What Does it Mean to be a Computer?
A Sociolinguistic Analysis of Social Meaning
in Perceived Human and Computer Voices

By

ASHLEY ROSE KEATON
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Linguistics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA
DAVIS

Approved:

_____
Georgia Zellou, Chair

_____
Santiago Barreda-Castañón

_____
Katharine Graf Estes

Committee in Charge

2025

Abstract

Speakers vary in how they produce language, and this variation can be stylistic, socially motivated, or based on individual differences. In human language, all communication has social meaning (Giles & Ogay, 2013; Eckert, 2012). But, it is unknown how the same variation would be perceived in "intelligent" computers like voice AI assistants and chatbots. Previous research has shown that when faced with increasing cues of humanity, users apply gender stereotypes, show politeness, linguistically align with, and anthropomorphize social computers (Cohn et al., 2023; Nass, et al., 1997; Sundar & Kim, 2012). Yet, it is unknown whether users evaluate social computer behavior as *socially meaningful*. Because of the importance of sociolinguistic variation in the generation of social meaning, the use of social-indexical variation by computer voices is an apt way to investigate whether users evaluate computers as social actors.

In three experiments, this dissertation research investigated listener evaluations of the sociolinguistic (ING) variable when used by human and computer talkers. Experiment 1 investigated how listeners perceive (ING) variation in human voices compared to computer ones. I found that listeners transfer the social meaning of the standard -*ing* variant to their evaluations of appropriateness for broadcasting and social status, even when the voice was a computer. However, (ING) variation was more influential on listeners' evaluations when it was used by human talkers, and human talkers were evaluated more positively regardless of (ING) use. Experiment 2 tested how college-age students perceive variation in synthetic voices in two different speaking styles. In Experiment 2, which used only synthetic voice stimuli, I found that participants' perception of a voice as a human was primarily influential on their social evaluations, but interactions between speaking style, participant gender, and (ING) variation were also impactful. Experiment 3 used the same design as Experiment 2 to test how older adults evaluate synthetic voices. Older adults evaluated voices they perceived as computers especially

negatively, particularly when the voices used nonstandard *-in'*. I found that older adults' personal voice AI use habits and the talker's speaking style were influential on their evaluations, but I did not find large gender effects like in Experiment 2.

Across studies and across voice types, listener evaluations interact with listener and speaker indexical traits like age, gender, and speaking style, similar to previous research on (ING) in human speech. Belief that a voice was a computer was one of many interrelated social perceptions that influenced listeners' rating behavior, suggesting that computer voices have their own "social class" in the mind of listeners. In other words, synthetic voices that use sociolinguistic variation are evaluated with the social meaning of (ING) from human communication but mediated by perception of humanlikeness and listeners' personal traits.

Looking forward, the ever-increasing sophistication of synthetic voices and social computer design will have complex implications for listener perception of computer voices. Future linguistically informed research in the fast-advancing field of human-computer interaction is essential to understand how people socially relate to nonhuman interactants.

Acknowledgments

First, to Georgia. In every possible way from academic to emotional to physical wellbeing, Georgia has supported me through this phase of life—not just dissertating, not just graduate school, but the last five years in every way. It is not an understatement, it is *the truth*, that I could not have finished graduate school without her. Georgia is so much more than my advisor. I could call her so many wonderful things, but today, I would like to call her my friend.

To Santiago, Katie, and Kelly for their advice and support in completing this dissertation. To UC Davis DataLab, particularly Elise, and every blog post on Statology. To my family, for nodding supportively when I explained my research every Thanksgiving and Christmas. To all my friends and colleagues who made writing this possible by paving the path before me or walking it with me—Chloe, Sophia, Dante, Jeremy, and Catherine. To the watercolor quote on the Communications floor of Kerr Hall that says, "*Science is a team sport*." To every day of my life that I have lived as a bonus day, even the ones that did not feel like bonus days. To Toby.

Table of Contents

Chapter 1: Introduction

Literature Review

Social Meaning in Linguistic Variation

Language variation is a natural property of human speech. All language users speak differently, and listeners are sensitive to differences in how people speak. While some speaker-specific linguistic differences are biological, e.g., age, physical size, and variation in vocal tract length, much speech variation is socially motivated (Barreda, 2020). Socially meaningful variation can occur at all levels of linguistic representation, from prosodic, lexical, to phonological segments (Holliday, 2021; Gessinger, 2010; Vaughn, 2022). Linguistic variation that primarily serves to index a speaker's social traits is known as sociolinguistic variation.

Much phonological variation is systematic; for instance, it can be a result of contextual overlap of adjacent sounds (Zellou, 2022) or can carry meaningful information about a speaker's social identity (Labov, 1986; Foulkes & Docherty, 2006). Sociophonetic variation, or changes in how speakers produce individual phonological segments, is one way that they can signal who they are in the context of the broader social landscape. Phonetic variants can be used to index a myriad of personal traits. Speakers vary their language use systematically based on their gender, socioeconomic status, and belonging to a particular geographic area (Drager, 2010; Kerswill & Trudgill, 2005). As an example of geographic variation, most speakers in Southern California are part of the California Vowel Shift, where back vowels are systematically more fronted than in other regions of the country (Podesva & Callier, 2015; Villarreal, 2018).

Sociophonetic variation can also be used to communicate socioeconomic status and social prestige. Labov (1986) observed that speakers in New York City systematically varied their use of postvocalic (r) based on their socioeconomic status. But speakers varied their use of

(r) when they wanted to be perceived as belonging to a more prestigious social group, suggesting that speakers are aware of the connection between postvocalic (r) and its social-indexical association. In a similar study in England, Trudgill (1972) found that working-class speakers in Norwich rarely pronounced (yu) with the Received Pronunciation (RP) form [yju], but that working-class women were more likely to use the prestige form than working-class men. He also found that women overreported while men underreported how often they used the prestige variant in their everyday speech and proposed that working-class women desired to be perceived as using prestigious language variants. This suggests that socioeconomic status and gender can interact to influence how speakers use sociolinguistic variation.

Sociophonetic variation is also a key factor in how speakers communicate their social and personal traits to the world. Speakers who belong to the same social group converge on their use of the same phonological segments, both to convey to their peers that they identify with that particular social group, and to people outside their social group to signal difference from them (Eckert, 1989). Labov (1963) found that fishermen on the island of Martha's Vineyard increased their use of centralized /ai/ and /au/ as the proportion of mainlanders on the island increased to show their social difference from them. Islanders who were most strongly opposed to mainlander influence were most likely to use the centralized forms and were using their speaking style to assert their social positioning against the mainlanders. Other studies found that the use of nonstandard variation for in-group and out-group behavior can be very fine-grained, down to what lunch table high school students sit at (Bucholtz, 1999a). Eckert (2000) found that students within one suburban high school used backed [ɔ] to signal whether they belonged to the social group of 'burnouts' or 'jocks.' Moreover, the backed [ɔ] is part of the Northern Cities Vowel Shift, and burnouts—high schoolers who identified with the urban areas of Detroit, even though

they did not live there—converged towards the Northern Cities Vowel Shift. On the other hand, jocks signaled their identity with the suburban areas around Detroit by maintaining the more fronted [ɔ] vowel. Students who were "in-betweens" had a more moderate use of [ɔ] than the leading members of the burnout group, but used [ɔ] more than the jocks, suggesting that the adolescents calibrated their use of backed [ɔ] based on how strongly they identified with a social group. Mendoza-Denton (1997) observed a similar phenomenon, where Chicana girls in the same high school used the lax high-front (ɪ) vowel to signal whether they belonged to a gang. Preppy girls never used raised (I), but "wannabe" gang girls used raised (I) an intermediate amount between preppy girls and gang girls.

These studies illustrated a connection between a speaker's variable use and their social positioning. Eckert (2000) found that backed [ɔ] was best predicted by the high schoolers' social group affiliation, not their parents' socioeconomic status, and Mendoza-Denton (1997) found that Chicana girls increased their use of raised (I) once they were officially initiated into the gang. In both cases, the adolescents' variable use did not have to do with who they frequently associated with; it had to do with who they *wanted* to be associated with.

Stylistic variation can even be used by speakers to index themselves in relation to the members of their own social group. Returning to the study of the adolescents at Belten High, Eckert (2008) found that "burnt-out burnouts" used tag questions more than other burnouts in the group. These few members within the larger burnout group wanted to highlight their risk-taking behavior and association with the townies of Detroit. Mendoza-Denton (2011) found that Chicana gang members used creaky voice to signal they were 'hardcore,' meaning they were particularly willing to fight. In both cases, sociolinguistic variables were used to signal an

individual's status within their social group, in addition to indexing their place in the larger community.

The intricate, highly complex nature of sociolinguistic variation makes it clear that speakers' use of linguistic variables to convey their social identity is agentful; they actively select which sociolinguistic variants to use to present a version of themselves to the world (Eckert, 2012). This presentation can also change based on conversational context or who they are interacting with (Eckert, 2008). For example, Podesva (2007) observed that Heath, a gay man, dynamically varied his use of falsetto depending on the situational context. Heath rarely used falsetto voice in conversation while at medical school, but when at social gatherings, his use of falsetto swiftly increased. The degree of falsetto use differentiated a "medical student" persona and a "diva" persona, and Heath, like all speakers, actively selected which persona to portray based on the social context and people he was speaking to. Martín Rojo and Molina (2017) found that interlocutors dynamically update their variable use turn-by-turn throughout a conversation, providing further evidence that speakers actively use sociolinguistic variation for persona construction. Importantly, these character associations are built as an interaction unfolds and in connection with the social positioning of their conversation partner (Eckert, 2012).

Eckert (2008) argued that use of one particular variable is not what indexes speakers to a particular social group. Instead, each sociolinguistic variable has a range of related meanings that a speaker can use dynamically and piece together to create their persona. In one example, Bucholtz (1999b) found that White adolescent men would borrow variants associated with African American English in order to index themselves as tough or masculine, but that they only used these variables when they wanted to invoke that specific association. As shown below, some sociophonetic variants, such as the (ING) variable, are associated with a wide array of

different social groups, and their social meaning shifts significantly based on its relationship to other sociolinguistic variants used by the speaker.

Listener Evaluations of Linguistic Variants

Since speakers vary their speech to signal their belonging to particular social groups, listeners infer a speaker's attributes from their variable use (Cheshire, 2009). Listeners make assumptions about a speaker's social class and group membership from mere seconds of speech (Drager, 2010; Kerswill and Trudgill, 2005). These judgments lead to assumptions about the speaker's personal qualities, with potentially major societal consequences, from likelihood of being hired for a position to potential bias in legal proceedings (Dunbar, et al., 2024; Levon et al., 2021).

The "standard dialect" is an idealized dialect that is evaluated by listeners as more prestigious and socially preferred, while nonstandard variants tend to be evaluated negatively by listeners (Labov, 1969; Levon, et al., 2022). All languages have their own version of a standard dialect; in the United States, linguists refer to the notion of the idealized standard dialect as Standard American English. All speakers believe there is a particular dialect or speaking style that is the "standard" even though people vary in what dialect they perceive to be the standard one (Preston, 1996). Similarly, listeners do not always agree on what dialect is the prestige variant but consistently agree that a prestige variant does exist. Preston (2011) finds that some regional dialects of American English are consistently evaluated more negatively than others, but that participants do not agree on one region where the "best" English is spoken. These findings suggest that all listeners understand the notion of linguistic prestige but differ in which variables they consider prestigious.

The features of Standard American English are frequently contrasted against less prestigious varieties. In general, speakers who use the prestige or standard variant are more likely to be perceived positively by listeners (Preston, 1996). Different varieties of language that deviate from what a speech community deems as the standard variety often have negative social prestige (Preston, 1996). In contrast, nonstandard or less prestigious variants result in more negative character evaluations of the speaker. For example, Received Pronunciation in British English is associated with prestige, politeness, and chivalry, while in the United States, the African American English dialect is judged as inappropriate or even unintelligible (Agha, 2003; Milroy, 2002). These attitudinal judgments occur automatically based on a speaker's voice, without conscious thought from the listener, and can lead to lasting assumptions about the personal characteristics of a speaker (Coupland and Bishop, 2007; Labov et al., 2011; Levon et al., 2022).

Importantly, social-indexical evaluations of language use are not about language—they are about the *person* who uses the social-indexical variant. These evaluations extend beyond the concept of education or social class; speakers who use prestige or standard variants are associated with personality traits like moral goodness, intelligence, and attractiveness (Gordon, 1997; Schleef, et al., 2015). Listeners' judgments of group membership based on sociolinguistic variable use can result in inaccurate or oversimplified assessments of a speaker's characteristics. These beliefs feed forward into stereotyping behavior, so that listeners make attitudinal judgments about speech communities based on already inaccurate beliefs (Baugh, 2016; Campbell-Kibler, 2012; Preston, 2013).

Finally, listeners are not only sensitive to the presence of nonstandard variation, but also how often it is used and how different the pronunciation of that variant is from the standard.

Speakers' usage of a standard or nonstandard variant is often gradient rather than categorical; most speakers use the standard variant in their speech to varying degrees rather than only the standard or only the nonstandard (Eckert, 1989; Labov et al., 2011). If a speaker uses nearly exclusively the nonstandard variant, listeners evaluate that speaker substantially more harshly than a speaker who uses mostly the standard variant and only a few instances of the nonstandard one (Labov et al., 2011).

Listener Age and Gender in Sociolinguistic Perception

The same sociolinguistic variant does not have the same indexical meaning to all listeners (Eckert, 2008). Two frequently explored areas of systemic variation based on listener traits are listener gender and listener age. Below, I provide a brief and focused explanation of the literature on gender and age effects in sociolinguistic perception.

A frequent finding in sociolinguistic research is that women evaluate nonstandard variation differently from men. Chappell (2016) found that women evaluate intervocalic /s/ voicing in Costa Rican Spanish negatively compared to men and argued that women evaluate nonstandard variants more harshly because they are unable to evoke the positive associations of the variant when it is used by men. Eckert (2008) argues that women lead in sound changes, but only when the nonstandard variant has associations that are beneficial to their constructed social-stylistic persona. Gordon (1997) suggested that women use prestige forms to avoid negative social stereotypes associated with the working class. She argues that this may explain why women evaluate other women that use non-prestige forms negatively; they are flouting the social constraints that are enforced for women to a greater degree than men.

Many sociolinguistic studies have found that listener age plays an important role in listeners' evaluations of nonstandard variants. Labov et al. (2011) found that high schoolers were

less critical of the presence of nonstandard variation in their social evaluations compared to undergraduate and graduate students. However, the age of participants in the Labov et al. (2011) study ranged from 18 to 31, with one 46-year-old outlier. Coupland and Bishop (2007) found in their study that participants 45 and older begin to have different evaluations of vernacular versus standard pronunciations. Participants 45 and older were more likely to evaluate vernacular English speech negatively, while participants under 45 did not show a significant difference in their evaluations of vernacular versus Standard English language. Levon et al. (2021), in a large study of UK English speakers, also found that an accent bias distinction emerged around age 45. Levon et al. (2021) and Eckert (2017) posit that the link between older adults and bias against vernacular forms is the result of years spent in a professional environment that enforces standard language norms. These studies suggest that listener evaluations of nonstandard speech become more negative around middle age, but Eckert (2017) points out that drawing perceptual distinctions at a particular age can be an oversimplification. Instead, she argues that generational shifts are an interaction between life stage and chronological age, which are not always the same from person to person.

