clopper & Pisoni 2004, Bradlow & Bent 2008). In the current study, we focus on adaptation specifically as boundary-retuning of a phoneme category in the spirit of Norris et al. (2003).

Although many perceptual-learning studies have employed fricative consonants, some have used stop consonants, and a few have examined vowels, as we do in the current investigation. Maye et al. (2008) exposed listeners to a spoken passage where words containing the front vowel /ɪ/ were modified to contain a lowered vowel [ɛ] (e.g. *wicked witch* /wɪkəd wɪʧ/ shifted to /wɛkəd wɛʧ/). After exposure, listeners completed a word-identification task on items with lowered vowels. Results showed that they accepted stimuli such as *wetch* as real English words, indicating that they had learned the vowel shift. Two aspects of Maye et al.'s (2008) work are particularly relevant for the current study. First, it demonstrates that perceptual learning occurs not just with fricatives, but also with vowels. Note that fricatives and vowels both employ spectral cues that provide speaker-specific information, a point that is relevant when we consider whether perceptual learning generalizes from one speaker to another (Kraljic & Samuel 2006). Vowels, in particular, play an important role in differentiating regional dialects of American English (Clopper & Pisoni 2004); indeed, Maye et al. refer to their experimental manipulation as a 'dialect difference'. Second, Maye et al. (2008) actually used a synthesized voice for their stimuli, one of the original Apple text-to-speech (TTS) synthesized voices ('Bruce'), thereby demonstrating that perceptual learning is possible from speech generated by a device.

More recently, perceptual learning has been observed for both human and TTS talkers. Ferenc Segedin et al. (2019) exposed listeners to CVC words with cross-spliced nasal vowels (e.g. shifted *dead* [dɛ̃d], compared to unshifted [dɛd]). Some of the words were produced by human voices, and others by TTS voices. At test, participants heard CV syllables with nasal vowels ([dɛ̃]) and indicated whether the fragment was from a CVC or CVN word. A higher proportion of CVC responses was interpreted to indicate that adaptation had occurred. Results showed that perceptual learning for TTS voices did occur; furthermore, it occurred to a greater extent for TTS voices than for human voices, although the authors note that distinct phonetic characteristics of the voices could account for this difference.

1.2. Guises and perceptual learning. Within the literature on adaptation to globally accented speech, guises have served as a very effective experimental manipulation to investigate listener beliefs about the speaker. For example, Vaughn (2019) asked American English listeners to transcribe sentences produced by an L1 Spanish/L2 American English speaker. Information about the speaker was presented in written form, in one of three different conditions: a no-guise condition, an L1 accent condition (speaker is monolingual, but his parents speak Spanish and English), and an L2

accent condition (speaker's first language is Spanish, and he learned English at school). Results showed significantly better accuracy for L1 accent and L2 accent conditions, compared to the no-guise condition. This suggests that any information about the source of the speaker's accent leads to greater accuracy, compared to no information. Similar studies on globally accented speech and speaker guises (e.g. Rubin 1992, Yi et al. 2013, Babel & Russell 2015, McGowan 2015, Pycha et al. 2022) all point to a similar conclusion, namely, that guises affect listeners' overall interpretation of the speech signal.

For studies that have concentrated specifically on perceptual learning, which is our focus here, the results of guise manipulations are more mixed. In previous guise studies of perceptual learning, the essential question is whether the guise provides listeners with a reason to disregard the altered linguistic input. For example, upon hearing the word *epiXode*, where 'X' is ambiguous between [s] and [ʃ], the listener may believe that the speaker has a characteristic means of producing [s] and adjust their representations accordingly. But if the word is presented alongside video of a speaker with a pen in her mouth, the listener may plausibly attribute the altered pronunciation to the pen, and not the speaker herself, in which case perceptual learning does not occur (Kraljic, Samuel, & Brennan 2008).

The pen-in-the-mouth is one example of a visual guise that has been employed in classic perceptual-learning paradigms. Results are mixed. Although Kraljic, Samuel, and Brennan (2008) demonstrated that the pen-in-mouth guise inhibited learning, their follow-up work showed that pen-in-the-mouth videos could lead to recalibration of the [s]-[ʃ] contrast under certain circumstances (Kraljic & Samuel 2011). A replication by Liu and Jaeger (2018) found that perceptual learning was indeed suppressed in the presence of a video, but only for the [ʃ]-label condition and not the [s]-label condition; the authors speculate that participants considered the pen a sufficiently likely cause for alteration of [ʃ] pronunciations, but not for alteration of [s] pronunciations.

In addition to visual guises, studies have also employed auditory guises that manipulate the context or quality of the spoken stimuli. Kraljic, Brennan, and Samuel (2008) focused on American English dialects in which a context-specific shift is found: [stɹ] sequences change to [ʃtɹ], such that words like *street* [stɹit] are pronounced as *shtreet* [ʃtɹit] due to coarticulation. In the dialect condition, they embedded ambiguous fricatives exclusively in the [_tɹ] context, as in *Xtreet*. In the control condition, they embedded them in locations that were not context-specific, as in *halluXinate* (which would not be due to coarticulation). Results indicated that perceptual learning did not occur for the dialect group, suggesting that listeners did not attribute the altered pronunciations to the individual speaker, but instead compensated for the effect of phonetic context (cf. perceptual compensation for coarticulation; Zellou 2017). Thus, like the pen-in-the-mouth study, if there is an articulatory explanation for why such a shift might be occurring, listeners do not show adaptation. Note, however, that the mere presence of dialect or accent features need not inhibit perceptual learning on its own. Reinisch and Holt (2014) and Xie et al. (2017) embedded ambiguous segments into globally accented words, such as the word *seaX* (where 'X' is ambiguous between [t] and [d]) spoken in Mandarin-accented English. Results showed that participants did undergo perceptual learning, suggesting that the process is hindered only when dialects or accents employ a context-specific shift (e.g. one attributable to coarticulation or a pen in the mouth).

In addition to accents and dialects, researchers have also used other auditory-guise manipulations. Liu and Jaeger (2019) examined three incidental causes for atypical pronunciations, including intoxication, fast speech rate, and tongue-twisters. They found

evidence for perceptual learning in every condition. Thus, at least in some cases, listeners recalibrate phonetic categories even when altered pronunciations are presented as transient characteristics of the speakers. In sum, although there are relatively few studies investigating the role of guises in perceptual learning, these studies fit within a growing body of work examining how different types of cues influence the 'social priming' effect on speech perception (e.g. Johnson et al. 1999, Niedzielski 1999, Hay, Nolan, & Drager 2006, Hay, Warren, & Drager 2006, Drager 2010).

In order to examine how listeners perceptually adapt to speech that they believe was produced by a device, the current study uses both visual guises (experiment 1), in which participants see an image of either a human speaker or a device, and auditory guises (experiments 2 and 3), in which participants hear either unaltered stimuli or 'roboticized' stimuli with a flattened F0 and a slight echo. The latter manipulation follows an approach taken in work exploring the role of 'voice anthropomorphism' in speech perception (e.g. Cowan et al. 2015, Zellou, Cohn, & Block 2021).

1.3. Generalization of learning to novel talkers. If perceptual learning of a shift does occur, under what circumstances should it generalize to other talkers? On the one hand, if other talkers do not share the idiosyncratic pronunciation of the original talker, generalizing can actually be detrimental, because other talkers might not exhibit the shifted variants (Eisner & McQueen 2005, Kleinschmidt 2019). On the other hand, if other talkers do share this pronunciation—for example, if the pronunciation is a feature of a particular dialect or foreign accent—generalization could be useful, provided that the listener knows to apply their adjusted representations only when communicating with talkers of that group (Reinisch & Holt 2014, Kleinschmidt & Jaeger 2015, Kleinschmidt 2019).

Only a handful of previous studies have investigated these issues. Eisner and McQueen (2005) exposed listeners to Dutch words containing a sound ambiguous between [s] and [f], and later, at test, asked them to categorize sounds on an [ɛs]–[ɛf] continuum. When test fricatives from the original talker were spliced next to vowels produced by a novel talker, a shift in the continuum was apparent, indicating that generalization did occur. However, when the test continuum was made entirely from a novel talker's speech, there was no shift in the continuum, and generalization did not occur. Kraljic and Samuel (2006, 2007) replicated this finding for [s]-[ʃ] fricatives in English words; for these sounds, generalization to a novel talker did not occur. However, for [d]-[t] stops embedded in English words, their findings were different. In this case, generalization to a novel talker did occur. The authors speculate that acoustic differences may account for these different results: fricatives contain more speaker-specific cues than stops do.

Two studies of phonetic recalibration have also examined generalization in the context of talker groups, such as talkers with a foreign accent. Reinisch and Holt (2014) embedded a sound ambiguous between [s] and [f] into English words, which were produced by a Dutch native speaker and therefore 'globally accented'. At test, a shift in the [s]–[f] continuum was apparent, both for the original accented talker and also for a novel Dutch-accented talker whose fricatives were sampled across a similar perceptual space. However, for a Dutch-accented talker whose fricatives fell outside of this perceptual space, generalization did not occur. Reinisch and Holt (2014) speculate that, in theory, the presence of a foreign accent should facilitate cross-talker generalization of recalibrated sound categories. In practice, this idea could not be evaluated in their particular study, because the English-speaking participants did not hear the Dutch accents as the same.

In a similar vein, Xie and Myers (2017) exposed listeners to pronunciations of words like *seed*, spoken in Mandarin-accented English where final [d] is realized similarly to [t]. Across three experiments, there were three different conditions: (i) multiple exposure talkers with a novel test talker, (ii) a single exposure talker whose [d] was realized similarly to the novel test talker, and (iii) a single exposure talker whose [d] was realized differently from the novel test talker. Results showed that adaptation extended to a novel talker in conditions (i) and (ii), but not in condition (iii). The authors conclude that successful generalization does not require exposure to multiple talkers, but rather depends upon acoustic similarity between exposure and test talker. A post-hoc questionnaire revealed that participants were largely unaware of the fact that the speech was accented, suggesting that listeners did not generalize on the basis of top-down categories ('speakers with a Mandarin accent'), but rather on the basis of bottom-up information ('speech with similar acoustic patterns').

While less studied, there is some work examining how speakers generalize a novel shift based on both auditory and visual guises. For example, Lai (2021) found that both auditory guises (manipulating the speaker's voice to sound like a different gender) and visual guises (showing an image of a different person) inhibited generalization of perceptual learning to novel talkers, suggesting that if the cue provides clear enough details indicating that the novel talker is in a different social category from the exposure talker, generalization will not occur.

In the current study, we ask whether generalization to novel talkers is mediated by 'device' versus 'human' guises. Here, we might predict generalization of perceptual learning across talkers, provided that those talkers share the same social category; that is, a pattern learned in a device guise extends only to novel talkers presented in a device guise, and a pattern from a human voice generalizes only to other humans. This would be consistent with the ideal adapter framework (Kleinschmidt & Jaeger 2015, Kleinschmidt 2019). In line with prior work, we predict that auditory guises might show more auditory-similarity effects (e.g. Xie & Meyers 2017), while visual guises might show more distinct patterns (e.g. Lai 2021).

Importantly, however, we do not actually know if listeners will form a 'social category' for speech produced by devices. Devices are not humans and therefore, by some definitions, are fundamentally not social. If that is the case, perceptual learning with a device guise should not generalize to new voices. But of course, many of our interactions with voice-AI devices closely resemble the interactions we have with other humans (e.g. Yu et al. 2019) and, under this view, may be considered perfectly social (cf. Nass et al. 1994). In that case, perceptual learning with a device guise should indeed generalize to other voices in a device guise similarly to how we see learning from a human guise generalizing to novel human talkers.

1.4. Effect of task type. In most perceptual-learning studies, participants are exposed to altered speech patterns via a lexical decision task. For example, after hearing a stimulus such as *halluXinate*, participants must indicate if it is a real word of English. This is a metalinguistic task, in the sense that participants are asked to reflect upon the English lexicon. A handful of other studies have employed exposure tasks that did not require this level of lexical engagement. McQueen et al. (2006) asked participants to count exposure items, rather than make a lexical decision. Several other studies asked participants simply to listen to spoken passages or individual sentences, either with no follow-up task (Eisner & McQueen 2006, Maye et al. 2008, Babel et al. 2021) or with a sentence-level comprehension task (Zhang & Samuel 2014). In these studies, despite the lack of lexical engagement in the task, perceptual learning still occurred.