Listener age and gender can interact to influence how voices are evaluated. Wollum (2019) also found that older women evaluate younger women's speech more harshly than younger women who are evaluating their peers. Levon et al. (2021) found that middle-aged men were considered young adult speakers "less hirable" when they used nonstandard variation compared to younger men. Mechler (2025) found that older men were most critical in their evaluations of talkers compared to other age and gender groups, and that young women gave the most positive ratings.

Finally, research has shown that individual listeners differ in how they evaluate the exact same voice. For example, Campbell-Kibler (2008) found that listeners vary in both how strongly they disfavor nonstandard variation for a particular person and which indexical qualities they associate with that speaker. For example, in her study, some participants evaluated the same speaker as a "surfer dude," while others thought the speaker sounded Southern. Listeners also individually vary in how harshly they rate speakers based on their nonstandard variable use (Preston, 2011). Individual variation among listeners has been shown to be important to understanding social-indexical evaluations, which further illustrates the complex factors that go into the evaluation of a person's voice.

Social-Indexical Variation in American English: The Case of (ING) Variation

A significant amount of work in sociophonetics has focused on the variable (ING) across contexts and speakers. The word-final apical nasal variant [n] is pervasive across English varieties, most prevalently in the context of present progressive verbs (i.e., "talking" pronounced as *talkin'*). Houston (1985) found that (ING) variation in American English is primarily constrained to present progressive verbs ("let's go *swimmin'*"). More rarely, speakers use the nonstandard –in' variant on gerunds ("I have *swimmin'* lessons"), and even rarer still on proper nouns and place names ("he's from *Reddin'*").

The nonstandard variable *–in'* tends to occur more in casual (as in, not careful) speech, but can also carry social meaning (Fischer, 1958). For example, more frequent use of the nonstandard *–in'* variant can be negatively perceived by listeners as informal, unintelligent, and unprofessional (Campbell-Kibler, 2007; Fischer, 1958; Labov et al., 2011). Higher prevalence of the *–in'* variant is associated with certain regional and economic groups. The nonstandard variant has been shown to be perceived by listeners as associated with lower socioeconomic status, the

American South, and African American English speakers (Campbell-Kibler, 2006; Labov, 1972; Trudgill, 1974). Conversely, higher use of standard –ing is sometimes used to signal an individual's alignment with institutions of power. Users of the *-ing* variant are signaling that they follow prestigious English language norms, while the *-in'* variant can be used to evoke a sense of anti-establishmentarianism (Eckert, 2008). This may stem from the fact that speakers are aware of the association of *–in'* with unintelligence and are signaling that they defy this indexical meaning.

More recent research found higher prevalence of the *-in'* variant among adolescents, fraternity brothers, and the queer community, suggesting that use of the *–in'* variant is widespread in many speech communities (Kiesling & Coupland, 2009; Schleef, et al., 2011). (ING) variation occupies an indexical field, and speakers can use (ING) variation to highlight specific indexical characteristics. For instance, fraternity brothers increase their use of the *-in* variant when they want to invoke associations of power or blue-collar masculinity, while gay men use the nonstandard *–in'* variant to signal a "diva" persona (Kiesling, 1998; Kiesling et al., 2009). In this way, the (ING) variable can be used to index very different social traits.

Campbell-Kibler (2006) found that listener associations with *-ing* are multilayered and context-specific. For example, speakers who use the *-in'* variant are perceived as more friendly due to the *-in* variant's association with the American South (Campbell-Kibler, 2006). In another study, when listeners heard a voice they believed was Southern, they evaluated that voice as pretentious; when the voice sounded Northern, use of the *-in'* variant was evaluated as inauthentic (Campbell-Kibler, 2008).

(ING) variation is not only associated with indexical groups, but also personal qualities. The *–in'* variant is associated with lower intelligence and unprofessionalism (Campbell-Kibler

2009; Eckert, 2008; Labov et al., 2011). But speakers who use the *-in'* variant are considered more friendly or easygoing than users of *-ing* (Campbell-Kibler, 2009). Labov et al. (2011) found that listener evaluation of (ING) variation is also conditioned by a speaker's frequency of use of the *-in'* variant. Speakers who use *-in'* are evaluated as unprofessional, and evaluations become increasingly negative with each additional instance of the *-in'* variant in the talker's speech.

Just as speakers vary across contexts and communities in frequency of *-in'* usage, listeners are highly sensitive to (ING) variation and evaluate it differently based on the context. Labov et al. (2011) demonstrated that listeners evaluate speakers who use the *-in'* variant as less professional but more friendly when spoken in a news broadcasting context, while Campbell-Kibler (2010) found that listeners evaluated speakers who used *–in'* comparatively more positively when they were talking about hobbies. Vaughn (2022) found that listeners place greater weight on where they expect variants to be used more than the frequency of the production itself, further providing evidence that listener evaluations of (ING) are context-specific, like other sociolinguistic variables.

Finally, listener perceptions of (ING) are individual and can vary widely. Taken together, studies on (ING) variation illustrate its powerful and complex potential for social meaning.

Listener and Speaker Age and Gender in Perception of (ING) Variation

Like other sociolinguistic variants, (ING) variation is evaluated differently by listeners of different ages and genders. Labov et al. (2011) explored the influence of listener age on (ING) variation. They found that college-age listeners rated speakers increasingly negatively for every new token nonstandard –in' use, so that speakers who used exclusively the –in' variant were heavily penalized compared to speakers who used 70% the standard –ing variant. Meanwhile,

high school listeners responded to increased rates of nonstandard –in' use with either a linear decrease or no decrease in evaluations of social status compared to the standard –ing condition. But, Labov et al. (2011) only assessed young adults and adolescents on their perceptions of (ING), so their study did not address how older adults might evaluate (ING).

A study by Mechler (2025) investigated how (ING) variation is perceived by listeners across the lifespan, including a younger adult group (in their 20s) against a middle-aged group (in their 40s). In her study, Mechler (2025) cross-spliced the –ing and –in' variant into naturalistic language productions by Tyneside vernacular English speakers of different ages and collected listener evaluations from a range of ages and genders. She found that both listener and speaker gender and age were influential on listeners' evaluations of the talkers, but (ING) had only a small impact. She proposed that gender and age were more relevant to listener evaluations of the talkers than (ING) in her study. One explanation for this finding might be that listeners made judgements about the speakers based on their identification of them as Tyneside English speakers, which is a highly salient vernacular style that is the subject of polarizing social perceptions. Speakers in the Mechler (2025) study were also evaluated as having a high degree of overall vernacularity, in contrast to other studies such as Labov et al. (2011) and this dissertation research. I also use synthetic voices that are designed to speak Standard American English, because it is considered the most prestigious dialect of American English (Sutton et al., 2019).

Human-Computer Interaction

Because all language is social, a natural question is how language is perceived when it is used by non-human agents, or computers. Human-computer interaction (HCI) is a broad field of study that examines how humans interact with intelligent machines and compares those

behaviors to human-human interactions (Clark et al., 2019). Within this broader field, many human-computer interaction studies have focused on language use by computers. Some of the first work in human-computer interaction and language was pioneered by Nass and colleagues. In a series of studies, they found that users transfer many socially motivated behaviors from human communication into interactions with computers. Nass et al. (1997) found that users applied gender stereotypes when interacting with female- and male-sounding voices, while Nass et al. (1994) found that users evaluated computers using face-saving strategies. Participants gave more positive evaluations to a computer's competence on tutoring topics when using that same computer, compared to when they gave evaluations of that computer's performance when answering survey questions on a different computer. Nass and Lee (2001) found a similarity-attraction effect where people gave higher social ratings to computers that appeared to align with the user's own personal degree of extroversion or introversion. In another study, Moon and Nass (1996) found that that were evaluated as extroverted or introverted were rated higher when their degree of extroversion matched the conversational context. For example, introverted voices were more positively rated when giving book reviews, while extroverted voices were more positively rated when they were giving travel advice. Importantly, Nass and Steuer (1993) showed that users primarily distinguish social computers based on voice rather than the physical computer itself.

In the late 1990s, Nass and colleagues proposed the *computers as social actors* (CASA) theory to explain the findings of their array of studies on human users' evaluations of social computers. This theory applied findings from human social psychology onto human-computer interactions. They proposed that users' social behavior with computers is the result of unconscious anthropomorphism, particularly when devices use speech (Nass et al., 1994). This

was the first of many frameworks in human-computer interaction that have developed over the past several decades. The *media equation* theoretical framework is a development of the CASA framework that extends the scope from computers to a broader class of "media" (Reeves & Nass, 1996). Media can include computers, digital avatars, robots, or chatbots that have characteristic cues of humanity such as the presence of a face, emotions, or anthropomorphic movement (Banks, 2020).

Sundar and Kim (2019) propose that there are routinized patterns of a "machine heuristic" based on users' previous experiences with the role and capability of 'smart' computers based on their previous experiences with computers and that this heuristic is activated when an interaction with a computer fits users' prior expectations. A heuristic can be thought of as a mental shortcut that users take to make assumptions about subsequent actions when a frequently repeated experience fits with previous experiences (Tversky & Kahneman, 1974). Social behavior towards computers increases the longer that an interaction goes on, suggesting that longer interactions present more opportunities for users to observe cues of socially directed speech and thus more strongly activate routinized social behaviors (Kim & Sundar, 2012).

Recent work in HCI introduced the concept of "social scripts," linguistic behaviors that are similar to human social communication but constrained to the contexts and routines where users most typically interact with social computers (Gambino et al., 2020). Since script-specific behaviors from human communication like politeness are transferred to computer interactants, script-specific behaviors with computers might branch off into a unique social script themselves. One example where linguistic scripts have been explored in human-computer interaction studies is the extension of the "small talk" heuristic, where highly frequent exchanges of polite greeting are routinized and enacted without conscious effort from the interactants (Kádár & House, 2021).

Endrass et al. (2011) found that users' expectations for culture-specific small talk with humans are transferred to interactions with chatbots, providing evidence that script-specific behavior is transferred to computers.

Studies on social scripts for computer interactions have historically primarily analyzed human-computer communication through text, but Sundar et al. (2015) point out that recent developments in social computer design have become increasingly interactive, which further increases the possibility for social behavior to be transferred to social computers. Indeed, the number of interactions that a person in the United States has with talking computers for social purposes has increased dramatically in the last ten years (Olmstead, 2017).

Yet, it is currently understudied how computer-directed speech fits within the social scripts framework. In linguistics, studies have shown that people have script-specific behavior for different speaking contexts, like child-directed speech or foreign-accented speech (Aoki and Zellou, 2024; Foulkes et al., 2005). Zellou et al. (2023) found a key difference in human-computer linguistic behavior where listeners expect synthetic voices to produce homogenous and invariant linguistic patterns than what would typically be seen in human-human interaction. This pattern has been found for listeners perceptions of a socialbot's proficiency in comprehending human speech as well. Aoki et al. (2022) found that users' language production patterns predict that people consider voice AI technology as less communicatively competent than human talkers. This effect was also found for in a matched-guise experiment with synthetic and human speech where listeners considered both voice types less intelligible if they believed them to be computers.

Finally, the machine heuristic and social scripts theories are characterized by the assumption that users overtly know that the social potential of computers differs significantly

from humans and therefore do not attempt higher-level elements of human social behavior like relationship maintenance or face-saving (Sundar and Kim, 2019; Gambino and Liu, 2022). In contrast, speech is often used for relationship maintenance or development in human interactions (Giles & Ogay, 2013). This highlights a potentially important difference between human-directed and computer-directed speech; language is inherently prosocial, and the absence of sociality in human-computer interaction has implications for our understanding of social meaning in spoken communication.

Interaction of Age on Social Perception of Synthetic Voices

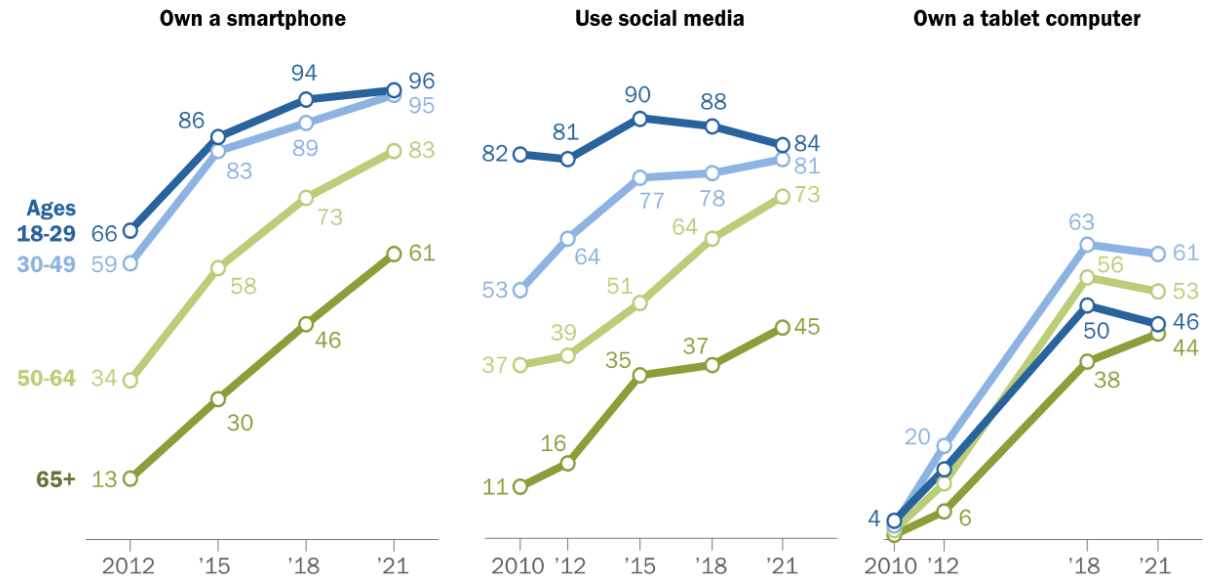
As mentioned in above, listener age has a significant influence on their evaluation of a speaker's language use. Moreover, listeners' personal beliefs, attitudes, and societal values strongly influence both their use and perception of language (Eckert, 2012). It is not surprising, then, that adults in different age groups have vastly different perceptions of the role of computers in society, and by extension their expectations for the linguistic behavior of computers.

Older adults use technology in markedly different ways than younger adults, particularly for social purposes (Auxier & Anderson, 2021). These findings are reflected in how many adults 50 and older use smartphones and social media compared to people ages 16 to 49, as shown in Figure 1.1 (Pew Research, 2021). Users 16–49 have very similar rates of smartphone and social media use across young and middle ages; 96% of people between 16 and 49 own a smartphone and around 82% use social media at least monthly. However, users 50 to 64 are markedly less likely to own a smartphone (83%) and use social media (73%), and another marked decrease is shown for users 65 and older, of whom 61% own a smartphone and less than half (45%) use social media.

Figure 1.1
Smartphone and Social Media Use Divided by Age Cohorts

Smartphone ownership and social media use among older adults continue to grow

% of U.S. adults who say they ...



Note: Respondents who did not give an answer are not shown.
Source: Survey of U.S. adults conducted Jan. 25-Feb. 8, 2021.

PEW RESEARCH CENTER

Note. Data is from a sample of 1,502 respondents. *Pew Research Center, 2021.*

Beyond technology usage, there are also stark ideological divides between older and younger adults on the role of artificial intelligence and social computers in society. Older adults are less trusting of generative AI and are more concerned about ethics of generative AI than younger adults (Thormundsson, 2024). Strikingly, adults from 18 to 44 overall consider AI to have a positive impact on their life compared to a negative one, while adults 45 and older consider voice-AI to have a negative effect on their lives (Thormundsson, 2024).

While the attitudinal judgments of older adults have been explored, little is known about how these expectations transfer to their language use with computers. Some work has already investigated the potential interaction between speaker and listener age in human-computer interaction. Zellou et al. (2021) found that people linguistically align more to synthetic voices that have similar apparent ages to the participant's than voices that sound older or younger. This

is similar to findings from sociolinguistic studies in human-human communication described above.

The CASA, media equivalence, and routinized social script theoretical frameworks from human-computer interaction present testable ways to make predictions about how people will interact with computers through language. I propose that human-computer interaction theories can be extended to linguistic behaviors with computers, taking into consideration the use of sociolinguistic variation in computer speech. This dissertation synthesizes human-computer interaction studies, which historically were largely undertaken in the field of computer science, and extends them in meaningful ways to linguistic theory of social communication.

Societal Issues in the Development of Computer Voices

Understanding the relationship between computer speech and social perception extends beyond theoretical implications; the emergence of computers as a new social actor will result in lasting changes to our social landscape. Even in the early studies of human-computer interaction, there is a long-standing finding that users interact with computers differently based on the perceived gender of the computer voice (Nass et al., 1997). Voices that sounded feminine are considered more knowledgeable in stereotypically feminine topics such as love and relationships compared to male voices. Computer voices that were giving instructions were rated more positively when those voices sounded masculine. Their findings showed clear patterns of gender-stereotypic behavior transferred to interactions with computer voices.

It is not coincidental, then, that intelligent voice assistants like Siri and Alexa are developed with feminine voices (Chen et al., 2022). Although tech companies claim that users prefer listening to a female voice, Nass and Brave (2005) showed that users specifically prefer feminine voices when they are in helpful or submissive roles (West et al., 2019). In 2019,

UNESCO published a report detailing how the stereotypical feminized and submissive voices of digital voice assistants were perpetuating harmful stereotypes about the role of women in society. The report argues that the the typical configuration of a voice assistant as a young unassertive and unconditionally positive woman influences user expectations for subsequent voice assistant interactions, and by extension people's expectations for real women (West et al., 2019). The interaction of female submissive stereotypes and the design of voice assistants have created a feedback loop where gender stereotypes led to the development of gendered voice assistants, and gendered voice assistants reinforce social stereotypes about the role of women in society.

Beyond the feminine and deferential stereotypes associated with device voices, there is also a tendency for synthetic voices to be White, speak "standard" English, and be hyperarticulate. For example, the development team of Google Voice created a "backstory" for the AI voice as part of their development process. Google Voice is personified as a young woman from Colorado who attended Northwestern University, enjoys kayaking, and, as a child, competed on Jeopardy Kids and won $100,000 (Shulevitz, 2018). While this persona is not explicitly described as "White," many of the personal qualities of Google Voice are associated with Whiteness, or at least privilege, and the association of Whiteness with "smart" computers can reinforce stereotypes that White voices are more prestigious or powerful. Sutton et al. (2019) proposed that users have the same implicit association for device voices, and that the predominant use of White-sounding, standard English voices reinforces social stereotypes of the White ideal. Lippi-Green (1997) proposed that the voices used for fictional characters in media are based upon and reinforce stereotypes about sociolinguistic variation and prestige. In her study, she found that characters in Disney animated movies that used nonstandard variation were more likely to play villainous or ethically grey characters. In fact, Holliday (2023) found that

listeners evaluate the Apple Siri voices with different demographic traits, and the voice that was most often perceived as Black, male, and young was also the voice judged as less professional and less competent than the other Siri voices. Other studies have investigated listener perceptions of marginalized populations such as nonbinary or queer speakers, but more research is necessary to understand how users would evaluate finer-grained linguistic differences in device speech (Danielescu, 2020). Pradhan and Lazar (2021) point out that persona design is becoming a central part of the development of conversational agents, so it is an especially timely moment to examine how linguistic biases might emerge in the development of computer voices.

Studies on linguistic variation in computer voices to date have largely focused on listener evaluations in terms of macrosocial categories. Sutton et al. (2019) also point out that it is unknown how listeners would respond to fine-grained variation in computer speech rather than macrosocial categories like gender, age, or extroversion. In their article, they advocate for the use of nonstandard linguistic variation and non-White or foreign accented speech in device voices in order to combat the prevailing stereotype that powerful voices use standard English.

Another consideration in user perception of computer voices is the relatively small amount of social agency that voice assistants hold. Voice assistants are developed to be unconditionally positive, nonconfrontational, and deferential to the user's opinions (Gambino & Liu, 2022; Ta et al., 2020). This impacts how listeners approach interactions with voice assistants and oftentimes emboldens users to speak abusively to chatbots because there are not negative consequences as in human-human interactions.

While it is presently unclear whether users consider interactions with smart computers to be socially meaningful, the constructed persona of a "smart" computer is prefabricated by developers, and the conversational dynamics of a human-computer interaction are also

stipulated. In human communication, a person's social positioning against their interlocutor is established as the conversation unfolds, and persona construction occurs in reference to their interaction partner's social positioning (Eckert, 2012). Interactions with intelligent voice assistants are comparatively very constrained. Users know that intelligent voice assistants will not be confrontational, dominant, or even disagree with them, which decreases the likelihood that any of the remaining aspects of an interaction with a voice assistant would be considered socially meaningful. This dissertation investigates whether listeners consider sociolinguistic variation in computer speech to be socially meaningful.

To answer these broader questions in human-computer interaction and sociolinguistics, this dissertation explores how one sociolinguistic variant—(ING)—might influence listener's perceptions of computer voices. Rather than attempt to create synthetic voices that are free of stereotypes, which is inherently impossible, my research more precisely maps our understanding of how sociolinguistic variation in synthetic speech influences listeners' social connections to real human talkers.

Language is a unique human ability and is closely tied to our perception of what makes us "human." However, synthesized speech is modeled on human voices and has become increasingly humanlike, even to the point that "voice deepfakes" can be perceptually indistinguishable from human voices to some listeners (Müller et al., 2022). Developers of voice-AI technology for intelligent voice assistants purposefully develop voices to be personable and humanlike (Sutton et al., 2019). In many directions, listeners are receiving cues to interact with voice-AI as though they are humans. Yet, how listeners evaluate sociolinguistic variation spoken by computer voices, and its implications for our understanding of what it means for phonetic variation to be "socially meaningful" is understudied in scholarly work. This dissertation

research refines the media equivalence routinization, and social scripts frameworks of human-computer interaction by investigating the nature of social scripts in human-computer interaction from a sociolinguistic perspective.

The Dissertation Experiments

The overarching question of this dissertation is:

How does gradient use of sociolinguistic variation affect listener evaluations across human and computer voices?

The dissertation studies are designed to address gaps in the literature on both human sociolinguistic perception and human-computer interaction. They advance our understanding of how listeners perceive computer voices, and whether social-indexical language use in computer speech is considered socially meaningful. I use Bayesian statistically informed analysis to answer my research questions with statistical robustness, a step that is often overlooked in traditional sociolinguistic studies. I also examine how the findings of the dissertation studies can be integrated into technology for societal benefit.

Research Questions and Predictions

In three experiments, I ask research questions that address how (ING) variation is perceived, how listener and speaker social-indexical traits interact in listeners' social evaluations of (ING), and whether evaluation patterns for (ING) variation are similar across human and synthetic voices. I ask 6 specific research questions:

Research question 1.
Do listeners notice (ING) variation when used by synthetic voices?

Research question 2.
Do listeners evaluate (ING) variation in synthetic voices similarly to human ones?

Research question 3.
Does social context of (ING) variation influence listener evaluations of synthetic voices?

Research question 4.
When listeners mistake a synthetic voice for a human one, does this result in evaluation patterns that are similar to those for human voices?

Research question 5.
For each of the above research questions, are these evaluations influenced by speaker and listener traits of age and gender, and can these indexical traits interact?

Research question 6.
Does a listener's perception of a voice as a human or computer interact with the listener's own social-indexical traits, particularly listener age?

Hypothesis 1.
I expect that listeners will unconsciously attend to (ING) variable use in synthetic speech, similar to unconscious evaluation of sociolinguistic variants in human speech.

Hypothesis 2.
I predict that some sociolinguistic evaluations of (ING) variation would transfer to synthetic voices, but in a more limited way than for human voices. For example, listeners expect synthetic voices to be professional, and nonstandard *-in'* use will likely violate their expectations for synthetic voices and cause them to rate them negatively. On the other hand, listeners might not think of synthetic voices as friendly, and therefore the association between *-in'* and friendliness would not be invoked.

Hypothesis 3.
I expect that the *belief* that a voice is a human or computer would be the primary factor in listener evaluations of synthetic voices, not the acoustic features of a synthetic voice. If this is the

case, I expect to find that listeners would evaluate computer voices using typical sociolinguistic evaluation of (ING) when spoken by humans. In other words, I expect that sociolinguistic evaluations would be top-down and based on listeners' assumptions about the world, not bottom-up and based on auditory-acoustic input.

Hypothesis 4.
I predict that, in line with existing knowledge from the field of sociolinguistics, listener and speaker social-indexical characteristics would interact. More specifically, I expect more negative evaluations from older adults and women, particularly older women.

Hypothesis 5.
I expect to find that older adults will evaluate computer voices that use the nonstandard variant more negatively compared to younger adults, and that they will evaluate computer voices more negatively overall due to their beliefs about the role of voice AI in society.

Hypothesis 6.
I expect more negative evaluations for nonstandard variation usage in the formal speaking context, and either neutral evaluations or more positive social evaluations for nonstandard usage in a more informal speaking context.

Outline of Chapters

In Chapter 1, I lay out relevant knowledge from the fields of sociolinguistics and human-computer interaction and contextualized them within my current research questions. Chapter 2 outlines the methods that are used in all the dissertation experiments: methods that are specific to one particular experiment are described in that chapter.

In Chapter 3, I present Experiment 1, where I investigate how listeners perceive synthetic voices that vary in their productions of (ING) directly compared to human speakers. This study provides a baseline of understanding whether there are broadly similar patterns of social

evaluation for synthetic voices compared to human ones. In Chapter 4, I present Experiment 2, and in Chapter 5, I present Experiment 3. In these experiments, I explore social evaluations of synthetic voices only. I examine whether speaker and listener social-indexical traits of gender and age interact, and if they follow the same evaluations as for listener social evaluations of human voices compared to existing research knowledge from sociolinguistics. I also investigate if listener evaluations of synthetic voices change based on whether they thought the voice was a human or computer, compare whether there are differences in identification accuracy between age groups, and investigate if listener age and perception of a voice as a computer interact.

Chapter 6 synthesizes the findings of Experiments 1, 2, and 3 and interprets them both in the lens of sociolinguistic and human-computer interaction theory. I also propose broader applications of the findings of the dissertation experiments for the development of voice AI technology in consumer products. Finally, I discuss limitations of the dissertation studies and future directions for research.

Chapter 2: Explanation of Methods

Relatively little work has investigated how sociolinguistic variants might be evaluated in human-computer interaction, and experimental approaches in social perceptions of computers are not always sociolinguistically informed (Sutton et al., 2019). Sutton et al. proposed that studies on fine-grained sociolinguistic variation in computer speech would contribute to our understanding of human-computer social behavior more generally. With this recommendation in mind, the present work combines techniques from sociolinguistic and human-computer interaction studies to examine listener evaluations of device voices. I extend the sociolinguistic paradigm of matched guise studies to this research which includes both human and naturalistic synthetic voices.

Experiment Design

Matched Guise Technique

In sociophonetic studies, the matched guise technique is used extensively to test how speakers' variable use affects listeners' social evaluations. In this approach, a speaker is recorded producing both the standard and the nonstandard forms of the variant. Then, listeners are instructed to listen to the two recordings of the same script under the guise that they were produced by two different talkers and asked to give social evaluations of the two recordings. Listeners remain unaware that the different versions of the script were spoken by the same person (Ball, 1983; Lambert et al., 1960). This technique was first developed by Lambert et al. (1960) to test whether English speakers and French-Canadian speakers would evaluate a speaker as more friendly when they spoke the same first language as them. In fact, listeners unknowingly heard the same French-English bilingual speaker but preferred the speaker that read the script in

the participants' first language. This approach allows researchers to isolate language use as the conditioning factor in listener evaluations, rather than any speaker-specific voice qualities.

Subsequent studies used the matched guise technique for increasingly fine-grained variables of interest. Ball (1983) used the matched guise technique to compare listener evaluations of two Australian dialects, and Purnell et al. (1999) used the matched guise technique to evaluate differences in how landlords responded to apartment applicants that had different ethnic dialects. The matched guise technique has contributed to our understanding of listener judgments of class membership, ethnicity, social groups, and personal qualities (e.g., Loureiro-Rodriguez et.al, 2012).

Cross-splicing sociolinguistic variants

While the classic matched guise technique uses one speaker that "code-switches" between the standard and nonstandard dialect, studies have also cross-spliced small phonological segments into the same passage to elicit participant responses to segmental variants (Campbell-Kibler 2007, 2009; Labov et al., 2011).

Advancements in voice editing technology (namely Praat; Boersma & Weenink, 2023) enable researchers to investigate very small segmental differences using cross-splicing techniques. In cross-splicing, the same speaker records a passage multiple times, using the standard or nonstandard variant. Then, the researcher cuts the segment of interest from one of the recorded passages and replaces it into the body passage (Scarborough, 2005). The same body passage is used, and the standard and nonstandard segments are spliced into it, so that the only differences in the passages are the variables of interest. This cross-splicing technique has been effective in eliciting participant evaluations of subtle linguistic differences, such as the same speaker differing only in their pronunciation of the California vowel shift (Villarreal, 2018).

Cross-splicing has been used in studies on perception of (ING) variation, such as Labov et al. (2011) and Campbell-Kibler (2007; 2009). Vaughn (2022) cautions that cross-splicing itself can be a factor in listener evaluations of segmental variation. Although cross-splicing techniques are not completely naturalistic, they do allow for highly controlled experiment design.