Drouin and Theodore (2018:1095) explicitly drew attention to the altered speech, informing listeners that 'this speaker's [sounds] are sometimes ambiguous, or sound funny, so listen carefully so as to choose the correct response'. Perceptual learning did occur in this condition; importantly, however, the magnitude of learning was no greater than that which occurred in the control condition, where the presence of altered speech was not highlighted. Thus, the basic phenomenon of perceptual learning appears to be relatively robust to differences in exposure tasks. Other types of speech-perception experiments, however, do show sensitivity to such task-based differences. For example, McGuire and Babel (2020) tested participants on how well they could identify old versus new voices. In the exposure phase, they listened to ten voices and completed either a lexical decision task or a dialect-rating task ('How likely is it that the speaker is from California?'). Results showed better accuracy for listeners who completed the dialect task, compared to those who completed the lexical task. The authors argued that the nature of attention given to a voice influences how speech patterns are encoded in long-term memory.

Does the greater attention that comes with an indexical task boost learning for human talkers only, or does it extend to device talkers as well? Work by Huyck and Johnsrude (2012) suggests that attention to indexical features of synthetic speech, in particular, helps listeners adapt. They found that successful comprehension of noise-vocoded speech occurred only when listeners were explicitly instructed to attend to speech (and not when they were instructed to attend to visual information or other auditory distractors). Thus, there is some evidence that the nature of exposure tasks—in particular, whether the task is oriented toward features of a speaker versus features of the language itself—affects how listeners adapt to speech. This is an issue we explore in the current study by manipulating task type across experiments 1 and 2 (word identification) versus experiment 3 (ratings for dialect and robot-likeness).

1.5. THEORIES OF HUMAN-COMPUTER INTERACTION. In almost all of the studies reviewed above, participants believed they were listening to a human talker. The novel question we pose in the current study is: what happens when listeners hear a phonetic shift produced by an apparent device? Two theoretical proposals from human-computer interaction can guide our predictions.

The COMPUTERS ARE SOCIAL ACTORS (CASA) theory proposes that people behave similarly to computers as they do toward humans, provided that the system displays a 'cue' of humanity (Nass et al. 1994). Indeed, there is work showing that people are prone toward anthropomorphizing machines (Bartneck et al. 2009, Waytz et al. 2010), and the transfer of human-based interaction schemas to machines has been demonstrated in a wide range of behaviors. For instance, in a seminal study, Nass et al. (1999) observed that participants are more likely to give higher performance ratings to a computer when that computer is in the room, compared to when the computer is in another room. This mirrors the 'politeness norms' that people apply to other humans, where people tend to be more positive when describing another person when that same person asks directly, compared to when they are asked about that person by a third party (e.g. Finkel et al. 1991). Of relevance to spoken interactions with voice-AI, CASA predicts that humans are particularly prone toward anthropomorphizing computers when the machines use language (Nass et al. 1994). Indeed, there is ample evidence that people apply human social behaviors during interactions with voice-AI, such as gender asymmetries (Ernst & Herm-Stapelberg 2020), responses to emotional expressiveness (Cohn & Zellou 2019, Cohn et al. 2021), politeness norms (e.g. saying 'Please' and 'Thank you' in Lopatovska & Williams 2018), and gendered personal pronouns (Purington et al. 2017). Additionally, people ask Alexa personal questions (e.g. 'What's your favorite

color?', 'Do you have any pets?') and engage in non-task-related chitchat (Yu et al. 2019). In sum, the CASA framework argues that people treat computer and human interlocutors in a similar manner. For the current study, then, a broad CASA prediction is that listeners should exhibit perceptual learning for device guises, similarly to what they exhibit for human guises.

In contrast to CASA, TECHNOLOGY-SPECIFIC ACCOUNTS propose that people have distinct expectations when interacting with technology, compared to when interacting with humans. For example, Gambino et al. (2020) argue that people develop technology-specific 'scripts', or modes of behavior, with devices. Part of what shapes these scripts is an expectation that the system has a reduced communicative competency. In several studies, top-down knowledge that the speaker is a 'device' versus 'human' has been shown to affect people's language patterns in ways consistent with this belief. For example, in a series of studies, Branigan and colleagues (2003, 2011) compared participants' linguistic alignment toward idiosyncratic syntactic and lexical patterns of an (apparent) computer and (apparent) human interlocutor. They found that people are more likely to adopt the linguistic patterns of the (apparent) machine, rather than those of the (apparent) human, as a way of increasing mutual comprehension with an interlocutor they deem to need more communicative support. Similar effects have been shown by cueing apparent roboticism via voice features. For example, some modern devices produce TTS voices with distinct speech characteristics, such as prosodic incongruencies, which can signal that the talker is a 'device' (Németh et al. 2007). Pursuing such issues, Cowan et al. (2015) found that more-robotic-sounding TTS voices were deemed to be less communicatively competent than more human-like TTS voices. For the current study, a technology-specific prediction is that listeners show differences in perceptual learning, depending on whether the guise is a device versus a human.

1.6. CURRENT STUDY. The overarching goal of our study is to investigate whether perceptual learning is modulated by the listener's belief that the speaker is a human versus a device. To pursue this goal, we use a traditional perceptual-learning paradigm, in which participants are trained with exposure words containing a phonetic shift (e.g. *bib* /bɪb/ produced as *beb* [bɛb]). During the subsequent testing phase, participants hear individual word stimuli that span a five-step continuum from *sit* [sɪt] to *set* [sɛt] and classify each item as either *sit* or *set*. As in any perceptual-learning experiment, the key question is whether participants' category boundaries shift as a result of exposure. Specifically here, we ask whether exposure makes listeners more likely to classify items on the continuum as *sit*.

Our crucial manipulation involves the presentation of a guise, which indicated whether the talker is a human or a device. The guise is either visual or auditory. In experiment 1, the guise is visual. Listeners were informed that the voice came from either a human or a device, and visual images of either a human or a device are presented alongside the exposure stimuli. In experiments 2 and 3, the guise is auditory. In the human condition, the exposure voice is unaltered, producing human-like speech. In the device condition, the exposure voice is altered to sound 'robotic', with a flattened F0 and a slight echo. In all three experiments, the key question is whether the presence of a human versus a device guise during exposure affects subsequent categorization of the *sit–set* continuum. Comparing across experiments 1 and 2 for the effect of visual versus auditory guises, we additionally test if speaker grouping effects are stronger based on the modality (visual or auditory). In all three experiments, listeners hear a single talker during exposure, and multiple talkers at test. Specifically, there are three different types

of test talkers: (i) those with the same voice and the same guise as exposure, (ii) those with a different voice but the same guise as exposure, and (iii) those with a different voice and a different guise. This design allows us to examine whether perceptual learning generalizes to novel talkers, and the extent to which guises affect this generalization.

Finally, we also investigate the role of task type. In experiments 1 and 2, the exposure phase consisted of a lexical identification task. Meanwhile in experiment 3, the exposure phase consisted of an indexical task in which participants rate socioindexical aspects of the voices (i.e. 'How likely is this speaker from California?' or 'How robotic does this voice sound?'). This manipulation allows us to test whether increased attention to indexical information boosts perceptual learning for human guises only, or boosts learning for device guises as well.

Note that in all three experiments and for both types of guises, the speech stimuli were generated using neural TTS synthesis. This is the most naturalistic and human-like speech technology available (van den Oord et al. 2016), and previous work has already demonstrated the feasibility of inducing rapid adaptation using TTS stimuli (Maye et al. 2008, Ferenc Segedin et al. 2019).

In examining the statistical results of each study, we evaluate whether learning has occurred in two ways: (i) if there is an increase in the proportion of *sit* responses from no exposure to an exposure condition overall, this indicates a shift occurred across the entire categorization curve (i.e. participants categorize more vowels as *sit*, including lowered and nonlowered vowels); and (ii) if there is a positive interaction between an exposure condition with Step, this means that the category boundary between [ɛ] and [ɪ] sharpened. In particular, if perceptual learning has occurred, then with each additional step toward [ɪ], *sit* responses should increase more steeply for participants who were exposed to the shift, compared to participants who were not.

Data and code for all three experiments are provided in an Open Science Framework (OSF) site.[1]

2. Experiment 1: visual guises and word-identification task. In experiment 1, listeners were exposed to altered speech (e.g. [bɛb] for *bib*), accompanied by either a visual image of a device or a visual image of a human. The exposure task was word identification: participants heard a sentence (e.g. *The baby got sauce all over her beb*) and identified the final word in a two-option forced-choice task. At test, participants categorized items from talkers' *set–sit* ([sɛt]–[sɪt]) continua. These items were presented in three different test speaker conditions: the exposure speaker's voice and guise, a new speaker's voice in the exposure guise, and a new speaker's voice in a different guise.

2.1. Methods.

Exposure-phase stimuli. Fifty monosyllabic lexical items containing /ɪ/ (e.g. *bib*) were selected as target words. The words all had high familiarity (mean familiarity rating of 6.9 out of 7 from the Hoosier mental lexicon; Nusbaum et al. 1984) and low age-of-acquisition ratings (mean rating of 7.8 years old from Kuperman et al. 2012). Fifty sentences were constructed such that the target word occurred in phrase-final position and in a highly predictable semantic context (e.g. *The baby got sauce all over her* ***bib***; full list provided in Appendix A). Sentences consisted of five to thirteen words in total and were designed to avoid nontarget words containing /ɛ/ and /ɪ/. The fifty sentences were generated in four female Amazon Web Services (AWS) Polly voices

[1] https://doi.org/10.17605/OSF.IO/2XQ9Z

(US-English: Salli, Joanna, Kendra, Kimberly; these were the only adult female US-English Amazon Polly voices available at the time of the study) using neural TTS synthesis. The target word in each sentence was modified orthographically in the plain text AWS Polly console to generate the /ɪ/-to-[ɛ] shift (e.g. for target word *twin*, <twen> yields [twɛn]). Sentences were downloaded from the AWS console, and amplitude was normalized to 60 dB. Average formant frequencies for [ɛ] vowels in target words, as well as the formant frequencies for original [ɪ] productions of the same words by these voices, are provided in Appendix B.

TEST-PHASE STIMULI. The words *sit* and *set* were generated in the same four Amazon Polly voices used to generate the exposure sentences, and amplitude was normalized to 60 dB. For each pair, the vowels were spliced out of their consonant contexts. The vowels were then gradually blended in equal proportions to generate five steps along a continuum using a Praat script (Winn 2014). The resulting vowels were spliced back into the original *sit* frame.

GUISE IMAGES. Eight images were created, consisting of either a female human silhouette or a digital device silhouette. The background of the images varied across four colors (red, blue, green, yellow) to align with different voices (examples shown in Figure 1 below). The four voices were randomly assigned one color identity, which remained consistent within and across lists (e.g. 'Salli' was always paired with a blue-background image, in both the human- and the device-guise conditions in both tasks).

PARTICIPANTS. One hundred and sixty-six students (mean age: 19.6 years old ± 2.41; two gender-fluid, 133 female, thirty-one male) from the UC Davis subject pool completed the study and received course credit for their participation. All participants identified themselves as native speakers of English with normal hearing.

PROCEDURE. The experiment was conducted online, using Qualtrics. Participants were instructed to complete the experiment in a quiet room with no background noise, where they would be undisturbed for the duration of the study. Participants were also instructed to wear headphones. The experiment began with a sound calibration procedure: participants heard one sentence amplitude normalized to 60 dB (*She asked about the host*), which they could play up to five times. They were instructed to identify the sentence from three multiple-choice options, each containing a phonologically close target word (*host*, *toast*, *coast*). After, they were instructed to not adjust their sound levels again during the experiment. (Note that during experimental trials, a stimulus item was played to participants only once, with no option for listeners to replay an item).

EXPOSURE + TEST CONDITION. Participants (n = 127) who were assigned to the EXPOSURE + TEST CONDITION then completed the EXPOSURE PHASE (schematized in Fig. 1.A). On a given trial, participants heard one of the fifty sentences with a final target word containing an /ɪ/-to-[ɛ] vowel shift. They were instructed to identify the last word of the sentence from one of two forced-choice options: one option was the correct target word, and the other option was another target word from the stimuli (pairings of target words were pseudo-randomized to avoid highly confusable options as much as possible, e.g. *chin* vs. *sick*). Participants identified the target word by clicking one of the two options that appeared on the screen. No feedback was given. During each trial, participants were given both visual and written cues to the voice guise (either human or device) with the presentation of the image and instructions to that effect (i.e. 'Listen to a digital assistant voice' or 'Listen to one of our research assistants'). Exposure trial order was randomized.