In Experiment 1 of my study, I use the cross-splicing methods for the human voices described above. For the synthetic voices in Experiments 1, 2, and 3, I generated the stimuli passage each of the four levels of (ING) variation tested in the study. I chose to generate the passage anew rather than cross-splice for synthetic voices because the smoothness of neural synthetic speech makes it difficult to isolate small segments (Tobing et al., 2019). Directly generating the *–in'* segment led to better intersegmental continuity in the synthetic voices.

Human-Computer Interaction Experimental Techniques

Because human-computer interaction is a relatively novel field and interdisciplinary topic among cognitive science, computer science, and engineering fields, a wide variety of methods are used that are often developed ad-hoc for individual studies (Clark et al., 2019). In studies on social evaluations of computers, a frequently used paradigm developed by Clifford Nass and colleagues asks participants to evaluate computer voices after interacting with the voice for a short task. For example, participants might listen to book reviews (Gong & Nass, 2007; Nass & Lee, 2001) or tutoring sessions (Nass et al., 1994; Nass et al., 1997) and then offer feedback on the efficacy or likability of the computer voice. This method asks participants to evaluate the voices in a concrete scenario where listeners can imagine the utility of each voice in a given context. These dissertation studies follow this paradigm by asking participants what voice they would like to hear in a computer application ("app"). This prompt, which can be found in Appendix Figure A (Experiment 1) and Appendix Figure C (Experiments 2 and 3), cues listeners

to think about the actual usefulness of a voice and implies that they have an influence over how voices should be designed. As suggested in the studies by Nass and colleagues, I expect that giving participants a sense of agency in the design of voice AI will result in better attention to the evaluation task.

Stimuli

In each of my experiments, participants listened to four voices read the same script. Experiment 1 used a different script from Experiments 2 and 3, which used identical passages.

Experiment 1 was adapted from Labov et al. (2011), using the same script passage as theirs for my recordings, which describes the congressional debate of the EGTRRA Bush administration tax cuts (GovInfo, 2001). The script used in Experiment 1 can be found in Appendix Figure B.

During Experiment 1, some participants commented that the political topic made them feel stressed or sad and therefore did not want to listen closely. For Experiments 2 and 3, I developed a news broadcast that describes a high school student winning a scholarship, with the idea that this would be a neutral topic. I also developed a radio DJ broadcast passage that describes a fictional playlist created by the speaker and the speaker's current music interests. Experiments 2 and 3 included two different types of broadcasts to investigate whether differences in speaking context would influence listeners' evaluations of (ING). The passages designed for Experiments 2 and 3 were developed to be as syntactically similar to the Labov et al. (2011) experimental stimuli as possible. The passage contains ten instances of (ING), seven total sentences, and is roughly one minute long. In both passages, there are no present progressive verbs as the first or last word of the sentence and no other words that end in [ɪn] or

[ɪŋ]. The passages used as stimuli in Experiments 2 and 3 were developed by A.R. Keaton in 2023 and can be found in Appendix Figure D.

Selection of Synthetic Voice Stimuli

In the same way that sociolinguists carefully select the speakers they use to record stimuli for matched guise techniques, we also must make informed decisions about how the synthetic voice stimuli should be produced. In contemporary text-to-speech generation, two main synthesis methods are available. Concatenative speech synthesis combines phone-sized segments spoken by a human talker into new combinations of segments based on the input text (Hunt & Black, 1996). One benefit of the concatenative method is that the individual phones sound naturalistic, but the major drawback is the difficulty in aligning adjacent phones without perceptible discontinuities (Stylianou & Syrdal, 2001). Although voice naturalness from concatenative methods has improved over time, disfluencies between phones can still be perceived by listeners (Kuligowska et al., 2018).

In contrast, neural speech synthesis methods use machine learning to generate synthesized speech modeled on human speech data. Many machine learning networks for text-to-speech have been developed, but they share some basic principles (Ning et al., 2019). Deep learning models are given pairs of acoustic waveforms and input text features as training data. The model estimates parameter weights for the relationship between the input features and acoustic waveforms. After training, the model weights are used to generate new acoustic parameters and are adapted into waveforms through vocoding (Ning et al., 2019). A key benefit of neural synthesis methods is that the generated speech is consistently smooth and continuous (Tobing et al., 2019). Additionally, given enough input data and training time, large language models can produce highly naturalistic speech (Perera & Ganegoda, 2023).

I selected neural speech synthesis methods for use in this study because of its improvements over concatenative synthesis methods for naturalistic prosody generation and intersegmental continuity. Discontinuities would create difficulties in manipulation of the stimuli, given that I was investigating a single segmental variant. Additionally, Cohn and Zellou (2020) generated neural and concatenative synthetic voices from the same speech generation platform and found that the neural synthetic voices were perceived by listeners as more naturalistic than concatenative ones.

However, selection of synthetic voice stimuli goes beyond the choice between concatenative and neural synthesis methods. Speech generation continually improves year over year; neural synthetic voices from previous years sound noticeably less natural than more recent ones. In Experiment 1, I used the Amazon Web Services (AWS) voices "Kendra" and "Kimberly," generated from the most current version available in 2022. Both synthetic voices used in Experiment 1 were feminine to have the same apparent gender as the female human speakers.

In Experiments 2 and 3, I used Google WaveNet neural synthetic voices, generated from the most current version available in 2023. I elected to use a different set of synthetic voices than in Experiment 1 for two reasons. First, the main goal of this research is to understand how naturalistic synthetic speech is changing our social landscape, and therefore it is appropriate to work with the most naturalistic synthetic voices commercially available at the time of synthetic speech generation. Second, as a practical consideration, the AWS voices produced artifacts when producing the nonstandard *-in'* on some words (*awardin'*, *addin'*) in the broadcast passages whereas the WaveNet voices synthesized all of the nonstandard variants naturalistically. Experiments 2 and 3 used four synthetic voices, two of each apparent gender. Voice D and Voice

I were designed by Google to sound like masculine voices, while Voices F and H were designed to sound feminine.

For all three experiments, I generated speech from each synthetic voice reading the passage with the *-ing* variant then generated the same passage with the *-in'* variant, which resulted in a more naturalistic result than cross-splicing. All stimuli were amplitude normalized to 65 dB. Because human voice stimuli were only used in one of the three experiments, I describe the generation of the human voice stimuli in detail in Chapter 3 (Experiment 1).

Participants

Participant Recruitment and Demographics

Participants in this dissertation research were either undergraduate students recruited from the University of California, Davis (UCD) or users of the online survey collection platform Prolific. Undergraduates recruited at UCD were compensated with partial course credit for their participation in the study. Participants recruited from Prolific were compensated $6.00 for their responses ($18/hour based on a 20-minute average experiment run time). Experiments 1 and 2 surveyed only UCD undergraduates, while Experiment 3 surveyed only Prolific participants. Participants were eligible to participate in the experiments if they spoke English as their strongest language and had no hearing impairments. Participants were only able to participate in one of the experiments.

At the end of the experiments, participants completed a demographic questionnaire. Participants were asked to self-specify their gender. I collected responses from people of all genders, but of the 606 participants across the three experiments, only 4 were nonbinary or genderqueer. Thus, I was not able to include nonbinary people in our statistical models of listener gender because there were so few nonbinary participants.

For the purposes of analysis, participants from 18 to 39 are considered "young adults," and participants 40 and older are "older adults." I made this decision based on general technology use by different age cohorts as well as previous knowledge from sociolinguistics that suggests that people begin to evaluate the nonstandard variant more negatively once they move into the working professional portion of their life (Coupland & Bishop, 2007; Labov, 1972; Levon et al., 2021; Pew Research Group, 2021). I use this definition to place participants into Experiment 2 and Experiment 3.

Table 2.1
Participant Age Cohorts in the Dissertation Studies

Experiment	Collection Platform	Age Group	Age Range
Experiment 1	UCD	Young Adults	18 to 24
Experiment 2	UCD	Young Adults	18 to 33
Experiment 3	Prolific	Older Adults	40 to 86

Data Collection Quality

Whenever data is collected from online survey platforms, an important consideration is whether participants paid close attention while completing the experiment. To check that participants were paying attention, we required subjects to complete a word identification task before the start of the task to ensure they had the audio at a listenable volume and paid attention to the words that were said. Further, Prolific allows researchers to select participants that have good ratings for quality attention to the survey tasks, so I selected participants with 95% quality scores or better. Participants were instructed that the survey would take about 20 minutes and had to be completed in one sitting. Participants were also instructed to wear headphones, silence outside noises and distractions, and maintain a stable internet connection throughout the course

of the experiment. Participants selected a box agreeing to these terms before the start of the experiment.

Task Design

Survey questions

I designed my survey questions to assess the most important factors of my research questions but also kept them brief to minimize the possibility that listeners will be fatigued by answering a repeated battery of questions. Because listeners complete the exact same procedure for four voices and answer the questionnaire for each of them, I chose to ask nine questions for each voice.

The questions assess four themes: social status, trustworthiness, likability, and social presence. I collected ratings of intelligence, professionalism, and appropriateness for newscasting (dimensions related to social status), friendliness and trustworthiness (dimensions related to likability), naturalness, authenticity, and the two social presence questions adapted from Lee and Nass (2005) (dimensions related to humanlikeness). Lee and Nass' social presence questions were developed to index how "real" listeners considered a computer voice. From their broader set of questions, I chose to use the two that test vividness and specificity. All rating questions in the study used the same question frame, *"How __ is this voice?"* with the exception of appropriateness and the two social presence questions. For every evaluation question, participants responded using a sliding scale from 0–100 that was centered at 50. After the sliding scale questions, there was a free-response question where they could optionally include any additional thoughts they had about the voice. The questions used in the study are provided in Table 2.2.

Table 2.2

Survey Questions Used in the Dissertation Experiments

1. How professional is this voice? 2. Is this voice good for news broadcasting/a radio DJ? 3. How friendly is this voice? 4. How natural is this voice? 5. How authentic is this voice? 6. How trustworthy is this voice? 7. How intelligent is this voice? 8. While you were hearing the broadcast, how much did you feel as if the voice was talking to you specifically? 9. While you were hearing the broadcast, how vividly were you able to mentally imagine the person who was speaking? 10. Do you have any other thoughts about this voice that you would like to share with us?

Note. Participants viewed this question list before the start of the experiment and after listening to each voice.

For each experiment, I use Principal Component Analysis (PCA) independent of the other studies to investigate whether the nine rating questions are clustered similarly to the four themes I set out to investigate. As described in the experiment design, when the results of the PCA differ from my preconceived thematic grouping, I use the groups identified by PCA for statistical modeling because they more accurately represent that participant group's interpretation of the nine rating metrics. PCA is described in more detail in the statistical analysis methods section below.

Experiment Procedure

All data in the three dissertation experiments follow the same task design and all participant responses were collected remotely. Each task began by prompting participants that I was collecting survey responses on which voice would be best to use for an app I was developing. Prior to the start of the task, participants were informed what rating questions they would be answering for each voice and were asked to keep those criteria in mind when listening to the voices. The experiment prompt for Experiments 1, 2, and 3 can be found in Figures A and C of the Appendix.

After reading the task instructions, participants were directed to a blank page that contained the text “Please listen to this voice,” and a recording of the first voice played automatically. When the recording ended, the page auto-advanced to a new screen where participants used a sliding scale from 0–100 to respond to each of the nine rating questions. Participants listened to each recording separately and answered all rating questions for that voice before moving onto the next one. Participants completed this procedure for each voice. The order that each voice appeared in was randomized across participants.

After completing the evaluation phase of the experiment, participants heard each voice repeat a seven-second clip of the passage and then selected whether they thought that voice was a human or computer. Participants had the opportunity to listen to the clip an unlimited number of times while making their decision. Participants were asked to identify the voices as human or computer at the end of the study rather than within the evaluation blocks to avoid cueing participants that I was testing non-human voices.

Lastly, participants completed a demographic questionnaire where they self-identified their gender, age, what languages they speak and what language they consider their strongest language, their voice AI use frequency, and what they use voice AI for (if applicable). Participants also selected whether they think synthetic or human voices are better for news or radio broadcasting overall, depending on whether they were in the news or radio listening group. As a last step, participants had the opportunity to leave any additional comments about the experiment.

Statistical Analysis Methods

Principal Component Analysis

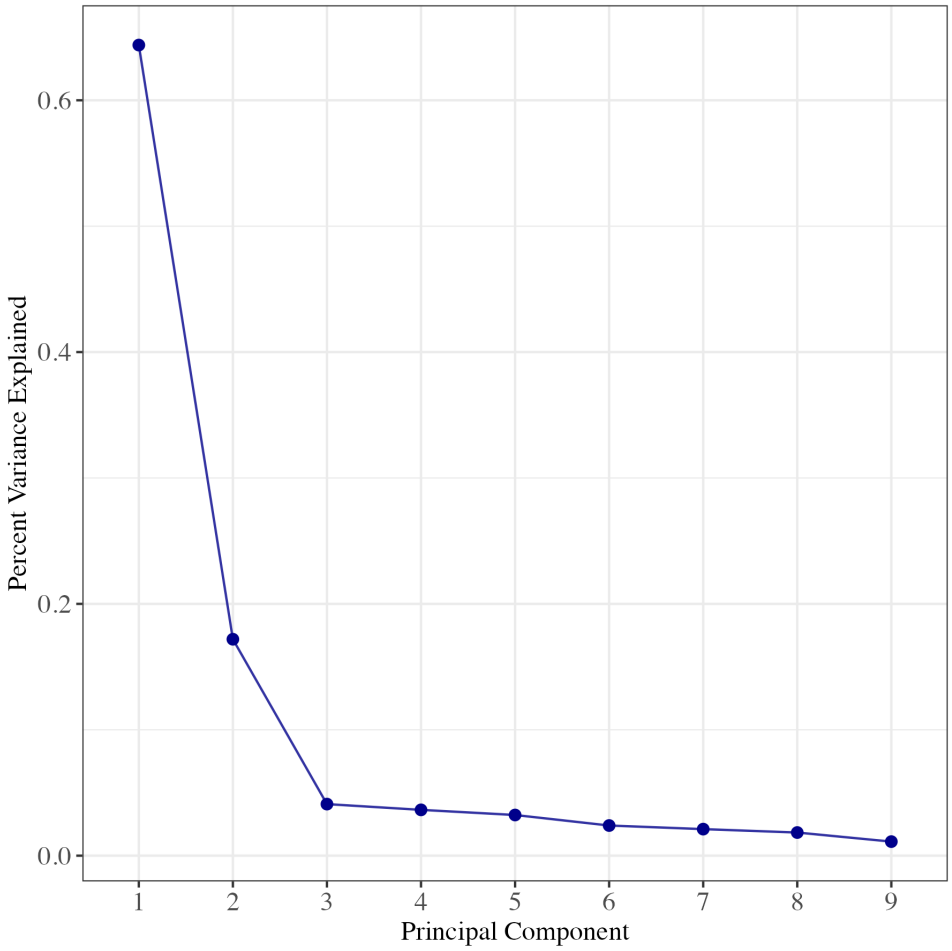
Principal Component Analysis (PCA) is an unsupervised machine learning technique that reduces high-dimensional data into a smaller set of component parts. PCA is an ideal choice for understanding relationships between variables when there is a high degree of collinearity and when the variance in the data can be well explained by a small number of dimensions, especially when the number of observations is relatively small compared to the number of variables. PCA does not predict the response variable; it estimates the relationships between a participant's evaluation of each rating metric. Because PCA can only estimate directions, the researcher must interpret what the resulting dimensions mean.