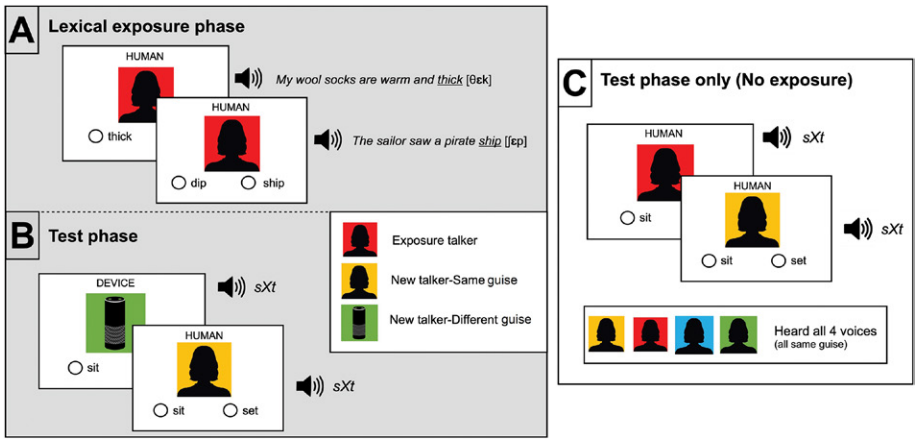

FIGURE 1. Perceptual-learning paradigm for experiment 1 based on visual guises. (A) Lexical exposure phase consisted of a single talker (human or device guise) producing a sentence with a final target word shifted from /ɪ/ to [ɛ]. Participants identified the target word from two options. (B) Test phase consisted of the talker from the exposure phase, as well as a new talker–same guise (e.g. human if exposure guise was human) and new talker–different guise (e.g. device if exposure guise was human). Most participants (n = 127) completed both the (A) lexical exposure phase, and then the (B) test phase. A subset of participants (n = 39) completed the (C) test phase only (no exposure) as a baseline, with all four talkers presented in a single guise.

Then, participants completed the TEST PHASE (schematized in Fig. 1.B), consisting of a categorization task. On a given trial, participants heard one of five items from a *set–sit* continuum and identified whether they heard 'sit' or 'set' from two alternative options. Participants categorized three voices: first, they completed a block with the exposure talker. Next, they categorized continua from two new talkers—one new talker with the same guise as the exposure talker (new talker–same guise) and the other new talker with a different guise from the exposure talker (new talker–different guise). The new talkers were presented in two separate blocks, and order of presentation of new talker–same guise and new talker–different guise was counterbalanced across participants. Participants heard the five continua items from each talker four times, for a total of sixty trials (3 voices * 5 steps * 4 repetitions; presented randomly within each block).

Assignment of the voices to exposure talker, new talker–same guise, and new talker–different guise was balanced equally across eight versions: 4 voices * 2 guises (correspondence of exposure talker to voice was pseudo-randomized across the versions).

TEST-ONLY CONDITION. Participants (n = 39) in the TEST-ONLY CONDITION (schematized in Fig. 1.C) did not complete an exposure phase. In the TEST-ONLY PHASE, listeners were presented with all four talkers in a single guise (i.e. either all device or all human guises) using the same images as in the exposure version of the experiment. Participants in all lists were presented with the five continua items from each of the four talkers four times (eighty total trials; presented in a single block, randomized). Participants were randomly assigned to either the device- or human-guise condition.

2.2. ANALYSIS AND RESULTS. The data from the 166 participants' responses to the test-phase items were binomially coded (1 = *sit*, 0 = *set*). We analyzed response data with a mixed-effects logistic regression model using the 'glmer()' function in the 'lmer' R package (Bates et al. 2015). The model included three fixed effects. The first effect was

TEST TALKER CONDITION (four levels: No exposure, Exposure talker, New talker–Same guise, New talker–Different guise; treatment-coded with No exposure as the reference level). The second effect was TEST GUISE (Device guise, Human guise; sum-coded). The third effect was STEP (five steps, standardized), testing whether probability of *sit* categorizations increased as the talkers' continua go from [ɛ] to [ɪ]. An interaction between Test talker condition and Test guise was also included. The random-effects structure of the model, which was maximal given the between-subjects design of the study, included random intercepts for Participant and Test talker and by-participant random slopes for Step.

We evaluated the addition of two- and three-way interactions with Step and the other predictors via model comparisons with the 'MuMIN' package in R (Bartoń 2015), based on the corrected Akaike information criterion (AICc), a measure of model fit, while penalizing overparameterization. The model that results in the lowest AICc is best supported by the data (Burnham et al. 2011). The model with the three-way interaction (Test talker condition * Test guise * Step) had the lowest AICc (ΔAICc = −19.39) and was thus retained.

Table 1 provides the summary statistics from the model, and Figure 2 shows the mean proportion of *sit* responses for each condition. There is an effect of Step, with greater *sit* responses as the continua shifted from [ɛ] to [ɪ]. There are also effects that reveal a shift in all or part of the curve, indicating learning. First, with respect to shifts across the entire continua, there are two effects of Test talker condition. As seen in the comparison between No exposure and Exposure talker in Fig. 2, listeners' overall *sit* responses increase after being exposed to a shifted talker. Thus, when listeners hear the same voice at exposure and test, they show evidence of perceptual learning: they accept as *sit* a greater number of vowels varying in their lowering than individuals without exposure do. Additionally, we see an increase in *sit* classifications in New talker–Same guise (relative to No exposure), indicating that listeners generalize this shift to other speakers who are presented in the same category as the exposure talker (e.g. device-to-device generalization to new talkers).

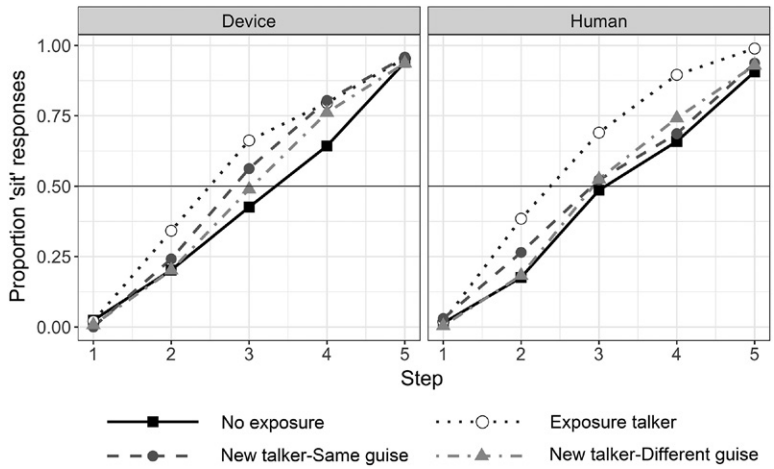

FIGURE 2. Experiment 1: visual guises. Proportion *sit* responses to [sɛt]–[sɪt] continua across test talker conditions as a function of Step for the device guise (left panel) and the human guise (right panel) following a lexical exposure task. No-exposure responses were provided by an independent group of participants who were not exposed to the vowel shift.

	Coef.	*SE*	*z*	*p*
(intercept)	−0.41	0.25	−1.66	0.10
Test talker condition (Exposure talker)	**1.18**	**0.21**	**5.62**	**< 0.001**
Test talker condition (New talker–Same guise)	**0.47**	**0.21**	**2.27**	**0.02**
Test talker condition (New talker–Different guise)	0.27	0.21	1.27	0.20
Test guise (Device)	−0.02	0.11	−0.19	0.85
Step	**2.14**	**0.10**	**21.77**	**< 0.001**
Test talker (Exposure talker) × Guise (Device)	−0.20	0.16	−1.26	0.21
Test talker (New talker–Same guise) × Guise (Device)	0.15	0.16	0.94	0.35
Test talker (New talker–Different guise) × Guise (Device)	0.00	0.16	0.00	1.00
Test talker (Exposure talker) × Step	**0.49**	**0.14**	**3.63**	**< 0.001**
Test talker (New talker–Same guise) × Step	**0.32**	**0.13**	**2.42**	**0.02**
Test talker (New talker–Different guise) × Step	**0.42**	**0.13**	**3.10**	**< 0.010**
Guise (Device) × Step	0.00	0.09	−0.04	0.97
Test talker (Exposure talker) × Step × Guise (Device)	−0.18	0.13	−1.42	0.16
Test talker (New talker–Same guise) × Step × Guise (Device)	**0.31**	**0.13**	**2.48**	**0.01**
Test talker (New talker–Different guise) × Step × Guise (Device)	−0.03	0.13	−0.22	0.83

N observations = 11,060, *N* participants = 166, *N* talkers = 4

Retained model syntax: Test talker * Test guise * Step + (1+Step|Participant) + (1|Test talker)

Model comparisons	AICc	*df*	ΔAICc
Model with no interaction with Step	8525.28	13	0
Model with two-way interactions with Step	8519.21	17	−6.07
Model with three-way interactions with Step	8505.89	20	−19.39

Table 1. Summary statistics for the glmer run on *sit* categorization responses in the test phase of experiment 1. Model comparisons are also provided, with the retained model underlined.

Step also interacted with Test talker condition and Test guise, indicating that with each additional step, the likelihood of a *sit* classification increases more sharply for all types of exposure conditions. This can be seen in Fig. 2, where there is a steeper increase in *sit* responses for all exposure conditions compared to No exposure.

Finally, a three-way interaction between Test talker, Step, and Guise reveals that the increase for Step for New talker–Same guise is even steeper for the device guise. Thus, device-to-device cross-talker generalization was stronger than human-to-human generalization.

2.3. Interim discussion. Experiment 1 tested perceptual learning for a vowel shift, with visual images indicating that the speakers were either ‘device’ or ‘human’. As expected, we observe a basic effect of a shift in *sit* responses, which confirms that perceptual learning can occur even when exposure stimuli are delivered by synthesized speech (Maye et al. 2008, Ferenc Segedin et al. 2019). More importantly, a key finding of experiment 1 is that perceptual learning of an exposure talker (i.e. talker-specific learning) occurs equivalently for both human and device guises—thus, when listeners believe that the speaker is a device, they recalibrate to the same extent as when they believe that the speaker is human.

There is also generalization to new talkers, but in ways mediated by the social categories of the talkers. Specifically, listeners show generalization following exposure to a ‘device’ guise, when generalizing to a novel ‘device’ talker (i.e. New talker–Same guise). This finding is consistent with theoretical accounts which predict that generalization of learning occurs most strongly for talkers within a social category (e.g. Kleinschmidt & Jaeger 2015, Kleinschmidt 2019), and it also suggests that listeners may apply the concept of a distinct social category robustly to voice-AI devices.

In addition to a broadening of acoustic values accepted as *sit*, the interaction with Step also shows evidence for sharpening category boundaries between [ɛ] and [ɪ] following any type of exposure to a shift: for the same talker as in the exposure phase, a new talker but same guise (e.g. human-to-human), and even a new talker in a different guise (e.g. human-to-device). This boundary is even sharper for the device guise in the new talker–same guise condition (i.e. device-to-device). This finding of a stronger device-to-device generalization parallels findings by Ferenc Segedin et al. (2019), who also observed device-to-device learning. Here, we extend this finding, showing that an apparent device triggers stronger generalization of learning to another apparent device, showing how strongly a visual guise can affect perceptual recalibration.

3. Experiment 2: auditory guises and word-identification task. The overall design of experiment 2 was identical to that of experiment 1, except that we used auditory guises instead of visual guises. Specifically, in order to create a robust 'device' voice guise, we resynthesized the original speech stimuli to sound 'robotic', flattening the pitch and adding a slight echo. The 'human' guise consisted of unaltered voices, synthesized from neural TTS, as in experiment 1. Also as in experiment 1, listeners were exposed to a vowel shift while completing a word-identification task, and then were tested on *set–sit* continua in the same voice and guise, as well as in different voices and guises. Thus, a primary goal of experiment 2 is to evaluate whether an auditory guise can modulate listener beliefs about the speaker and, in turn, affect perceptual learning.