Principal components are rank ordered based on the direction of the data that explain the most variance. The first principal component accounts for the greatest amount of variance, the second principal component accounts for the second greatest amount of variance, and so on. The total number of principal components that can be estimated is equal to the number of variables.

The amount of variance explained by the principal components is relative to how many principal components have been estimated. For example, in Experiment 1, it is apparent that the first principal component explains a large amount of the variance (64.49%) out of a possible nine components. The second principal component explains 17.20% of the variance. Subsequently, the third to tenth principal components each explain a small amount of variance in the data (see Figure 2.1). I considered the first and second principal components only when clustering the ten rating metrics into indexical fields. In Chapters 3, 4, and 5, I use the {ggfortify} package in R to present data visualizations of the principal component relationships (Tang et al., 2016).

Figure 2.1

Variance Explained by Each Principal Component in Experiment 1's Data.



Theoretical Motivation for PCA

Because the main aim of this study is not to evaluate "intelligence" per se but rather the broader indexical associations associated with the *-ing* variable, it is beneficial to collapse the rating metrics into meaningful groups rather than nine interrelated categories, thereby increasing the generalizability of the results. In fact, the use of PCA to aggregate response variables that were perceived by listeners as equivalent decreases the likelihood of finding false positive effects; if "trustworthiness" and "appropriateness" both represent an underlying judgement of credibility but "trustworthiness" happens to appear statistically meaningful and "appropriateness"

does not, this could lead to the incorrect interpretation that variation in *-ing* usage affects listeners' evaluations of credibility.

Studies in sociolinguistic perception that have found collinear relationships between response variables have also used dimension reduction techniques to create a smaller set of evaluation categories. For example, Levon et al. (2021) completed a sociolinguistic perception study where listeners assessed speakers' performances in a job interview on five rating metrics (overall response quality, expertise, likelihood to succeed, personal likeability, and overall rating) and collapsed these metrics into "hirability" based on a finding that all five rating metrics had a correlation >.96. Correlation is another way to identify when participants essentially consider a rating metric to mean "the same thing," but PCA is more sophisticated because the relation between variables can be examined in multiple different directions.

Bayesian Regression

In this research, I chose to use Bayesian regression models to statistically analyze the effects of interest in the studies. Bayesian regression differs from frequentist mixed-effects regression models in that it samples from the data and prior distributions to estimate posterior distributions of the data (Barreda and Silbert, 2023). Bayesian regression is especially useful for studies in the social sciences where there is typically wide variation between participants' social evaluations. It is also particularly useful for studies that have repeated measures, such as the present research, where multiple samples are taken from the same participant. Further, the Bayesian approach places moves away from the concept of "significance" in frequentist methods, where statistically significant effects are considered sufficient evidence to reject the null hypothesis. Instead, uncertainty is central to the Bayesian analysis framework by design.

In Bayesian analysis, whether an effect is "real" is not determined by reference to a

particular numerical value but instead is evaluated against a region of practical equivalence (ROPE) (Krushcke, 2018). Schwaferts and Augustin (2020) argue that both a benefit and drawback of evaluating statistical reliability based on ROPE is that because there is no predetermined cutoff point for statistical significance, decisions based on ROPE are determined by the researcher. Bayesian ROPE recontextualizes the concept of "true positives" as evidence that a hypothesis is likely to be true. It is both the responsibility of the researcher and the reader to carefully and critically consider whether the argument put forth by the statistical analysis is sufficiently strong evidence for whether a hypothesis is considered true. Determinations of the strength of an argument based on statistical tests should be evaluated against real-world knowledge of the topic of study and compared against relevant work in the field that concurs or is at odds with the proposed findings.

However, an unfamiliar reader might refer to confidence intervals as one index of the reliability of an estimated effect, analogous to *p* values in frequentist statistics. An effect which has a positive or negative value within both the upper and lower confidence intervals (l-CI and u-CI, respectively) suggests a 95% probability that the effect is not zero.

In all the models in this dissertation, I used the {brms} package in R to model the data (Bürkner, 2017).

Beta and Zero-One Inflated (ZOIB) Regression

Part of statistical model development requires the choice of the distribution on which the data will be modeled. For all the studies in this dissertation, I chose to use a zero-one inflated beta (ZOIB) distribution to model the data.

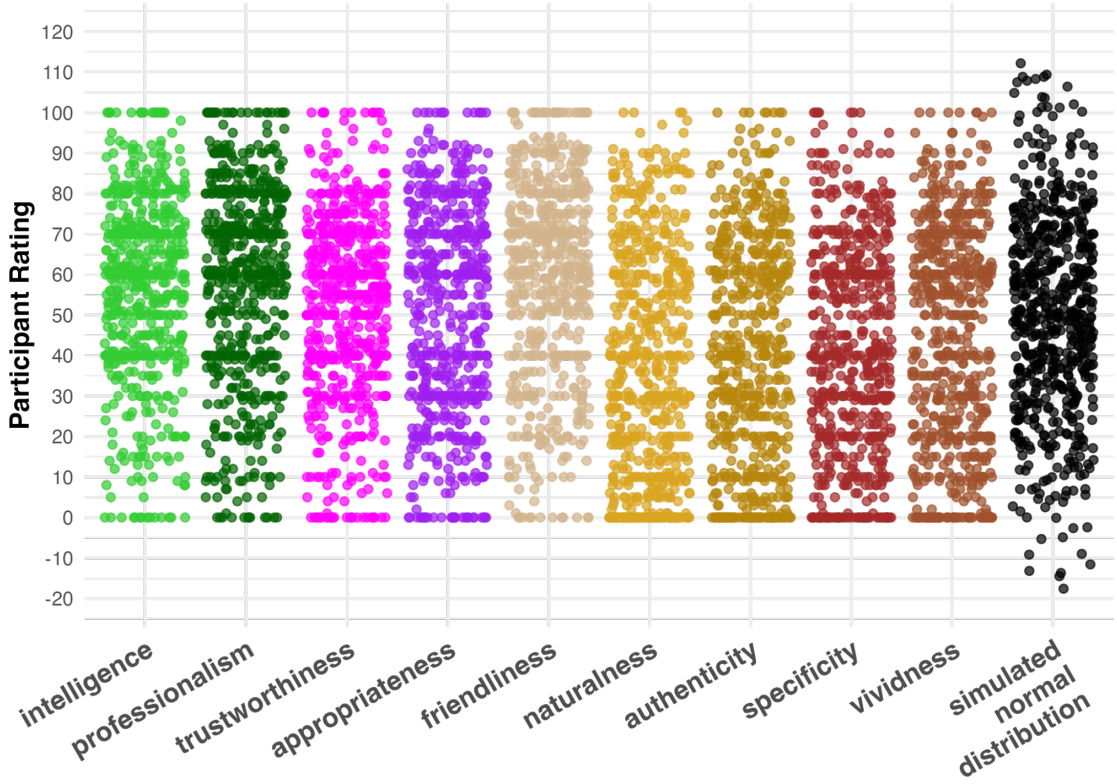
ZOIB regression is an appropriate choice for sliding-scale data because participant responses are unlikely to be normally distributed for two reasons: first, it is impossible to have

values less than zero and greater than one hundred; second, participant responses to sliding scales tend to be skewed (Vuorre, 2019). Figure 2.2 illustrates participant ratings from Experiment 2 compared against a normal distribution with the same standard deviation as the collected data. There are two clear issues with modeling the data with linear regression: First, a normal distribution estimates values above 100 and below 0, which are impossible on a sliding scale. Second, participants' rating behavior is clearly not normally distributed. There are more responses at the boundaries (0 and 100) than would be expected, and it appears that many participants evaluated the voices using ten-point increments as benchmarks. This is illustrated on the graph as the unusual clustering of points at 40, 50, 60, etc. A similar distribution is found for the participant ratings in Experiment 1 and Experiment 3.

Instead, the beta distribution is an appropriate choice when modeling proportion data. The beta distribution models two parameters, mu and phi. The mu parameter models the mean of the distribution estimated from the data, while phi is the precision, which can be described as how closely the distribution is clustered around the mean (Murphy, n.d.; Ospina & Ferrari, 2012; Robinson, 2014). However, the beta distribution can only estimate values that are within the open [0, 1] interval; in other words, if a participant gave a perfect 0 or perfect 100 rating on a particular rating metric, these kinds of values cannot be directly estimated by beta regression.

Figure 2.2

Actual Response Data from Experiment 2 Compared to a Simulated Normal Distribution



There are two general approaches taken to this problem. First, the researcher can adjust 0 and 1 values to be very close to the boundary, such as 0.0001 or 0.9999. Heiss (2021) illustrates that this approach can be acceptable in cases where very few boundary values are present in the data; in his example, less than 2% of the total values are zero. However, in the present study, it is evident that boundary values, particularly zero, are abundant in the data as illustrated in Figure 2.2 above.

Due to the observed shape of my data, I chose Bayesian zero-one inflated beta (ZOIB) regression modeling in the dissertation studies. ZOIB regression is a mixture model that is designed to model proportion data that has response values at the (0, 1) boundary. In addition to the beta distribution model, ZOIB regression includes two additional parameters, alpha and gamma (Heiss, 2021). The alpha parameter uses a logistic regression model to estimate the

likelihood that participants gave a response that was at the 0 or 1 boundary. In other words, if participants rated a voice as truly zero or one hundred on a particular rating metric, that probability is estimated by the alpha parameter. Then, the gamma parameter is a logistic regression model that estimates the likelihood that a response was at the 1 boundary, or a participant rating of truly 100 (Vuorre, 2019). Thus, the ZOIB model accounts for the 0s and 1s in the data without "nudging" the data to be within the open (0,1) interval.

Cross Validation and Competing Model Selection

Competing model selection is a method which compares different model formulas against each other and informs the researcher's decision on which model is most appropriate for the explanation of their data (Barreda & Silbert, 2023). This is accomplished by calculating the expected log pointwise predictive density of a given model and comparing it to other possible models. In essence, expected log pointwise predictive density weighs the variance explained by a model (how closely the model predictions match the real data) against model complexity (how many parameters are estimated in the model). Model complexity is penalized, where models that include more parameter estimates are assigned a penalty. Conversely, models that explain less variance in the data compared to models with more parameter estimates are penalized. This approach creates a push-pull between model complexity and model prediction accuracy to select a model that maximizes performance on in-sample and out-of-sample data.

Cross-validation is accomplished by holding out some data from the model for fitting and then refitting the same model on the held-out data. The model that most accurately fits the data with the least number of parameters is considered the "best" model. However, cross-validation should not be the only reason that a model is selected over another (Barreda & Silbert, 2023). At times, a more complex model is beneficial because it better answers the research questions than a

less complex one. For example, it is beneficial to estimate an "unnecessary" parameter if it is informative to answering the research questions. In Experiment 1, for instance, I find that the effect on whether a voice is a human or computer is not predictive of listener evaluation of voices. However, this is a central research question of the study, and so I include this parameter in the model. Deciding which model is best for interpreting data should include reference to the predictive density but ultimately be informed by the research questions.

In each of the analyses, I used k-fold cross validation to assess different model formulas compared to other possible models. The models selected for use in the study are very close to what is considered the "best" model in terms of posterior predictive density, but with some exceptions. In Experiment 1, participant voice AI use does not increase the model explanation of variance and introduces an additional parameter, but I chose to include it in the final model because it informs what factors influence participant rating behavior based on their individual AI use. Understanding an individual's technology use frequency is a relevant area of interest in human-computer interaction studies, as described in Chapter 1.

One area where predictive density was particularly informative to my model building was in the selection of terms to include in the alpha and gamma parameters within the zero-one inflated beta (ZOIB) regression models. For data that has relatively few values at the boundary (0 or 100), estimation of multiple parameters for alpha and gamma might not be appropriate. In this dissertation study, I find through k-fold cross validation that the model parameters included in the alpha and gamma estimates are important to improving the predictive accuracy of the model. This fits with my prior understanding of the shape of the data, as illustrated above in Figure 2.2.

In the next three chapters, I explore three sociolinguistic evaluation studies using the methods described above.

Chapter 3: Exploring Variation in Sociolinguistic Evaluation Across Human and Computer Talkers

Abstract

Sociolinguistic variation is ubiquitous in human speech and is recognized by talkers and listeners alike to communicate social positioning to other interlocutors. However, little research has investigated whether listeners would perceive sociolinguistic variation in computer voices similarly to how they do for human voices. "Social" computers are becoming widespread in today's social landscape, and they are designed to exhibit many characteristics of human social communication in their speech, such as politeness and gendered language. Do listeners evaluate these sociolinguistic features as socially meaningful when used by a computer talker? The current study investigates this question by using the (ING) variable, a widespread variant in American English, as a case study of sociophonetic variation. Participants listened to a sample news report read by two human and two synthetic voices that varied in their (ING) usage rate, then provided social evaluations on rating metrics of status, appropriateness for news broadcasting, trustworthiness, and humanlikeness. I found that listeners transferred their social-indexical perceptions of (ING) variation to computer voices, but to a limited extent compared to human voices. These findings have implications for both theories of social meaning and human-computer interaction.

Introduction

Socially meaningful language variation is present in every spoken utterance. From mere milliseconds of speech, listeners intuit information about a speaker's size, gender, and dialect and create complex social evaluations of that talker (Barreda, 2020; Drager, 2010). Many researchers have begun to ask how sociolinguistic perception works during human-computer interaction, such as when an individual is using a voice-enabled device like Amazon Alexa or

Apple Siri (Zellou & Holliday, 2024; Zellou et al., 2023). Do listeners apply the same social evaluations to computer talkers? Are nonstandard linguistic variants (such as use of *-in'* for present progressive in English, instead of standard *-ing*), in particular, judged similarly or differently in a human-computer interaction context, compared to when they are produced by human talkers?

Socially conditioned (ING) Variation

I focus on one phonological variant well-examined in the literature on socially conditioned linguistic variation in English: variable (ING). While (ING) variation occurs across many lexical categories, the most widespread use of (ING) variation appears in the production of present progressive verbs (e.g., *talking* pronounced as *talkin'*) (Houston, 1985). Variable (ING) use is widespread in many speech communities (Kiesling et al., 2009; Schleef & Flynn, 2015). Listeners who hear speech with higher prevalence of the *-in'* variant rate those talkers as being less intelligent, less professional, and poorer than those who use *-ing* more frequently (Campbell-Kibler, 2007; Fischer, 1958; Trudgill, 1974). This effect is gradient: Listeners' evaluations become increasingly negative with each additional instance of *-in'* used by the speaker (Labov et al., 2011). (ING) variation is also evaluated differently based on context. Campbell-Kibler (2009) found listener evaluations of *-in'* were comparatively less negative when the speaker was talking about hobbies. Thus, in some contexts, greater *-in'* usage results in more positive evaluations.

Language Variation in Human-Computer Interaction (HCI)

How might (ING) variation be evaluated in a relatively novel social context—when produced by computers? In today's world, searching the web, weather forecasts, and news reports might be commonly done via voice-enabled digital devices. Do people apply the same

social evaluations for humans and computers based on their speech and language patterns? More specifically, I ask: Do people apply the same social evaluations of nonstandard (ING) variation when they are hearing the news produced by a computer?

A series of studies conducted by Nass and colleagues investigated how users socially behave towards computers in the late 1990s and early 21st century, when personal computers were a novel invention. They found that users were polite toward computers they interacted with, applied gender stereotypes to female- and male-sounding voices, and used face-saving techniques when evaluating a computer's competence on tutoring topics (Nass et al., 1994; Nass et al., 1997). They proposed the computers as social actors (CASA) theory to explain users' social behavior with computers as the result of unconscious anthropomorphism, particularly when computers use speech (Nass et al., 1994). Nass and Lee (2001) found a similarity-attraction effect when users interacted with social computers that aligned with their personal degree of extroversion; in other words, people preferred to interact with computers that appeared to align with their personality characteristics. For example, computer voices that are more emotionally expressive are rated more positively by listeners (Cohn et al., 2021). Taken together, these results indicate that people apply human-based social expectations and judgements to computers. From a sociolinguistic perspective, the CASA theory is consistent with sociolinguistics findings that people automatically and unconsciously evaluate social aspects of speech just by hearing a voice (e.g., Niedzelski, 1999). We might predict, then, that listeners will evaluate variable (ING) usage in a computer voice in the same way they evaluate variation produced by a human talker.

On the other hand, more recent research on human-computer interaction has proposed the concept of routinized social scripts as the underlying cause for what appears to be socially directed speech. This was observed in a recent study on perceptual adaptation to accented speech

(Zellou et al., 2023) which showed that people were more likely to generalize a novel accent to a new talker's voice when the novel talker matched the exposure talker's *humanness*; in other words, people behaved as if they expected an accent to be socially constrained to either human or machine speaker categories. So, another possibility is that the social evaluation of (ING) variation in computer speech will not be evaluated with the socially meaningful indexical associations with (ING) from the human speaker category onto the machine speaker category.

Current Study

In the current study, I use the widely studied (ING) variant in English to investigate whether listeners perceive nonstandard language use in computer voices as socially meaningful in the same ways as they do for human speakers. I played listeners a short news report that contained ten present-progressive verbs, produced either naturally by human talkers or generated using state-of-the-art synthetic speech. Listeners heard the same news report read by four different voices: natural productions by two American English-speaking women, and two female synthetic voices generated via text-to-speech (TTS). Each listener heard each speaker produce the passage once and across speakers there were variable rates of *-ing*. I varied the clips to contain different rates of the non-standard *-in'* variant from 0% to 100% of possible occurrences and counterbalanced speaker assignment to *-ing* rate across listeners. Listeners rated each speaker on several attitudinal dimensions (professionalism, intelligence, appropriateness for news broadcasting, trustworthiness, authenticity, naturalness, vividness, specificity, and perceived age).

Prior to beginning the experiment, I established a central question and several hypotheses. How will these ratings vary based on gradient *-ing* usage across human and device talkers? A finding that listeners evaluate nonstandard variation in computer speech similarly to

their evaluations of human voices would align with the CASA theory that users equivalently apply human social norms to computers. Another hypothesis was that listeners would socially evaluate the use of nonstandard variation in synthetic speech, but in a more limited way than in human speech. This would be in line with the "social scripts" account of HCI. My results speak to how speech and language use from computers is socially evaluated by people. Now that machines are regular actors in our speech communities, understanding the social evaluation of their language use is an important theoretical and practical issue.

Methods

Approach and Experimental Design

I use the matched-guide paradigm to test listeners' implicit associations of sociolinguistic variation (Baugh, 2003). For each of the voices in the study (human and synthetic), I created matched guises of the stimuli with four different rates of (ING) variation. Participants listened to each voice use a different rate of *-ing* and then provided social evaluations of each talker. This allows for the isolation of sociolinguistic variant use as the conditioning factor in listener evaluations, rather than speaker-specific voice qualities. This technique is effective in eliciting participant evaluations of subtle linguistic differences, including the *-ing* variant (Campbell-Kibler 2007, 2009; Vaughn, 2022).

In my study, participants listened to a passage spoken by two synthetically generated and two naturally recorded human voices. Participants listened to the news broadcast passage taken from Labov et al. (2011), which describes the Congressional debate of the EGTRRA Bush administration tax cuts proposed in 2001 (GovInfo, 2001). The one-minute passage contains ten different present progressive verbs. All voices in the current study were female.

There were four *-ing* use conditions: all *-ing* use (100% *-ing*), mostly *-ing* use (70% *-ing*), rarely *-ing* use (30% *-ing*), and no *-ing* use (0% *-ing*). In the 70% and 30% *-ing* conditions, the opposite variant always appears at the 3rd, 6th, and 9th instance to control for primacy and recency effects. For instance, in the 30% *-ing* condition, all present progressive verbs used the nonstandard *-in'* variant except for the 3rd, 6th, and 9th present progressive verbs, which used the *-ing* variant.

To mitigate participant fatigue from listening to the same passage multiple times, I used a counterbalanced design to evenly collect observations of each level of *-ing* use in both voice types. Each participant listened to the passage a total of four times. I blocked the synthetic and human voices to avoid cueing participants to directly compare human versus non-human voices. Participants were randomly assigned to one of eight different experimental lists: Half heard the synthetic voices first block, half heard the human voices first block. Within the voice-type block, one voice used more "standard" speech (either 100% or 70% *-ing*), and one voice used more "nonstandard" speech (either 30% or 0% *-ing*). The rate of *-ing* use was counterbalanced across blocks so that participants heard each *-ing* condition exactly once.

Materials

Two native American English speakers (both female) read the passage in a sound-attenuated booth. Recordings were digitized at 44.1 kHz and denoised using Praat Vocal Toolkit (Corretge, 2023). I had each speaker record the passage twice, once using only the standard variant and once with only the nonstandard variant. In order to ensure that the only difference in the standard and nonstandard passages was the variant itself, I cross-spliced a speaker's *-in'*/*-ing* segments into the body passage. To generate the recording of the passage with different (ING) rates, the segment of interest was spliced from the recorded passages and replaced into the body

passage. Both the standard and nonstandard variants are cross-spliced into the body passage to control for a possible effect of cross-splicing unnaturalness, as pointed out by Vaughn (2023).

Because the *-in'* variant is frequently shorter in duration compared to *-ing*, the difference in duration between the *-in'* and *-ing* segments was averaged for each word (Campbell-Kibler, 2009). When the duration of the *-in'* and *-ing* segments differed, I adjusted the durations of each segment in Praat so that the standard and nonstandard variant for each word would be of equivalent lengths. When a segment needed to be lengthened, cycles were copied then pasted adjacent to the copied cycle. Cycles were added or deleted from the (ING) variable at different time points throughout the segment to ensure the cross-spliced segments were uniform. Care was taken to paste segments at the nearest zero crossing to avoid clicks. (ING) variant segments for each word were matched in duration within .01 ms of each other. Each segment was amplitude normalized to the surrounding word.

To create the computer voice stimuli, I used two female TTS voices from the Amazon Web Services (AWS) Polly voices ("Kendra" and "Kimberly"). The AWS "Alexa" voice was not included in this study so that there would not be an effect of voice familiarity on participant ratings. Neural speech synthesis was chosen for use in this study because it produces more naturalistic synthesis of nonstandard pronunciations (such as *-in'*) and has better intersegmental continuity than concatenative synthesis methods (Cohn & Zellou, 2020; Kuligowska et al., 2018). I synthesized the passage with the *-ing* variant for each voice and generated the same passage with the *-in'* variant, which resulted in a more naturalistic result than cross-splicing. For both human and synthetic voices, all stimuli were amplitude-normalized to 65 dB.

Participants and Procedure

Participants (n = 191; mean age 20; 132 female, 57 male, 2 nonbinary) were UC Davis undergraduates and were compensated with partial course credit for their participation. All participants reported that they spoke English as their strongest language and had no hearing impairments. Before the experiment, participants completed a short word identification task to ensure they could hear and understand audio.

Participants were told that I was developing a news broadcasting app and would give feedback on which voice should be selected for use in the app. They were instructed to rate four voices on nine different rating metrics: professionalism, intelligence, appropriateness for news broadcasting, trustworthiness, friendliness, authenticity, naturalness, how vividly the participant could imagine the person speaking, and how much the participant felt the voice was speaking to them specifically.

Participants listened to each news broadcast recording one at a time. Following each recording, participants used a sliding scale from 0–100 to respond to each of the rating questions. There was an optional comment box where participants could write evaluations they had about a particular voice on the same screen as the nine sliding-scale evaluation questions. Participants completed this procedure for each voice.

After completing the evaluation phase of the experiment, participants heard each voice repeat a short clip of the passage and then selected whether they thought the voice was a human or computer. Lastly, participants responded whether they thought computer or human voices were better for news broadcasting in general.

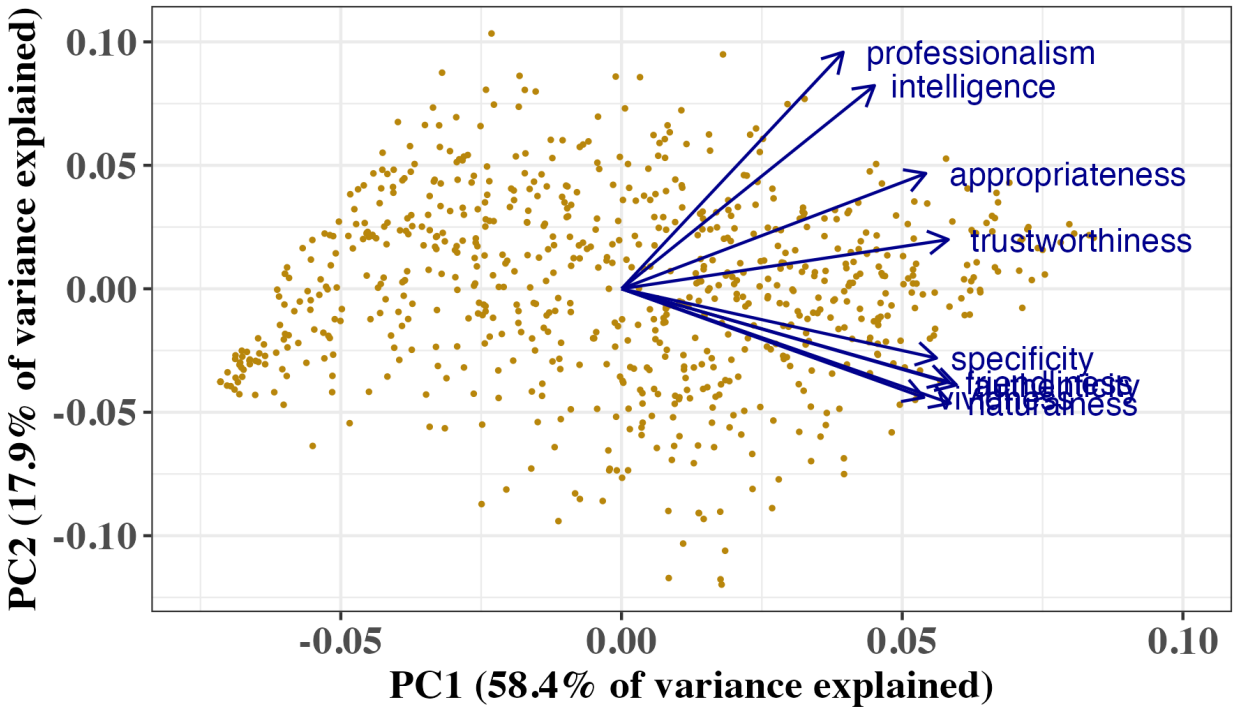
Results

Cognitive Organization of Social Evaluations

Because listener evaluations of sociolinguistic variation can be highly covariant, I investigated whether the nine rating metrics I collected could be reduced into a smaller set of dimensions. I used unguided principal component analysis (PCA) to evaluate whether reducing the rating metrics to a smaller set was appropriate based on the data I collected. I found that the first two principal components account for 58.4% of the variance in the data, suggesting that dimension reduction is appropriate.

Figure 3.1

Principal Component Analysis of the Nine Sliding-Scale Rating Metrics Presented to Participants



Note. Filled points represent each observation in the data, rotated to the shortest distance between points based on principal component 1 and principal component 2. Blue arrows represent how covariant rating metrics are with other metrics.

Figure 3.1 displays the eigenvector loadings and directions of principal components 1 and 2. The eigenvector directions reflect how the social characteristics are differentiated from each

other. The distances between the eigenvector loadings on the y axis indicate that some of the social evaluations covary with each other. In particular, professionalism and intelligence covary. Friendliness, authenticity, naturalness, vividness, and specificity show a high degree of covariation. The covariance explained by PC2 is consistent with previous research findings that intelligence and professionalism stem from the same indexical field in listener evaluations of *-ing* variation (Campbell-Kibler, 2009; Labov et al., 2011). The covariance between naturalness, authenticity, vividness, and specificity is consistent with previous findings from human-computer interaction research relating these evaluation categories (Nass & Lee 2001; Nass et al., 1994; Kim & Sundar, 2012). Based on the results of PCA, I collapsed the nine rating metrics into four indexical fields, provided in Table 3.1.