3.1. Methods.

Stimuli. For the human-guise condition, we used the same stimuli as in experiment 1. For the device-guise condition, we resynthesized the stimuli from experiment 1 using two Praat VocalToolkit (Corretge 2012) functions. First, the F0 was monotonized to contain 0% F0 variation. Second, an echo was added (delay: 0.01 s; 0.5 Pa). Listeners associate flattened pitch and echo with 'robot' voices (Wilson & Moore 2017), and prior work has shown that this procedure for resynthesis yields speech that is rated as significantly more robotic-sounding than unmodified neural TTS (Zellou, Cohn, & Block 2021).

Participants. One hundred and fifty-seven students (mean age: 19.7 ± 2.6 years old; 121 female, thirty-six male) from the UC Davis subject pool, who identified themselves as native speakers of English with normal hearing, completed the study and received course credit for their participation. None of the subjects had participated in experiment 1.

Procedure. The experiment was conducted online, using Qualtrics. Participants (n = 108) in the exposure + test condition completed a paradigm identical to that of the exposure + test condition in experiment 1, except that the device-guise conditions were replaced with the corresponding 'roboticized' voices. Additionally, no explicit information (e.g. images, labels) about the speaker was provided in experiment 2. Participants (n = 49) in the test-only condition heard all four unmodified voices or roboticized voices, paralleling the between-subjects design used for test-only in experiment 1. No images or explicit instructions about the speaker were provided.

3.2. Analysis and results. The test responses were binomially coded (selected *sit* = 1, *set* = 0) and modeled in a mixed-effects logistic regression. The fixed- and random-effects structure was identical to that of experiment 1, except that Test guise consisted of the voice conditions: device guise (roboticized) or human guise (unmodified) (two levels: Device, Human; sum-coded). (glmer syntax: Test talker condition*Test guise + Step + (1 + Step|Participant) + (1|Test talker).) As with experiment 1, the addition of Step interactions was evaluated via model comparisons, testing both two- and three-way

interactions. Model comparisons showed that the model including three-way interactions with Step had the lowest AICc.

Figure 3 displays the mean proportion of *sit* responses for each condition. Table 2 provides the output of the model. An effect of Step indicates that *sit* responses increase as the continuum shifts from [ɛ] to [ɪ]. There was one effect of Test talker condition. As seen in the comparison between No exposure and Exposure talker in Fig. 3, listeners' overall *sit* responses increase after exposure to a shifted talker.

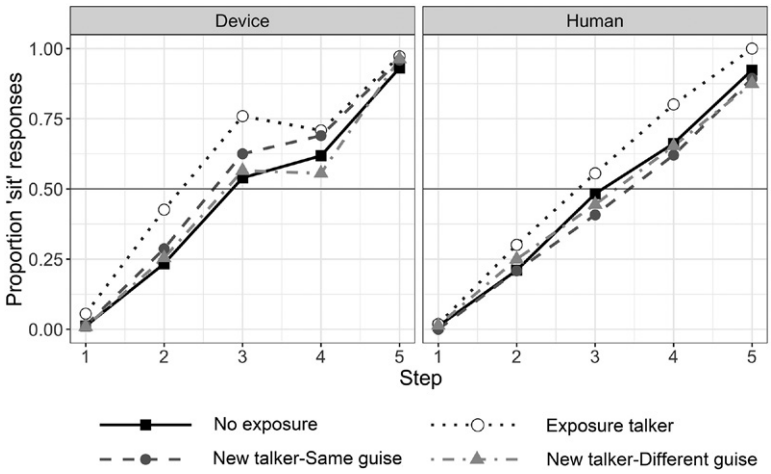

Figure 3. Experiment 2: auditory guises. Proportion *sit* responses to [sɛt]–[sɪt] continua across test talker conditions as a function of Step for the device guise (roboticized voice; left panel) and the human guise (unmodified voice; right panel) following a lexical exposure task. No-exposure data comes from participants who were not exposed to any vowel shift.

	Coef.	*SE*	*z*	*p*
(intercept)	−0.27	0.22	−1.25	0.21
Test talker condition (Exposure talker)	**0.90**	**0.20**	**4.55**	**< 0.001**
Test talker condition (New talker–Same guise)	0.06	0.20	0.33	0.74
Test talker condition (New talker–Different guise)	0.04	0.20	0.21	0.83
Test guise (Device)	0.03	0.16	0.22	0.83
Step	**2.04**	**0.09**	**22.25**	**< 0.001**
Test talker (Exposure talker) × Guise (device)	0.22	0.20	1.10	0.27
Test talker (New talker–Same guise) × Guise (Device)	0.36	0.20	1.84	0.07
Test talker (New talker–Different guise) × Guise (Device)	0.05	0.20	0.23	0.81
Test talker (Exposure talker) × Step	**0.37**	**0.13**	**2.77**	**< 0.01**
Test talker (New talker–Same guise) × Step	**0.26**	**0.13**	**1.99**	**< 0.05**
Test talker (New talker–Different guise) × Step	0.06	0.13	0.45	0.65
Guise (Device) × Step	−0.10	0.09	−1.07	0.29
Test talker (Exposure talker) × Step × Guise (Device)	−0.08	0.13	−0.57	0.57
Test talker (New talker–Same guise) × Step × Guise (Device)	**0.28**	**0.13**	**2.11**	**0.03**
Test talker (New talker–Different guise) × Step × Guise (Device)	0.12	0.13	0.95	0.34

N observations = 10,400, *N* participants = 157, *N* talkers = 4

Retained model syntax: Test talker * Test guise * Step + (1+Step|Participant) + (1|Test talker)

Model comparisons	AICc	*df*	ΔAICc
Model with no interaction with Step	8631.54	13	0
Model with two-way interactions with Step	8627.93	17	−3.61
Model with three-way interactions with Step	8623.44	20	−8.10

Table 2. Summary statistics for the glmer run on *sit* categorization responses in the test phase of experiment 2. Model comparisons are also provided, with the retained model underlined.

Additionally, we observe two interactions between Test talker and Step: an even steeper increase in *sit* responses occurs for Exposure talker as well as for New talker–Same guise. Finally, we observe one three-way interaction between Test talker, Step, and Guise, demonstrating that there is an even steeper increase in *sit* responses for New talker–Same guise in the device (roboticized) guise. (Note that while there appears to be a numerical dip in response values for the device guise for step 4 in Fig. 3, the model did not compute a main effect of Guise, or an interaction between Guise and Step, indicating that this is not a reliable pattern.)

3.3. INTERIM DISCUSSION. Experiment 2 tested perceptual learning with acoustic modifications indicating a speaker guise that was 'device' (i.e. modified to sound robotic) compared to a speaker guise that was more 'human' (unmodified). As expected, results show an overall effect of perceptual learning, and, just as in experiment 1, the degree of talker-specific learning is equivalent for human and device guises. This indicates that listeners quickly adapt to changes in the speech signal, even when they believe that the signal was produced by a nonhuman interlocutor, signaled by either visual or auditory cues.

Moreover, paralleling experiment 1, we observe that generalization of the shift to new talkers with the same guise is STRONGER for the device guise. Thus, both visual and auditory guises create conditions that favor generalization from one device to another, compared to generalization from one human to another.

One difference across the two studies is that in experiment 1, we observe some generalization to all new talkers (whether in the same guise or a new guise), while in experiment 2 we only observe generalization to new talkers in the same guise. Does an auditory guise provide a more robust activation of social categories? We explore this question in experiment 3 by testing learning in a different type of task where we manipulate explicit attention to different types of social features of the voices.

4. EXPERIMENT 3: AUDITORY GUISES AND SOCIAL-RATING TASK. Across both experiments 1 and 2, we find equivalent talker-specific learning for 'device' and 'human' exposure talkers. In both experiments, participants were exposed to the shift while completing a lexical identification task. As reviewed in §1.3, however, previous work provides reason to believe that the type of exposure task can shape learning. Therefore, the overall design of experiment 3 was identical to that of experiment 2, except that during exposure, participants completed a social-rating task rather than a word-identification task. Thus, after hearing each stimulus sentence, such as *The baby got sauce all over her* ***beb***, participants gave a rating indicating either how 'robotic' the sentence sounded (experiment 3a) or how 'Californian' it sounded (experiment 3b). As in experiment 2, sentences were presented with either an auditory 'device' guise (robotized stimuli) or an auditory 'human' guise (unmodified stimuli), and the test phase consisted of categorizations of *set–sit* continua. The basic goal of experiment 3 was to evaluate whether the presence of an auditory guise affects learning when participants complete a social-rating task.

4.1. METHODS.

STIMULI. The same stimuli from experiment 2 were also used in experiment 3.

INDEXICAL EXPOSURE TASK. Two different exposure conditions were used in experiment 3. In experiment 3a, after hearing each stimulus sentence, listeners were asked to rate the naturalness of the voice using a sliding scale, where 0 = 'very robotic',

100 = 'very humanlike'. In experiment 3b, listeners were asked to rate the likelihood that the talker was from California using a sliding scale 0–100, where 0 = 'very unlikely to be from California', 100 = 'definitely from California'.

Lists. The list structures for experiments 3a and 3b were identical to that from experiment 2.

Participants. A total of 141 participants (mean age = 19.4 ± 2.3 years; three nonbinary, one demi female, one genderqueer, 115 female, twenty-one male) completed experiment 3a (exposure + test condition), while 161 participants (mean age = 19.9 ± 2.0 years; one nonbinary, 129 female, one trans male, thirty male) completed experiment 3b (exposure + test condition). An additional fifty-six participants completed the test-phase-only condition (mean age = 19.9 ± 3.7 years; one nonbinary, forty-three female, twelve male).

4.2. Analysis and results.

Exposure-phase ratings. Separate mixed-effects linear regression models were fit to the exposure-phase ratings, one modeling human-likeness and one modeling Californian likelihood. In both, models included a fixed effect of Guise (Device, Human; sum-coded) and by-participant, by-TTS voice, and by-word random intercepts.

For experiment 3a, results showed that the humanlike ratings for the roboticized voice were lower on average (mean = 17.8, *SD* = 17.4) (*coef.* = −17.0, $t = -6.19$, $p < 0.001$); as the factor was sum-coded, the converse was true for the unmodified voice, with more humanlike ratings (mean = 39.9, *SD* = 24.5). Thus, the roboticization manipulation resulted in greater 'robotic'-sounding voices; the unmodified stimuli voices were more humanlike.

For experiment 3b, results revealed no difference in California ratings for the unmodified (mean = 43.5, *SD* = 26.4) and the roboticized (mean = 47.2, *SD* = 26.5) voices (*coef.* = 1.89, $t = 1.67$, $p = 0.10$).

Test-phase *sit* categorizations. Categorizations from experiments 3a and 3b were coded as binomial data (selected *sit* = 1, selected *set* = 0) and were analyzed with a mixed-effects logistic regression. Fixed effects included test talker condition (four levels, reference = No exposure; treatment-coded), test guise (Device, Human; sum-coded), exposure rating type (two levels: California indexicality, Bot indexicality; sum-coded), and all possible interactions. A fixed effect of step was also included (standardized). Inclusion of Step interactions was evaluated with model comparisons (testing two-, three-, and four-way interactions of Step with the other predictors). Model comparisons showed that the model with two-way Step interactions had the lowest AICc and was thus retained (the three- and four-way interaction models with Step increased AICc, as seen in Table 3).

The summary statistics of the model are provided in Table 3. Figure 4 displays the mean proportion of *sit* responses for participants who completed the robot indexicality exposure (experiment 3a; top panel) and the regional indexicality exposure ('How Californian?') (experiment 3b; bottom panel).

The model run on *sit* responses for experiment 3a–b reveals an effect of Step: more *sit* responses as the continua go from [ɛ] to [ɪ]. The model also shows an effect of Test talker condition: there is an increase in *sit* responses in the Exposure talker condition, relative to No exposure, providing evidence of talker-specific perceptual learning. Additionally, we observe new talker generalization, in accepting a wider

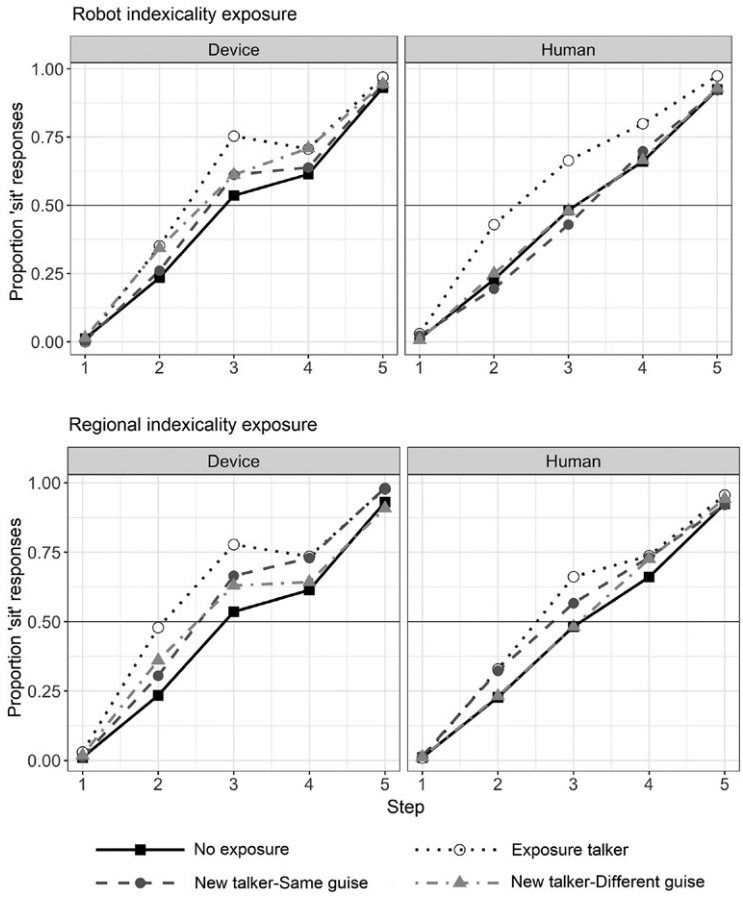

FIGURE 4. Experiment 3: auditory guises following indexical task. Proportion *sit* responses to [sɛt]–[sɪt] continua across test talker conditions as a function of Step for the device guise (roboticized voice; left panels) and the human guise (unmodified voice; right panels) following an exposure task where listeners rated the exposure voice on either 'robotic-sounding' (top panel) or 'How much from California' (bottom panel).

range of vowel lowering as *sit* for both New talker–Same guise and New talker–Different guise.

Test talker also interacts with Exposure rating type. First, when the exposure rating was to evaluate 'how Californian' the voice sounded in exposure (experiment 3b), we see more generalization to a new talker with the same guise (e.g. greater device-to-device and human-to-human generalization), illustrated in Fig. 4. Finally, there is a three-way interaction (Test talker condition * Test guise * Exposure rating type). As seen in the bottom panel of Fig. 4, talker-specific learning is stronger for the device guise in the California indexicality condition.

Finally, we observe interactions between Step and Test talker condition. Both Exposure talker and New talker–Same guise show even steeper *sit* responses with each increasing step, indicating sharper category boundaries. No other effects or interactions were observed.

	Coef.	*SE*	*z*	*p*
(intercept)	−0.27	0.20	−1.33	0.18
Test talker condition (Exposure talker)	**0.92**	**0.14**	**6.51**	**< 0.001**
Test talker condition (New talker–Same guise)	**0.31**	**0.14**	**2.19**	**0.03**
Test talker condition (New talker–Different guise)	**0.31**	**0.14**	**2.20**	**0.03**
Test guise (Device)	0.04	0.12	0.31	0.76
Exposure rating condition (How Californian, CA)	0.00	0.03	0.05	0.96
Step	**2.00**	**0.08**	**26.12**	**< 0.001**
Test talker (Exposure talker) × Guise (Device)	0.01	0.13	0.10	0.92
Test talker (New talker–Same guise) × Guise (Device)	0.08	0.13	0.63	0.53
Test talker (New talker–Different guise) × Guise (Device)	0.18	0.13	1.35	0.18
Test talker (Exposure talker) × Rating (CA)	0.02	0.07	0.23	0.82
Test talker (New talker–Same guise) × Rating (CA)	**0.23**	**0.07**	**3.25**	**< 0.01**
Test talker (New talker–Different guise) × Rating (CA)	0.02	0.07	0.26	0.80
Guise (Device) × Rating (CA)	0.00	0.03	0.00	1.00
Test talker condition (Exposure talker) × Step	**0.22**	**0.10**	**2.36**	**0.02**
Test talker condition (New talker–Same guise) × Step	**0.22**	**0.09**	**2.36**	**0.02**
Test talker condition (New talker–Different guise) × Step	0.10	0.09	1.11	0.27
Guise (Device) × Step	−0.04	0.03	−1.74	0.08
Rating (CA) × Step	−0.02	0.03	−0.56	0.57
Test talker (Exposure talker) × Guise (Device) × Rating (CA)	**0.14**	**0.07**	**2.02**	**0.04**
Test talker (New talker–Same guise) × Guise (Device) × Rating (CA)	−0.05	0.07	−0.69	0.49
Test talker (New talker–Different guise) × Guise (Device) × Rating (CA)	−0.01	0.07	−0.16	0.87

N observations = 27,080, *N* participants = 358, *N* talkers = 4

Retained model syntax: Test talker * Test guise + Step + (1+Step|Participant) + (1|Test talker)

Model comparisons	AICc	*df*	ΔAICc
Model with no interaction with Step	22738.19	21	0
Model with two-way interactions with Step	22736.00	26	−2.19
Model with three-way interactions with Step	22744.04	33	+5.85
Model with four-way interactions with Step	22743.83	36	+5.64

Table 3. Summary statistics for the glmer run on *sit* categorization responses after exposure via indexical rating tasks (experiment 3). Model comparisons are also provided, with the retained model underlined.

4.3. Interim discussion. Experiment 3 shows that perceptual learning is robust when the exposure task is a social rating, consistent with other studies showing learning across different types of exposure tasks (e.g. counting items in McQueen et al. 2006, listening to a passage in Maye et al. 2008).

Yet, unlike some prior work showing no difference in perceptual learning whether listeners are told the productions were 'altered' or not (Drouin & Theodore 2018), we see that learning in the present study is indeed mediated by the type of social attention in the exposure phase. In particular, when participants are attending to 'how Californian' the voice sounds, they demonstrate more talker-specific learning for a device guise than for a human guise, in contrast to patterns in experiments 1 and 2, where we find equivalent talker-specific learning for both guises. Thus, directing listeners' attention to an indexical California-rating task leads to stronger learning for the device guise.

Experiment 3 also reveals cross-category generalization (i.e. from device to human and vice versa), which is equivalent across the two types of indexical exposure tasks. This is similar to what we observe in experiment 1, where participants generalized patterns to novel talkers, both of the same guise and of different guises.

At the same time, we do see greater generalization to a new talker with the same guise when the exposure task consists of regional indexicality ratings (i.e. more device-to-device

or human-to-human generalization), compared to when it consists of 'robot-like' ratings. This finding suggests that directing attention to features associated more with humans (i.e. regional background) than machines (i.e. roboticism) shapes the way participants extend the pattern to other speakers sharing the same human social characteristic.

Unlike in experiments 1 and 2, experiment 3 does not demonstrate greater new-talker generalization in the device-to-device condition compared to the human-to human condition. Thus, directing listeners' attention to the social aspects of the signal leads to equivalent cross-talker generalization for device and human exposure talkers.

5. Post-hoc comparison of learning across lexical and indexical exposure tasks. In order to directly test whether perceptual learning differed as a function of type of exposure task (lexical vs. indexical), we ran a post-hoc mixed-effects logistic regression with the combined data from experiment 2 (lexical task) and experiment 3 (indexical task). The model structure was the same as in experiment 3, but instead of ratings type, we compared the effect of exposure task (two levels: Indexical, Lexical; sum-coded). Model comparisons revealed that including all two-way interactions with Step improved model fit (retained model glmer syntax: Test talker * Test guise * Exposure task + Step * (Test talker + Test guise + Exposure task) + (1+Step|Participant) + (1|Test talker)).

Across the two experiments, we observe a consistent effect of Test talker, revealing that there is learning for the talker from exposure (*coef.* = 0.93, $z = 6.11$, $p < 0.001$). There was also an effect of Step, wherein participants gave more *sit* responses as the continua moved from [ɛ] to [ɪ] (*coef.* = 2.03, $z = 25.88$, $p < 0.001$).

Furthermore, we observe interactions between Test talker and Exposure task. In the indexical tasks, we observe more *sit* responses overall for the new talkers in the same guise as exposure (*coef.* = 0.14, $z = 2.03$, $p < 0.05$) as well as new talkers in the different guise (*coef.* = 0.14, $z = 1.98$, $p < 0.05$), indicating generalization of learning.

We also observe two interactions between Test talker and Step: even steeper *sit* responses for the exposure talker (*coef.* = 0.26, $z = 2.80$, $p < 0.01$) and for new talkers in the same guise (*coef.* = 0.23, $z = 2.55$, $p < 0.05$).

Finally, we find one three-way interaction between Test talker, Guise, and Exposure task: in the device guise, there is a weaker shift for new talkers in the same guise in indexical tasks (*coef.* = −0.15, $z = -2.14$, $p < 0.05$). No other effects or interactions were observed.

6. Post-hoc comparison of learning across visual vs. auditory guises. We ran a post-hoc analysis to investigate learning and generalization based on the type of guise: a visual guise (images in experiment 1) or auditory guise (roboticized or unmodified TTS in experiment 2). We fit a mixed-effects logistic regression to the combined experiment 1 and 2 data, with fixed effects of test talker (four levels, reference = No exposure; treatment-coded), test guise (two levels: Device, Human; sum-coded), and guise type (two levels: Visual, Auditory; sum-coded), as well as their interaction. We also included the predictor of step (standardized). We tested including the interaction of Step with the other predictors (in four-, three- and two-way interactions) in separate models. The random-effects structure included random intercepts for Participant and Test talker, and by-participant random slopes for Step. Model comparisons confirmed that the model consisting of the three-way interactions with Step had the lowest AICc (ΔAICc = −37.86), and it was thus retained (retained model syntax: (Test talker+Test guise+Guise type+Step)^3 + (1 + Step|Participant) + (1|Test talker)).

The model revealed no effect of Guise type, and no interactions with Guise type. However, the general patterns from experiments 1 and 2 were confirmed. First, there was an effect of Test talker, indicating an overall effect of talker-specific learning (compared to No exposure) (*coef.* = 1.00, $z = 7.07$, $p < 0.001$). There is also generalization for New talker–Same Guise (*coef.* = 0.29, $z = 2.06$, $p < 0.05$). An interaction between Test talker and Test guise indicated even stronger generalization to a New talker–Same guise in the device guise (*coef.* = 0.24, $z = 1.97$, $p < 0.05$).

As with the other models, Step was a predictor: there were more *sit* responses with each increasing step (*coef.* = 2.06, $z = 31.19$, $p < 0.001$). Additionally, we observe interactions of Step with Test talker, indicating changes in the categorization function: listeners show steeper categorization functions for the exposure talker (*coef.* = 0.44, $z = 4.65$, $p < 0.001$), for a new talker with the same guise (*coef.* = 0.26, $z = 2.90$, $p < 0.01$), and for a new talker with a different guise (*coef.* = 0.25, $z = 2.68$, $p < 0.01$). There was one three-way interaction between Test talker, Test guise, and Step, with even steeper *sit* responses for New talker–Same guise for the device guise (*coef.* = 0.26, $z = 2.94$, $p < 0.01$). No other effects or interactions were observed.

7. GENERAL DISCUSSION. We began this study with the premise that speech perception is social and adaptive. Our first major finding is that this premise extends to nonhuman interlocutors. Across three different experiments, we have shown that perceptual learning occurs even when the listener has reason to believe that the talker is a device. In almost every case, talker-specific perceptual learning from device guises is just as strong as from human guises, and this finding holds regardless of whether the guise is visual or auditory. Perhaps more surprisingly, the second major finding of the current study is that the social, adaptive nature of speech perception is sometimes even MORE apparent in listeners' behavior toward devices, rather than toward other humans. In two out of three experiments, we find that generalization to new speakers is actually stronger when the listener believes the speaker is a device, compared to a human, suggesting that 'device' is a more salient social category for learning. In the sections that follow, we discuss these findings in more detail and interpret their implications for speech-perception models as well as theories of human-computer interaction.