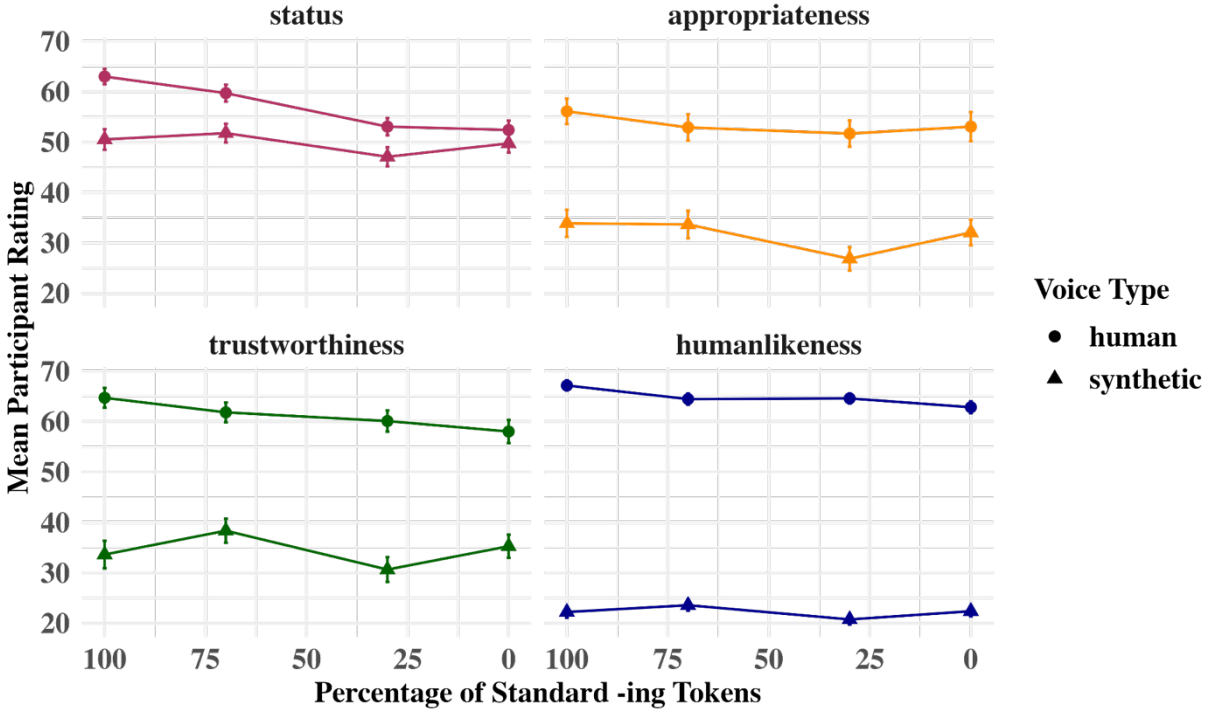
Table 3.1
The Nine Rating Metrics Collapsed into Four Rating Categories Based on Their Degree of Covariance

Rating Metric	Reduced Dimension
Professionalism, Intelligence	Social Status
Appropriateness	Appropriateness
Trustworthiness	Trustworthiness
Naturalness, Authenticity, Vividness, Specificity, Friendliness	Humanlikeness

Figure 3.2 illustrates the distribution of participant ratings based on which voice block (human or synthetic) they heard first. Figure 3.2 suggests that synthetic and human voice types are evaluated differently by listeners across evaluation categories.

Figure 3.2

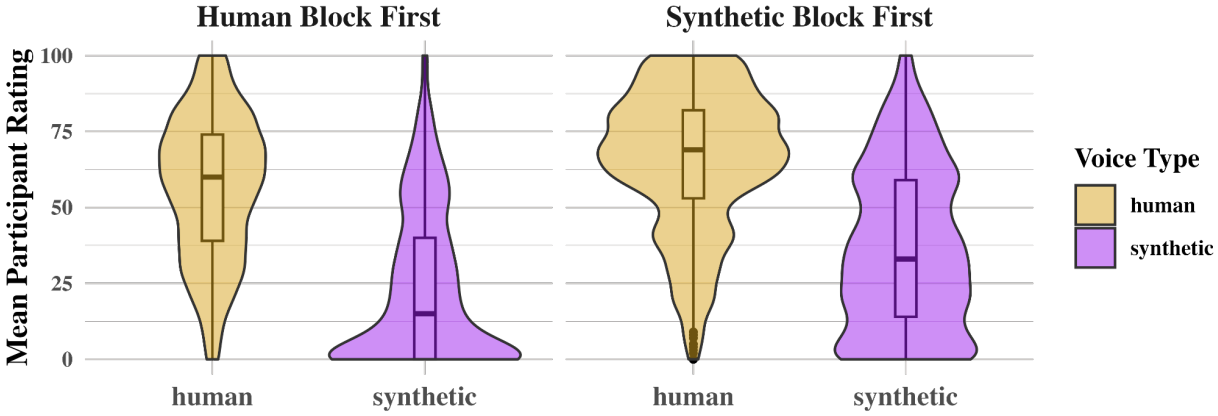
Mean Participant Ratings for Each Evaluation Category by -ing Rate for Each Voice Type



But, which voice block participants heard first influenced how they rated voices in the second block, as shown in Figure 3.3. When participants heard the human block first, they evaluated synthetic voices lower compared to when participants heard the synthetic block first. The distribution of participant ratings of the second block is also skewed; listeners became more likely to use all-or-nothing rating behavior in their evaluations of the voices in the second block. This effect varied based on the specific evaluation being made, as shown in Figure 3.4.

Figure 3.3

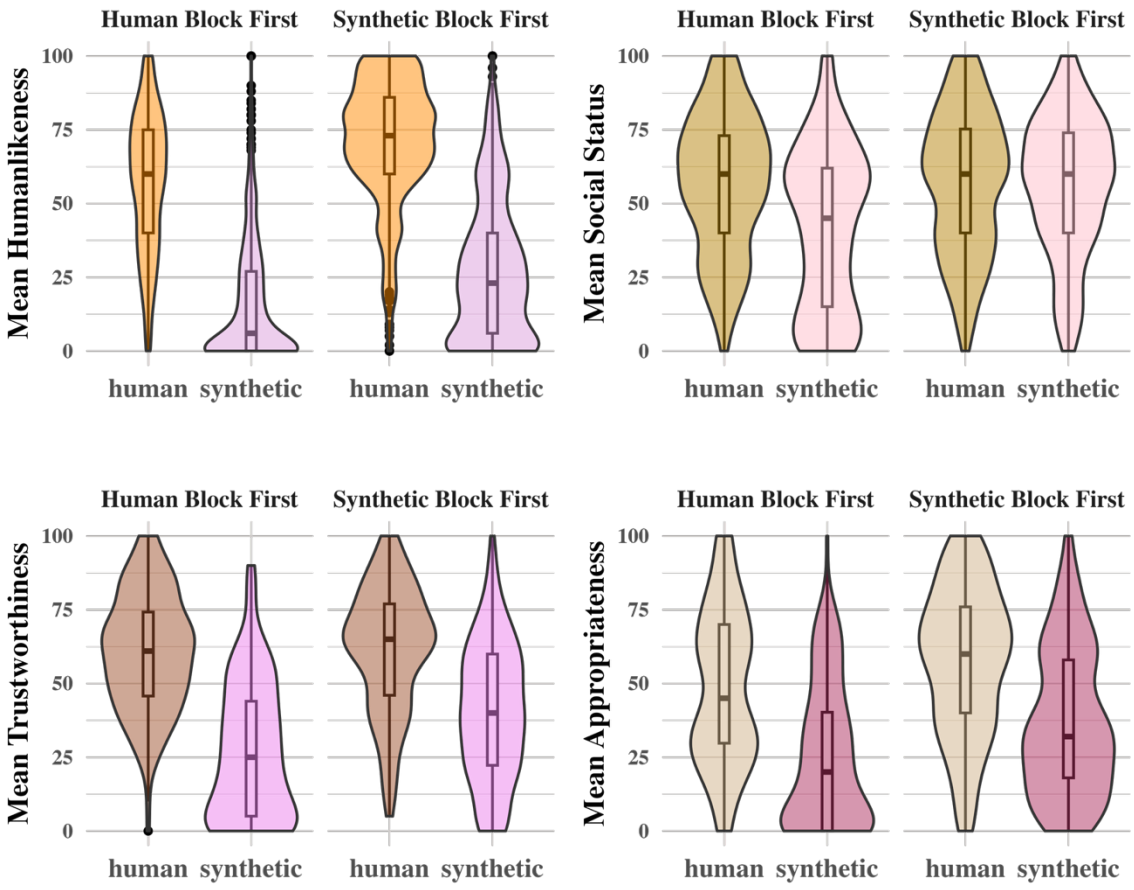
Total Participant Ratings Distribution Subset by Block Order



Note. Filled points are outliers.

Figure 3.4

Participant Ratings Distributions Subset by Block Order for Each Evaluation Category



Note. Filled points are outliers.

Bayesian Regression of Effects on Listener Social Evaluations

I used a Bayesian zero-one inflated beta (ZOIB) regression model to estimate the effects of (ING) variation, voice type, the talker (the four voices used in the stimuli), first block (human or synthetic), and the interaction of voice type and *-ing* use rate. I also modeled the effects of participant gender and participant's reported voice-AI usage frequency (either more than weekly or less than weekly). I included a random effect by-subject. Mu, alpha, and gamma are all estimated using the same formula. The phi parameter is estimated only as the intercept without any additional parameter estimates to maximize interpretability of the model. Each of the evaluation categories was fit to separate models with the following model syntax:

participant rating ~ voice type*ing use + voice AI use + voice type*first block + participant gender + voice + (voice type | id)

Generalized Effects on Listener Evaluations

Across the four models, I find some commonalities. Female participants rated voices lower than male participants in every evaluation category. I also find that some effects did not have an effect on any of the evaluation categories. The model estimated with confidence that participants' voice AI use frequency did not significantly affect their evaluations. The model estimated with low confidence the effect of individual talker, which suggests that the particular voice participants heard was not reliably influential on their evaluations. The model also estimated with low confidence the effect of voice type for all rating categories, suggesting that whether a voice was human or synthetic did not reliably influence participant evaluations. Instead, the effect of human-first block order is estimated with confidence to influence participant ratings for every rating metric. In other words, participants rated both human and synthetic voices lower when they heard human voices first compared to participants who heard

synthetic voices first. In the discussion I explore why human-first block order might have been more predictive of participant evaluations than human voice itself.

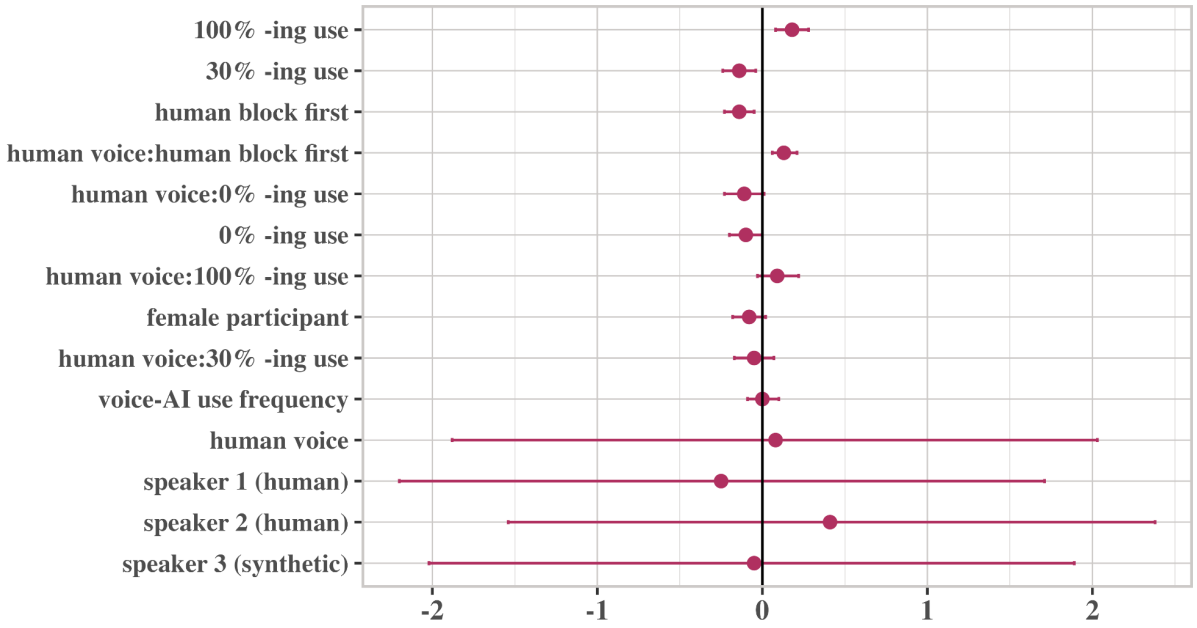
Evaluation Category-Specific Effects on Listener Evaluations

In the Social Status model, there was an effect of all levels of (ING) variation on participant evaluations. As a speaker's use of the nonstandard -*in'* variant increases, participant evaluations of social status decrease. I also observe an interaction between (ING) use and voice type: When human speakers used 100% the standard -*ing* variant, they were evaluated more positively than synthetic voices that used 100% -*ing*. The reverse is also true: When human voices used 0% -*ing*, this had a stronger negative influence compared to when synthetic voices used 0% -*ing*. For this interaction effect, it appears that the interaction between intermediate levels of nonstandard -in' use (30% and 70% -*ing* conditions) and voice type did not influence participant ratings.

Beyond differences in (ING) variation, I also find an interaction between human voice and human-first block order, indicating that participants in the human-first condition evaluated human voices lower on social status lower compared to participants in the synthetic-first block. Figure 3.5 plots the estimated effects and interactions in the social status model.

Figure 3.5

Plots of All Estimated Effects and Their Interactions for Social Status



Note. Effects above the bold horizontal line at zero had a positive effect on listener evaluations, while effects below the zero line had a negative effect on listener evaluations.

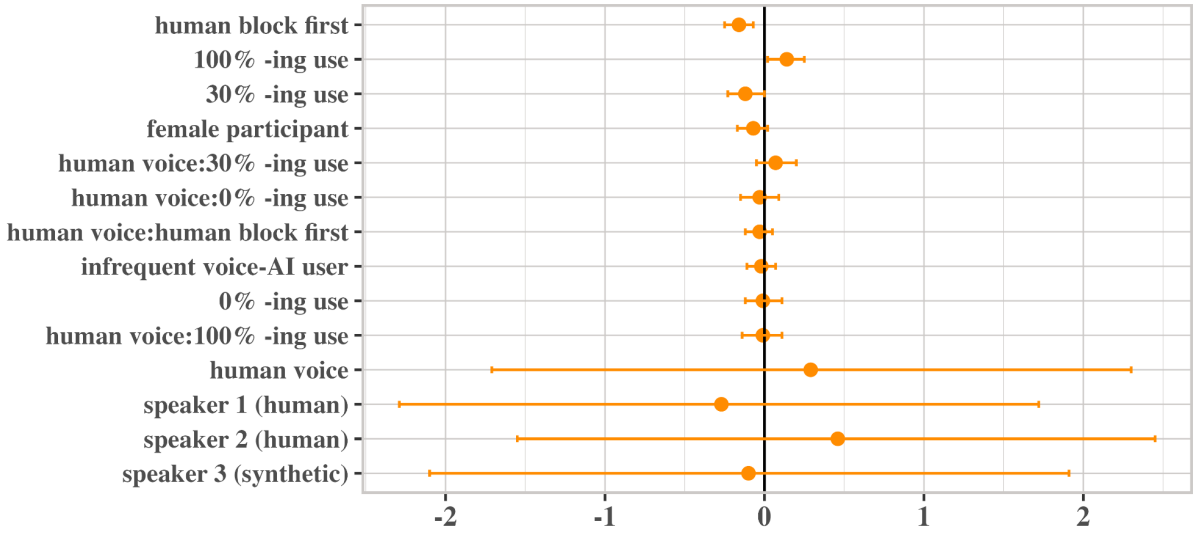
For appropriateness, there was a main effect of some levels of (ING) variation on participant ratings, but not others. Only 100% *-ing* use and 30% *-ing* use reliably influenced listener evaluations. 100% *-ing* use was perceived as more appropriate for news broadcasting for both voice types, while an increased amount of nonstandard *-in'* lowered listeners' evaluations of appropriateness. However, 0% *-ing* use is estimated to not have a negative effect on listener evaluations; in other words, 30% *-ing* use was penalized more than 0% *-ing*.