7.1. PERCEPTUAL LEARNING FOR BOTH HUMAN AND DEVICE GUISES. In earlier sections, we speculated that human idiosyncrasy may lie at the heart of perceptual learning—the idea being that our perceptual system stands prepared to rapidly adjust because we know that humans produce tremendous variation. The findings of the current study explore this notion. When exposed to a shifted pronunciation, such as *beb* for the word *bib*, participants subsequently adapt their categorization functions: they accept more variants of lowered vowels as *bib* and/or shift their *set-sit* category boundary. They do so when they believe the speaker is a human, which was entirely expected. However, they also do so when they believe the speaker is a device, and this is a key finding. Thus, it may be the case that the notion of HUMAN idiosyncrasy is actually irrelevant for perceptual learning. Alternatively, perhaps human idiosyncrasy is indeed relevant, and listeners are simply willing to extend this notion to nonhumans, such as voice-AI devices. This possibility would be in line with the computers as social actors (CASA) framework (Nass et al. 1994, Nass et al. 1999), which argues that people interact with both computers and humans in a similar manner.

Importantly, perceptual learning occurs regardless of guise modality. Indeed, we see evidence of learning for both humans and devices when the guises are presented as a

visual image (experiment 1) and also when the guises are presented as an auditory modification (experiments 2 and 3), and post-hoc testing revealed no difference between the two. Of course, one possible interpretation is that the guises simply had no effect, but this does not square with the fact that human and device guises lead to differences in generalizations to new talkers. A more plausible interpretation is that guises exert equivalent learning effects regardless of whether they originate from top-down (visual image) or bottom-up (auditory modification) sources. Such a conclusion is relevant for the ideal adapter framework (Kleinschmidt 2019), which, to date, has focused exclusively on speaker models that are built using auditory information gleaned from the speech stream. Our results support prior work that finds a role for visual information in perceptual learning (cf. Kraljic, Samuel, & Brennan 2008), which is in the spirit of proposals that speech representations contain contextual details from many modalities (Pierrehumbert 2016).

Although the bulk of our results indicate that talker-specific learning for human and device guises is equivalent, there is one instance in which they differed. Specifically, in experiment 3b, where participants rated the likelihood that the speaker was from California, perceptual learning is stronger for the device guise than for the human guise. This finding is in line with technology-specific accounts (Branigan et al. 2011, Gambino et al. 2020, Zellou, Cohn, & Kline 2021), which claim that people have distinct expectations when interacting with technology, compared to when interacting with humans—but notably, our results reveal this difference only when listeners specifically attend to a regional dialect feature of the voice. This aspect of our findings would thus seem to contradict previous work showing that perceptual learning is relatively immune to task type (Eisner & McQueen 2006, McQueen et al. 2006, Maye et al. 2008, Zhang & Samuel 2014, Drouin & Theodore 2018, Babel et al. 2021), although it is line with research on other aspects of speech perception, which do exhibit sensitivity to task type (e.g. McGuire & Babel 2020).

Why should it be the case that talker-specific learning is stronger for the device guise, but only for a California-rating task? One possibility is that learning is stronger when there is a mismatch between social attention and voice features, as in experiment 3b, where listeners attended to the dialect features of a voice with roboticized characteristics, with a flat pitch and an echo. Indeed, previous work has found evidence for heightened attention to atypical scenarios (e.g. Näätänen et al. 1978), which is consistent with attentional accounts of speech learning (Huyck & Johnsrude 2012). An alternative explanation is that, rather than being particularly strong in the device guise, learning was actually relatively weak in the human guise. Recall that all of our speech stimuli were generated by neural TTS. While increasingly naturalistic, TTS voices are still rated as less humanlike than recordings of naturally produced voices (e.g. Cohn & Zellou 2021, Cohn et al. 2022). Thus, it is possible that the California-rating task draws participants' attention to the not-completely-human aspects of the TTS stimuli, leading to reduced learning for the human guise. Future work comparing recordings of naturally produced human voices, as well as other types of exposure tasks, could provide further insight into these possible factors. Regardless of the exact mechanisms, our findings from experiment 3 are broadly consistent with theories of speech representation in which episodic encoding is both socially weighted and modulated by attention (e.g. Sumner et al. 2014, Pierrehumbert 2016).

As a final note, the fact that our stimuli were generated by TTS confirms the basic feasibility of perceptual learning from TTS voices and extends prior work that had conflated guise with acoustic differences (Ferenc Segedin et al. 2019).

7.2. Generalization to novel talkers. In addition to testing for the basic phenomenon of perceptual learning, our experiments also investigate the extent to which participants generalized learning to new voices. In the talker-specific case, which we just reviewed above, listeners heard an exposure talker pronounce [bɛb] as an instance of the word *bib*, and, as a result, they shifted their categorization of that same talker's [ɛ]-to-[ɪ] continuum. In the generalization case, we ask whether listeners also shift their categorization of two new talkers' [ɛ]-to-[ɪ] continua—one of which presented with the same guise as exposure, and one of which presented with a different guise—even though they had never heard those particular talkers before. Our results show that this type of new-talker generalization does indeed occur, although in ways that differ from one experiment to the next.

In all three experiments, we find evidence for generalization to a new talker in the same guise. That is, when listeners are exposed to a talker with a human guise, they exhibit perceptual learning for a new talker that is also presented in a human guise. Similarly, when listeners are exposed to a talker with a device guise, they exhibit perceptual learning for a new talker that is also presented in a device guise. Unlike the perceptual learning for exposure talkers, however, generalization is not always equivalent for human versus device guises. In fact, in experiments 1 and 2, generalization to new speakers is stronger when the listener believes the speaker is a device, compared to a human.

To interpret this novel result, it may be useful to think about sources of variation in natural human speech. These can occur on an individual level, when speakers exhibit their own idiosyncrasies. They can also occur on a group level, when speakers exhibit patterns that are characteristic of their social groupings, such as those based on their region of origin or gender. Plausibly, people who listen to a single human will adopt a default assumption that any shifted pronunciations arise from individual-level variation and will not attribute them to group-level variation unless provided with explicit evidence that they should do so. Upon listening to a second human, then, the listener would show no evidence of perceptual learning, since the learning that occurred previously was speaker-specific. Crucially, the same logic need not apply to devices. Unlike humans, there are a limited number of TTS voices available in each language and a limited number of methods to synthesize them, so listeners have no particular reason to believe that devices exhibit individual-level variation. Plausibly, then, people who listen to a single device might assume that any shifted pronunciations actually arise from group-level variation (e.g. 'all devices sound like this'). In other words, listeners form a robust 'device' representation after a single exposure. This interpretation is consistent with an exemplar-theoretic approach (e.g. Pierrehumbert 2002, 2016): as people, conceivably, have fewer experiences with TTS productions than with human voices, a single exposure block by a 'device' voice can have a stronger impact on the entire category of 'device' productions than that for humans. In line with these explanations, we do see that participants readily generalize perceptual learning to a novel device talker in experiments 1 and 2.

Unlike the stronger same-guise generalization for devices, we see equal cross-category generalization in experiments 1 and 3. That is, when listeners are exposed to a shifted talker with a human guise, they extend the pattern they learned to a new talker in a device guise. Similarly, when listeners are exposed to a shifted talker with a device guise, they generalize for a new talker in a human guise. The finding of device-to-human generalization, which suggests that speech patterns learned from voice-AI can carry over to interactions with new human talkers, has particularly interesting

implications and suggests that our increasing interactions with voice-AI devices might have the potential to contribute to language change over time. This is a ripe area for future research.

In contrast to experiments 1 and 3, experiment 2 does not show evidence of different-guise generalization. This finding is consistent with previous studies of perceptual learning, which have reported generalization only in very narrow instances, specifically when the new talker's target sounds overlapped in their acoustic characteristics with those of the exposure talker (Eisner & McQueen 2005, Kraljic & Samuel 2006, 2007). Recall that experiment 2 uses auditory guises. This means that, in order to generalize to a new talker, listeners would need to take the pronunciation shift they learned in unmodified speech and apply it to modified, roboticized speech (or vice versa). Given that the acoustic characteristics of the new talker and the exposure talker have less overlap, we do not expect this to occur, and indeed experiment 2 shows that it does not. The open question concerns experiment 3, which also uses auditory guises but nevertheless shows evidence for different-guise generalization. Presumably, the indexical nature of the exposure tasks in experiment 3 somehow mitigates the acoustic differences between the different types of speech. Pinpointing how and why this occurs is an area for future research. The fact that we observe same-guise generalization across all three experiments provides robust support for theoretical proposals that incorporate social groupings into models of speech perception (e.g. the ideal adapter framework; Kleinschmidt & Jaeger 2015, Kleinschmidt 2019). At the same time, the fact that different-guise generalization occurs in some cases adds further nuance to such models and suggests the need for continuing research on talker types, listener beliefs, and exposure tasks.

7.3. Scope and future directions. For millions of people, nonhuman interlocutors have begun to exert an impact on communication during daily life. We seek to understand that impact: the current study is one attempt to understand how, when, and why interactions with devices affect language. We see several future avenues for pursuing this goal and building upon the present experiments. For example, the present study confirms that perceptual learning from TTS voices can occur, but also that generalization to new talkers is crucially modulated by listener beliefs. It will be important to gauge the extent to which these patterns of generalization are similar to, or different from, that which occurs with naturally produced human voices. Similarly, although we have demonstrated generalization effects based on a particular visual device guise (namely, an image of a round and cylindrical machine), many other visual images could be deployed that convey stronger anthropomorphic cues, such as a Nao robot or Furhat robotics bust. Indeed, prior work suggests that gradient differences in robot anthropomorphism affect listener behavior in a phonetic imitation task (Cohn et al. 2020), and we may expect such differences to affect perceptual learning as well.

Exploring the social characteristics of TTS voices will also be a crucial area for future research. The voices we employed in the current study were relatively homogenous: they were all female, with American English accent features. But recent work has shown that users are highly sensitive to indexical features of TTS voices, including the apparent gender (Cohn et al. 2019), age (Zellou, Cohn, & Ferenc Segedin 2021), degree of personification/humanlike embodiment (Cohn et al. 2020), idiosyncratic variation (Trude & Brown-Schmidt 2012), and race (Mengesha et al. 2021). These socioindexical

features could potentially modulate how people engage with technology. Furthermore, previous work has shown that individuals differ in the extent to which they anthropomorphize nonhuman entities, such as computers (Waytz et al. 2010), and it seems reasonable to expect the same to be true for TTS voices.

Other factors, beyond the scope of the current study, also are areas ripe for future work. For example, previous research has shown that trial-level effects, voice randomization, and block ordering can influence perceptual learning (e.g. Tamminga et al. 2020). Exploring how these factors affect learning and generalization with voice-AI devices is a direction for future studies.

On a larger scale, we may ask whether interactions with TTS voices exert any effect on the course of diachronic language change. It has already been proposed that perceptual learning, of the kind observed in the current study, can serve as a mechanism for sound change (Tamminga et al. 2020). Further work exploring whether, and when, adaptation can generalize from devices to humans will contribute to this question. Moreover, future work observing how human-device interactions change over time, across the lifespan, or across speech communities can speak to this issue.

Beyond perceptual learning of a particular speech sound, what are the broader implications of the current study for the domain of learning? From a practical point of view, it would be highly useful if voice assistants such as Siri, Alexa, and Google Assistant could be used as language training tools, and this is a growing area of research (e.g. Tai & Chen 2020, Li & Lan 2021). For example, children and adults could use the devices to acquire a new language. Or adults could use them to learn a new dialect—picture an American English user preparing for a trip to India by setting their device to an Indian English voice. Earlier work examining similar issues has not always been promising: for example, Kuhl et al. (2003) showed that infants learn sound categories only when interacting with a real human interlocutor, rather than a video of an interlocutor. The current study contributes a new perspective. We demonstrate a situation in which people can indeed learn new sound correspondences from a nonhuman interlocutor. Furthermore, their generalization of that learning to new contexts can be enhanced, or reduced, by their beliefs about WHO or WHAT the talker is. Exploring the optimal conditions for device-based speech and language learning will be a crucial avenue for future research.

8. CONCLUSION. Millions of people now interface with technology using spoken language—once considered to be an exclusively human domain. The broad adoption of voice-AI devices represents a fundamental change in this understanding and challenges us to reconceptualize our models of linguistic representation. In this study, we asked how listeners respond to speech stimuli when they are given reason to believe that the speaker is a device. Our results show that they rapidly adjust their category boundaries, just as they do when they believe the speaker to be a human, suggesting that social information about a device can structure linguistic representations in the same way that social information about humans does. Furthermore, listeners generalize these adjustments to new talkers, and do so particularly strongly when they believe that both old and new talkers are devices, suggesting that listeners may group nonhuman talkers according to metrics different from those used for human talkers. These findings strongly support models of speech perception that incorporate social information, and also indicate that we must continue to explore the very notion of what it means to be social.

Appendix A: Exposure stimuli

Full list of fifty sentences with the target word in phrase-final position in a highly predictable semantic context, with familiarity (FAMIL) and age-of-acquisition (AoA) ratings, number of words in each sentence (*N* WDS), and lexical task competitor option (COMP).

	RATINGS				
TARGET	FAMIL	AoA	SENTENCE	*N* WDS	COMP
bib	6.8	4.5	The baby got sauce all over her bib.	8	rib
bricks	7	6.4	The house was made out of bricks.	7	rinse
bridge	6.9	5.6	The car slowly crossed the bridge.	6	risk
chin	7	4.2	The baby stumbled and scraped her chin.	7	sick
chip	6.9	5.9	She had lots of salsa but only one chip.	9	sip
crib	6.8	6.1	The baby couldn't crawl out of the crib.	8	shin
dig	6.9	4.2	Paula had no shovel so she could not dig.	9	ship
dip	7	7	Raymond forgot to buy carrots and ranch dip.	8	skin
dish	7	4.9	Emmanuel heard a crash when he dropped the dish.	9	skip
drip	7	5.9	You could hear the water drip.	6	spit
fifth	7	5.4	Bob placed fourth but Steve placed fifth.	7	stick
fin	7	7.3	From the boat she saw a shark fin.	8	swim
fish	7	4.1	A large aquarium holds many fish.	6	switch
fist	6.5	4.6	To slap you use your palm, but to punch you use your fist.	13	tip
fit	7	5.7	The shoes were so small that her feet would not fit.	11	thick
fix	7	5	I brought you the cracked vase to fix.	8	trip
gift	7	5.1	Molly received a doll as a birthday gift.	8	twig
grin	6.8	5.8	She smiled with a cheeky grin.	6	twin
hint	7	5.9	He dropped a really subtle hint.	6	wig
hip	6.8	6.2	She tumbled and broke her hip.	6	wind
hit	7	4.8	Tom saw the ball before he got hit.	8	wish
kick	7	4.1	The boy gave the football a kick.	7	witch
kid	7	4.3	Tommy thought he was grown-up but he was just a kid.	11	rich
kiss	7	3.6	The mother gave the baby a kiss.	7	zip
pitch	6.9	6.4	The player threw out the last pitch.	7	bib
rib	7	6.3	On Saturdays, Joan always cooks prime rib.	7	bricks
rich	7	6.3	The man who owned the mansion was rich.	8	kiss
ridge	7	8.8	Gordon tried to climb over the ridge.	7	rinse
rinse	7	4.9	He had many plates to rinse.	6	bridge
risk	6.8	7.6	He wouldn't sky dive because of the risk.	8	spit
shin	6.9	8.4	At the soccer game, Kayla hurt her shin.	8	dig
ship	7	5.3	The sailor saw a pirate ship.	6	dip
sick	7	4.1	Amy came to the doctor because she was sick.	9	chip
sip	6.7	4.3	He didn't usually drink coffee so he just took one sip.	11	crib
skin	7	4.5	Laura put sunscreen on her skin.	6	dish
skip	7	4.7	Amanda had to choose an appointment to skip.	8	drip
spit	7	5.1	To defog goggles I usually just use spit.	8	risk
stick	6.9	3.9	The dog buried a stick.	5	fin
swim	7	4.2	She came out to the lake for a swim.	9	spit
switch	7	4.8	John had Jack's badge so they had to switch.	9	fist
thick	7	5.6	My wool socks are warm and thick.	7	fix
tip	6.9	5.4	The sailor could not see the whole iceberg, only the tip.	11	fit
trip	7	4.2	The class was going on a field trip.	8	gift
twig	7	6.3	The storm reduced the large tree into a twig.	9	grin
twin	6.5	6.1	James has a brother but he's not a twin.	9	hint
wig	7	5.6	Jonathan was bald but wore a wig.	7	hip
wind	7	3.9	George's kite was lost to the wind.	7	hint
wish	6.9	3.8	The genie announced he would grant one wish.	8	kick
witch	7	4.8	For the halloween party she came as a witch.	9	kid
zip	7	5.3	The woman worried when the backpack wouldn't zip.	8	pitch

Appendix B: Vowel formant frequencies

Average of vowel formant frequencies (Hz, taken at vowel midpoint) for target words in exposure stimuli that occurred with an /ɪ/-to-[ɛ] shift and original productions by the four TTS voices.

	F1	F2
original	614	1985
shifted	840	1844

REFERENCES

Ammari, Tawfiq; Jofish Kaye; Janice Y. Tsai; and Frank Bentley. 2019. Music, search, and IoT: How people (really) use voice assistants. *ACM Transactions in Computer Human Interaction* 26(3):17. DOI: 10.1145/3311956.

Babel, Molly; Khia A. Johnson; and Christina Sen. 2021. Asymmetries in perceptual adjustments to non-canonical pronunciations. *Laboratory Phonology* 12(1). DOI: 10.16995/labphon.6442.

Babel, Molly, and Jamie Russell. 2015. Expectations and speech intelligibility. *The Journal of the Acoustical Society of America* 137(5).2823–33. DOI: 10.1121/1.4919317.

Baese-Berk, Melissa. 2018. Perceptual learning for native and non-native speech. *Psychology of Learning and Motivation* 68.1–29. DOI: 10.1016/bs.plm.2018.08.001.

Bartneck, Christoph; Dana Kulić; Elizabeth Croft; and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1(1).71–81. DOI: 10.1007/s12369-008-0001-3.

Bartoń, Kamil. 2015. MuMIn: Multi-model inference. R package version 1.18. Online: https://cran.r-project.org/web/packages/MuMIn/index.html.

Bates, Douglas; Martin Maechler; Ben Bolker; Steven Walker; Rune Haubo Bojesen Christensen; Henrik Singmann; Bin Dai; Fabian Scheipl; and Gabor Grothendieck. 2015. lme4: Linear mixed-effects models using S4 classes. R package version 1.6. Online: https://cran.r-project.org/web/packages/lme4/index.html.

Bertelson, Paul; Jean Vroomen; and Béatrice De Gelder. 2003. Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science* 14(6). 592–97. DOI: 10.1046/j.0956-7976.2003.psci_1470.x.

Bradlow, Ann R., and Tessa Bent. 2008. Perceptual adaptation to non-native speech. *Cognition* 106(2).707–29. DOI: 10.1016/j.cognition.2007.04.005.

Branigan, Holly P.; Martin J. Pickering; Jamie Pearson; Janet F. McLean; and Clifford Nass. 2003. Syntactic alignment between computers and people: The role of belief about mental states. *Proceedings of the 25th annual meeting of the Cognitive Science Society (CogSci 2003)*, 186–91.

Branigan, Holly P.; Martin J. Pickering; Jamie Pearson; Janet F. McLean; and Ash Brown. 2011. The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition* 121(1).41–57. DOI: 10.1016/j.cognition.2011.05.011.

Burnham, Denis K.; Sebastian Joeffry; and Lauren Rice. 2010. Computer- and human-directed speech before and after correction. *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*, 13–17. Online: https://assta.org/proceedings/sst/SST-10/SST2010/PDF/AUTHOR/ST100077.PDF.

Burnham, Kenneth P.; David R. Anderson; and Kathryn P. Huyvaert. 2011. AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology and Sociobiology* 65(1).23–35. DOI: 10.1007/s00265-010-1029-6.

Clopper, Cynthia G., and David B. Pisoni. 2004. Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics* 32(1).111–40. DOI: 10.1016/S0095-4470(03)00009-3.

Cohn, Michelle; Bruno Ferenc Segedin; and Georgia Zellou. 2019. Imitating Siri: Socially-mediated vocal alignment to device and human voices. *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)*, Melbourne, 1813–17. Online: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2019/papers/ICPhS_1862.pdf.

Cohn, Michelle; Bruno Ferenc Segedin; and Georgia Zellou. 2022. Acoustic-phonetic properties of Siri- and human-directed speech. *Journal of Phonetics* 90: 101123. DOI: 10.1016/j.wocn.2021.101123.

Cohn, Michelle; Patrik Jonell; Taylor Kim; Jonas Beskow; and Georgia Zellou. 2020. Embodiment and gender interact in alignment to TTS voices. *Proceedings of the 42nd annual meeting of the Cognitive Science Society (CogSci 2020)*, 220–26. Online: https://cognitivesciencesociety.org/cogsci20/papers/0044/index.html.

Cohn, Michelle; Kristin Predeck; Melina Sarian; and Georgia Zellou. 2021. Prosodic alignment toward emotionally expressive speech: Comparing human and Alexa model talkers. *Speech Communication* 135.66–75. DOI: 10.1016/j.specom.2021.10.003.

Cohn, Michelle, and Georgia Zellou. 2019. Expressiveness influences human vocal alignment toward voice-AI. *Proceedings of Interspeech 2019*, 41–45. DOI: 10.21437/Interspeech.2019-1368.

Corretge, Ramon. 2012. Praat vocal toolkit. Online: http://praatvocaltoolkit.com.

Cowan, Benjamin R.; Holly P. Branigan; Mateo Obregón; Enas Bugis; and Russell Beale. 2015. Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human-computer dialogue. *International Journal of Human-Computer Studies* 83.27–42. DOI: 10.1016/j.ijhcs.2015.05.008.

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

Drager, Katie. 2010. Sociophonetic variation in speech perception. *Language and Linguistics Compass* 4(7).473–80. DOI: 10.1111/j.1749-818X.2010.00210.x.

Drager, Katie. 2011. Speaker age and vowel perception. *Language and Speech* 54(1). 99–121. DOI: 10.1177/0023830910388017.

Drouin, Julia R., and Rachel M. Theodore. 2018. Lexically guided perceptual learning is robust to task-based changes in listening strategy. *The Journal of the Acoustical Society of America* 144(2).1089–99. DOI: 10.1121/1.5047672.

Eisner, Frank, and James M. McQueen. 2005. The specificity of perceptual learning in speech processing. *Perception & Psychophysics* 67(2).224–38. DOI: 10.3758/BF03206487.

Eisner, Frank, and James M. McQueen. 2006. Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America* 119(4).1950–53. DOI: 10.1121/1.2178721.

Ernst, Claus-Peter H., and Nils Herm-Stapelberg. 2020. Gender stereotyping's influence on the perceived competence of Siri and Co. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 4448–53. Online: http://hdl.handle.net/10125/64286.

Ferenc Segedin, Bruno; Michelle Cohn; and Georgia Zellou. 2019. Perceptual adaptation to device and human voices: Learning and generalization of a phonetic shift across real and voice-AI talkers. *Proceedings of Interspeech 2019*, 2310–14. DOI: 10.21437/Interspeech.2019-1433.

Finkel, Steven E.; Thomas M. Guterbock; and Marian J. Borg. 1991. Race-of-interviewer effects in a preelection poll Virginia 1989. *Public Opinion Quarterly* 55(3).313–30. DOI: 10.1086/269264.

Gambino, Andrew; Jesse Fox; and Rabindra A. Ratan. 2020. Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication* 1(1).71–86. DOI: 10.30658/hmc.1.5.

Goldinger, Stephan D. 1996. Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 22(5).1166–83. DOI: 10.1037/0278-7393.22.5.1166.

Hay, Jennifer, and Katie Drager. 2010. Stuffed toys and speech perception. *Linguistics* 48(4).865–92. DOI: 10.1515/ling.2010.027.

Hay, Jennifer; Aaron Nolan; and Katie Drager. 2006. From *fush* to *feesh*: Exemplar priming in speech perception. *Linguistic Review* 23(3).351–79. DOI: 10.1515/TLR.2006.014.

HAY, JENNIFER; PAUL WARREN; and KATIE DRAGER. 2006. Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics* 34(4).458–84. DOI: 10.1016/j.wocn.2005.10.001.

HUYCK, JULIA JONES, and INGRID S. JOHNSRUDE. 2012. Rapid perceptual learning of noise-vocoded speech requires attention. *The Journal of the Acoustical Society of America* 131(3).EL236–EL242. DOI: 10.1121/1.3685511.

JOHNSON, KEITH. 1997. The auditory/perceptual basis for speech segmentation. *OSU Working Papers in Linguistics* 50.101–13. Online: http://hdl.handle.net/1811/81782.

JOHNSON, KEITH. 2006. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics* 34(4).485–99. DOI: 10.1016/j.wocn.2005.08.004.

JOHNSON, KEITH; ELIZABETH A. STRAND; and MARIAPAOLA D'IMPERIO. 1999. Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics* 27(4).359–84. DOI: 10.1006/jpho.1999.0100.

KLEINSCHMIDT, DAVE F. 2019. Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience* 34(1).43–68. DOI: 10.1080/23273798.2018.1500698.

KLEINSCHMIDT, DAVE F., and T. FLORIAN JAEGER. 2015. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review* 122(2).148–203. DOI: 10.1037/a0038695.

KRALJIC, TANYA; SUSAN E. BRENNAN; and ARTHUR G. SAMUEL. 2008. Accommodating variation: Dialects, idiolects, and speech processing. *Cognition* 107(1).54–81. DOI: 10.1016/j.cognition.2007.07.013.

KRALJIC, TANYA, and ARTHUR G. SAMUEL. 2006. Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review* 13(2).262–68. DOI: 10.3758/BF03193841.

KRALJIC, TANYA, and ARTHUR G. SAMUEL. 2007. Perceptual adjustments to multiple speakers. *Journal of Memory and Language* 56(1).1–15. DOI: 10.1016/j.jml.2006.07.010.

KRALJIC, TANYA, and ARTHUR G. SAMUEL. 2011. Perceptual learning evidence for contextually-specific representations. *Cognition* 121(3).459–65. DOI: 10.1016/j.cognition.2011.08.015.

KRALJIC, TANYA; ARTHUR G. SAMUEL; and SUSAN E. BRENNAN. 2008. First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science* 19(4).332–38. DOI: 10.1111/j.1467-9280.2008.02090.x.

KUHL, PATRICIA K.; FENG-MING TSAO; and HUEI-MEI LIU. 2003. Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences* 100(15).9096–9101. DOI: 10.1073/pnas.1532872100.

KUPERMAN, VICTOR; HANS STADTHAGEN-GONZALEZ; and MARC BRYSBAERT. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* 44(4).978–90. DOI: 10.3758/s13428-012-0210-4.

LADEFOGED, PETER, and DONALD E. BROADBENT. 1957. Information conveyed by vowels. *The Journal of the Acoustical Society of America* 29(1).98–104. DOI: 10.1121/1.1908694.

LAI, WEI. 2021. *The online adjustment of speaker-specific phonetic beliefs in multi-speaker speech perception*. Philadelphia: University of Pennsylvania dissertation.

LI, PING, and YU-JU LAN. 2021. Digital language learning (DLL): Insights from behavior, cognition, and the brain. *Bilingualism: Language and Cognition* 25(3).361–78. DOI: 10.1017/S1366728921000353.

LIU, LINDA, and T. FLORIAN JAEGER. 2018. Inferring causes during speech perception. *Cognition* 174.55–70. DOI: 10.1016/j.cognition.2018.01.003.

LIU, LINDA, and T. FLORIAN JAEGER. 2019. Talker-specific pronunciation or speech error? Discounting (or not) atypical pronunciations during speech perception. *Journal of Experimental Psychology: Human Perception and Performance* 45(12).1562–88. DOI: 10.1037/xhp0000693.

LOPATOVSKA, IRENE, and HARRIET WILLIAMS. 2018. Personification of the Amazon Alexa: BFF or a mindless companion? *CHIIR '18: Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, 265–68. DOI: 10.1145/3176349.3176868.

Maye, Jessica; Richard N. Aslin; and Michael K. Tanenhaus. 2008. The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science* 32(3).543–62. DOI: 10.1080/03640210802035357.

McClelland, James L., and Jeffrey L. Elman. 1986. The TRACE model of speech perception. *Cognitive Psychology* 18(1).1–86. DOI: 10.1016/0010-0285(86)90015-0.

McGowan, Kevin B. 2015. Social expectation improves speech perception in noise. *Language and Speech* 58(4).502–21. DOI: 10.1177/0023830914565191.

McGuire, Grant, and Molly Babel. 2020. Attention to indexical information improves voice recall. *Proceedings of Interspeech 2020*, 1595–99. DOI: 10.21437/Interspeech.2020-3042.

McGurk, Harry, and John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264(5588).746–48. DOI: 10.1038/264746a0.

McQueen, James M.; Dennis Norris; and Anne Cutler. 2006. The dynamic nature of speech perception. *Language and Speech* 49(1).101–12. DOI: 10.1177/00238309060490010601.

Mengesha, Zion; Courtney Heldreth; Michal Lahav; Juliana Sublewski; and Elyse Tuennerman. 2021. 'I don't think these devices are very culturally sensitive.'—Impact of automated speech recognition errors on African Americans. *Frontiers in Artificial Intelligence* 4:725911. DOI: 10.3389/frai.2021.725911.

Munson, Benjamin; Sarah V. Jefferson; and Elizabeth C. McDonald. 2006. The influence of perceived sexual orientation on fricative identification. *The Journal of the Acoustical Society of America* 119(4).2427–37. DOI: 10.1121/1.2173521.

Näätänen, Risto; Anthony W. K. Gaillard; and Sirkka Mäntysalo. 1978. Early selective-attention effect on evoked potential reinterpreted. *Acta Psychologica* 42(4).313–29. DOI: 10.1016/0001-6918(78)90006-9.

Nass, Clifford; Youngme Moon; and Paul Carney. 1999. Are people polite to computers? Responses to computer-based interviewing systems. *Journal of Applied Social Psychology* 29(5).1093–1109. DOI: 10.1111/j.1559-1816.1999.tb00142.x.

Nass, Clifford; Jonathan Steuer; and Ellen R. Tauber. 1994. Computers are social actors. *CHI '94: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78. DOI: 10.1145/191666.191703.

Németh, Géza; Márk Fék; and Tamás G. Csapó. 2007. Increasing prosodic variability of text-to-speech synthesizers. *Proceedings of Interspeech 2007*, 474–77. DOI: 10.21437/Interspeech.2007-222.

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

Norris, Dennis; James M. McQueen; and Anne Cutler. 2003. Perceptual learning in speech. *Cognitive Psychology* 47(2).204–38. DOI: 10.1016/S0010-0285(03)00006-9.

Nusbaum, Howard C.; David B. Pisoni; and Christopher K. Davis. 1984. *Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words*. (Research on speech perception, progress report 10.) Bloomington: Indiana University Press.

Pierrehumbert, Janet B. 2002. Word-specific phonetics. *Laboratory Phonology* 7(1).101–40. DOI: 10.1515/9783110197105.1.101.

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

Purington, Amanda; Jessie G. Taft; Shruti Sannon; Natalya N. Bazarova; and Samuel Hardman Taylor. 2017. 'Alexa is my new BFF': Social roles, user satisfaction, and personification of the Amazon Echo. *CHI EA '17: Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2853–59. DOI: 10.1145/3027063.3053246.

Pycha, Anne; Michelle Cohn; and Georgia Zellou. 2022. Face-masked speech intelligibility: The influence of speaking style, visual information, and background noise. *Frontiers in Communication* 7:874215. DOI: 10.3389/fcomm.2022.874215.

Reinisch, Eva, and Lori L. Holt. 2014. Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance* 40(2).539–55. DOI: 10.1037/a0034409.

Rubin, Donald L. 1992. Nonlanguage factors affecting undergraduates' judgments of nonnative English-speaking teaching assistants. *Research in Higher Education* 33(4). 511–31. DOI: 10.1007/BF00973770.

Samuel, Arthur G., and Tanya Kraljic. 2009. Perceptual learning for speech. *Attention, Perception, & Psychophysics* 71(6).1207–18. DOI: 10.3758/APP.71.6.1207.

Staum Casasanto, Laura. 2008. Does social information influence sentence processing? *Proceedings of the 30th annual meeting of the Cognitive Science Society (CogSci 2008)*, 799–804. Online: https://escholarship.org/uc/item/8dc2t2gf.

Sumner, Meghan; Seung Kyung Kim; Ed King; and Kevin B. McGowan. 2014. The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology* 4:1015. DOI: 10.3389/fpsyg.2013.01015.

Tai, Tzu-Yu, and Howard Hao-Jan Chen. 2020. The impact of Google Assistant on adolescent EFL learners' willingness to communicate. *Interactive Learning Environments* 31(3).1485–1502. DOI: 10.1080/10494820.2020.1841801.

Tamminga, Meredith; Robert Wilder; Wei Lai; and Lacey Wade. 2020. Perceptual learning, talker specificity, and sound change. *Papers in Historical Phonology* 5.90–122. DOI: 10.2218/pihph.5.2020.4439.

Trude, Alison M., and Sarah Brown-Schmidt. 2012. Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes* 27(7–8). 979–1001. DOI: 10.1080/01690965.2011.597153.

van den Oord, Aaron; Sander Dieleman; Heiga Zen; Karen Simonyan; Oriol Vinyals; Alex Graves; Nal Kalchbrenner; Andrew Senior; and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. arXiv:1609.03499 [cs.SD]. DOI: 10.48550/arXiv.1609.03499.

Vaughn, Charlotte R. 2019. Expectations about the source of a speaker's accent affect accent adaptation. *The Journal of the Acoustical Society of America* 145(5).3218–32. DOI: 10.1121/1.5108831.

Waytz, Adam; John Cacioppo; and Nicholas Epley. 2010. Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science* 5(3).219–32. DOI: 10.1177/1745691610369336.

Wilson, Sarah, and Roger K. Moore. 2017. Robot, alien and cartoon voices: Implications for speech-enabled systems. *Proceedings of the 1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots (VIHAR-2017)*, 40–44. Online: http://vihar-2017.vihar.org/assets/papers/VIHAR-2017_paper_1.pdf.

Winn, Matthew. 2014. Gradually blend two sounds. Praat script. Online: http://www.mattwinn.com/praat/GradualBlend_2_sounds.txt, accessed January 1, 2019.

Xie, Xin, and Emily B. Myers. 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. DOI: 10.1016/j.jml.2017.07.005.

Xie, Xin; Rachel M. Theodore; and Emily B. Myers. 2017. More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance* 43(1).206–17. DOI: 10.1037/xhp0000285.

Yi, Han-Gyol; Jasmine E. Phelps; Rajka Smiljanić; and Bharath Chandrasekaran. 2013. Reduced efficiency of audiovisual integration for nonnative speech. *The Journal of the Acoustical Society of America* 134(5).EL387–EL393. DOI: 10.1121/1.4822320.

Yu, Dian; Michelle Cohn; Yi Mang Yang; Chun-Yen Chen; Weiming Wen; Jiaping Zhang; Mingyang Zhou; Kevin Jesse; Austin Chau; Antara Bhowmich; Shreenath Iyer; et al. 2019. Gunrock: A social bot for complex and engaging long conversations. arXiv:1910.03042 [cs.CL]. DOI: 10.48550/arXiv.1910.03042.

Zellou, Georgia. 2017. Individual differences in the production of nasal coarticulation and perceptual compensation. *Journal of Phonetics* 61.13–29. DOI: 10.1016/j.wocn.2016.12.002.

Zellou, Georgia; Michelle Cohn; and Aleese Block. 2021. Partial compensation for coarticulatory vowel nasalization across concatenative and neural text-to-speech. *The Journal of the Acoustical Society of America* 149(5).3424–36. DOI: 10.1121/10.0004989.

Zellou, Georgia; Michelle Cohn; and Bruno Ferenc Segedin. 2021. Age- and gender-related differences in speech alignment toward humans and voice-AI. *Frontiers in Communication* 5:600361. DOI: 10.3389/fcomm.2020.600361.

Zellou, Georgia; Michelle Cohn; and Tyler Kline. 2021. The influence of conversational role on phonetic alignment toward voice-AI and human interlocutors. *Language, Cognition and Neuroscience* 26(10).1–15. DOI: 10.1080/23273798.2021.1931372.

Zhang, Xujin, and Arthur G. Samuel. 2014. Perceptual learning of speech under optimal and adverse conditions. *Journal of Experimental Psychology: Human Perception and Performance* 40(1).200–217. DOI: 10.1037/a0033182.

Zellou
469 Kerr Hall
One Shields Avenue
Davis, CA 95616
[gzellou@ucdavis.edu]
[mdcohn@ucdavis.edu] (Cohn)
[pycha@uwm.edu] (Pycha)

[Received 3 April 2022;
revision invited 20 September 2022;
revision received 7 December 2022;
accepted pending revisions 23 April 2023;
revision received 25 April 2023;
accepted 17 May 2023]