Unlike listener evaluations of social status, there is no apparent interaction between human-first block order and human voice. An interaction between human voice and (ING) variation for appropriateness appears minimal; there might be an interaction between human voice and 30% *-ing* use, where synthetic voices are rated even lower when they use 30% *-ing* compared to the other (ING) conditions. Moreover, (ING) variation was evaluated similarly by listeners regardless of whether the voice was human or computer, but for both voice types (ING)

variation influenced participant ratings. Figure 3.6 plots the estimated effects and interactions in the appropriateness model.

Figure 3.6

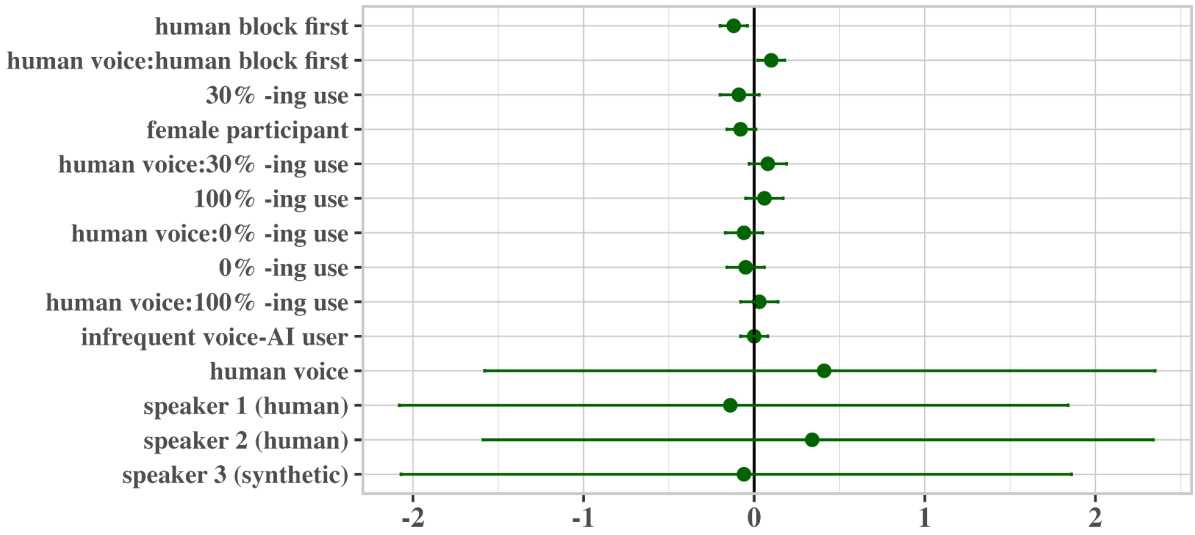
Plots of All Estimated Effects and Their Interactions for Appropriateness



For Trustworthiness, there is no clear effect of (ING) use or the interaction between (ING) and voice type on participant ratings. Compared to the other two social evaluation categories, estimations of (ING) use are closer to zero. 30% *-ing* use and the interaction between 30% *-ing* use and human voice might be influential on participant ratings of trustworthiness, but this effect appears small. Instead, evaluations of trustworthiness appear to be more influenced by whether a voice is human. Human voices were rated as more trustworthy when participants had heard synthetic voices first, and the interaction of human voice and human-first block resulted in an even bigger boost in how trustworthy participants considered the human voices. Figure 3.7 plots the estimated effects and interactions in the trustworthiness model.

Figure 3.7

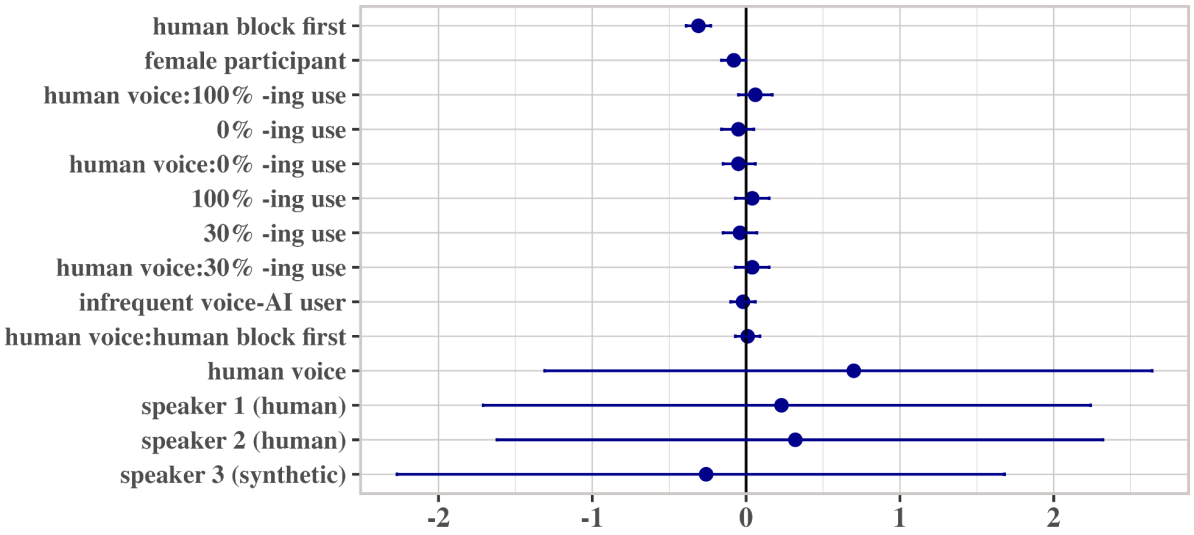
Plots of All Estimated Effects and Their Interactions for Trustworthiness



In the humanlikeness model, I find that (ING) use and the interaction of (ING) and human voice does not meaningfully influence participants' ratings. The effect of human-first block was reliably the most predictive of how listeners evaluated the voices. Interestingly, the interaction of human voice and human-first block did not change how humanlike participants considered the voices. Although human voice is estimated with very low confidence, the humanlikeness model has the largest estimated effect of human voice of the four evaluation categories. Figure 3.8 plots the effects and interactions estimated by the humanlikeness model.

Figure 3.8

Plots of all Estimated Effects and Their Interactions for Humanlikeness



Discussion

The current study explored whether listeners judge variable (ING) usage similarly for a computer versus a human talker producing a news report. Consistent with prior work, I find listeners rated voices that produced standard *-ing* as having higher social status and as more appropriate for reading the news (Labov et al., 2011; Campbell-Kibler, 2009). When comparing differences across humans and computers within these two rating categories, findings were somewhat mixed. Listeners perceived voices that used exclusively the *-ing* variant as most appropriate for news broadcasting and majority nonstandard use as less appropriate, regardless of whether the voice was human or synthetic. This suggests there is direct transfer of appropriateness evaluations from human-human interaction to computers based on their social-indexical language use patterns.

For social status, (ING) use influenced listener perceptions for every rate of (ING), where each nonstandard *-in'* token gradiently decreased listeners' evaluations of the talker, similar to findings in human speech (Labov et al., 2011). But (ING) variation had an even greater

effect on listener evaluations of human voices compared to computer voices. This suggests listeners do transfer the social meaning of (ING) variation to computer voices, but to a lesser extent compared to their perception of human voices. Taken together, then, the evidence of sociolinguistic transfer varies based on the type of social judgement being made: Linguistically mediated appropriateness judgements are the same across human and computer voices, while social status evaluations are influenced by sociolinguistic variation more when used by a human voice.

A novel finding is that (ING) variation did not affect participant evaluations of humanlikeness. The large difference in mean participant ratings of humanlikeness suggests that they identified human voices as socially different from computer voices. Yet, the use of a social-indexical trait—the *-in'* variant—did not further influence how natural they considered computer voices. I also found that participants evaluated voices as more trustworthy when they were human, regardless of any sociolinguistic speech differences.

An unexpected result is that I do not find a simple effect of voice type for any of the rating metrics. Instead, there is an effect of block ordering on social evaluations. Because participants were not informed that they would hear both synthetic and human voices in the experiment, they rated human voices relatively higher once they discovered that they were being compared against synthetic ones. Likewise, participants in the human-first voice condition rated the second block (synthetic voices) relatively lower. For some evaluation categories (Social Status and Trustworthiness), I find an interaction between human-first block and human voices. When participants were asked whether they thought human voices were better for reading the news at the end of the experiment, nearly all (98.4%) participants responded that they preferred human voices. This is consistent with the finding that listeners bring their script-specific

perceptions of computers' social potential to interactions with new computer talkers (Gambino & Liu, 2022).

Another interesting contribution of the current study is that participant voice AI use did not impact participants' evaluations of voices. While we might expect that listeners' evaluations are influenced by how often they themselves interact with synthetic voices, it appears that participants' overall beliefs about the social role of computers is more influential on their rating behavior.

My findings on voice effects on trustworthiness and humanlikeness are relevant to theories of human-computer interaction. The connection between humanness and trust has been widely studied in the field of human computer interaction, and findings are mixed on the "trustworthiness" of machines. On the one hand, studies on uncanniness in computer voices have found that listeners react with feelings of distrust and disgust when they interact with a computer that is imitating human social traits, including sociolinguistic variability (Clark et al., 2019; Tinwell, 2013). Other work has found that computers are considered more trustworthy (Sundar & Kim, 2019), suggesting that trust in computers is context specific. In the study by Sundar and Kim (2019), participants were asked to give credit card information to chatbots, while in the current study chatbots gave information to listeners. Whether the computer is giving or receiving the information may be related to how participants evaluate trust in machines.

From a sociolinguistics perspective, the findings of the current study support theories that suggest that social context is an important influence on listeners' social evaluations (e.g., Campbell-Kibler, 2009; Drager, 2010) and that linguistic variation is a relevant social factor in how people interact with and behave toward computers (Zellou & Holliday, 2024). In particular, the current study illustrated that listeners transfer the social meaning of (ING) variation to their

perceptions of computer voices. But the fact that listeners' evaluations were less influenced by (ING) variation when it was used by computer voices suggests that they consider it *less* socially meaningful than in human speech. Work on (ING) in human voices illustrates that listeners are less sensitive to (ING) variation when it is not stylistically congruent with the surrounding speech (Vaughn, 2022). In my study, I suggest the decreased sensitivity by listeners to (ING) variation is because sociolinguistic variation is stylistically incongruent with listeners' expectations for computer voices. Future work on other sociophonetic variants can further elucidate whether stylistic variation in computer voices is perceived by listeners as socially meaningful.

Chapter 4: Young Adults' Evaluations of Sociophonetic Variation in Synthetic Speech

Introduction

The main aim of Experiment 1 (Chapter 3) was to test how listeners socially evaluate synthetic voices that vary in their use of (ING) and directly compare their evaluations with human voices within-participant. Experiment 1 had the key finding that the connection between perceived intelligence and professionalism—the most common indexical associations with standard -ing—is extended by listeners to their evaluations of computer talkers as well. I also found an interaction effect between perception of (ING) variation and social status depending on whether the talker was a human or computer. Standard (ING) was considered more appropriate for news broadcasting for both human and computer voices, similar to Labov et al's (2011) findings for human voices. Moreover, I find that (ING) perception in my study was consistent with prior scholarly work and with the innovation that (ING) perception extends to non-human talkers. I found that nonstandard (ING) use did not influence participants' evaluations of humanlikeness. In other words, participants considered human voices more humanlike but did not consider computer voices even *less* humanlike if they used the nonstandard variant, a human social-indexical trait.

I also found that participants' evaluations of human and synthetic voices were primarily based on comparing one type against the other. Once participants realized the voices were being compared against each other, they adjusted their rating behavior in the subsequent block to either boost scores given to human voices or penalize synthetic voices, depending on which block they heard first. While Experiment 1 was able to directly compare human and synthetic voices in a controlled experiment, it was limited by the need for a blocked design and the inclusion of only female voices.

Current Study

The current study extends the knowledge gained in Experiment 1 by comparing human and synthetic voices as a baseline and allows me to extend that knowledge to a larger set of synthetic voices using more sophisticated voice synthesis technology. I included both stereotypically masculine and stereotypically feminine voices in the experiment to test for an effect of perceived speaker gender on listener evaluations of synthetic voices.

In the current study, Experiment 2, I ask how listeners would socially evaluate synthetic voices without direct comparison to human ones. This question is relevant for two reasons. First, I ask this question to assess how participants evaluate synthetic voices used in news broadcasting without applying their top-down knowledge that they prefer human voices overall, which was the primary influence on their rating behavior in Experiment 1, instead of the particular voice itself. Second, I ask how participant evaluations might be influenced if they incorrectly perceive a synthetic voice as human. Participants had near-ceiling performance on their ability to identify whether a voice was a human or computer in Experiment 1, so I selected newer, more naturalistic synthetic voices developed by Google WaveNet for use in the current study.

Like the previous study, I investigate if participant voice AI use frequency changes their evaluations of synthetic voices. While I did not find an effect of voice AI use in Experiment 1, in this experiment, I investigate whether participants' personal voice AI use habits will influence their evaluations when only synthetic voices are used in the study. I also investigate if frequent users of voice AI give different social evaluations to the more naturalistic voices in this study, and if differences in synthetic voices' apparent gender interacts with how often participants interact with voice AI.

I also investigate whether listeners' evaluations change based on their own gender identity because speaker gender and listener gender have been shown to interact (Eckert, 2008). Differences in (ING) usage perception by men and women have already been shown in studies on human voices (Kiesling & Coupland, 2009; Mechler, 2025). I might expect, then, that gender-based social perceptions transfer to synthetic voice evaluations, since Experiment 1 showed evidence that synthetic voices are socially evaluated. Could these social evaluations be finer grained to the extent that they apply to different apparent genders?

To create the most controlled experimental environment possible for optimal comparison to the findings between my studies, Experiment 2 follows a similar experimental design to Experiment 1 with new stimuli. Because experiment length and participant fatigue were a consideration in both the previous study and the current one, I elected to use four voices in this study.

Methods

Experiment Design

The present study follows the same general method as Experiment 1, where participants evaluated four voices on their reading of the same one-minute passage. Like Experiment 1, the current study used an adapted version of the matched guise technique to test listeners' social perceptions of (ING) variation while controlling for potential listener-specific differences in social evaluations. Rather than have participants listen to one synthetic voice four times producing varying rates of *-ing*, I generated four different synthetic voices available from Google WaveNet's neural text-to-speech voices. Two voices were feminine, and two voices were masculine. There are two benefits to this approach: First, it allows us to examine the interaction

of speaker gender and *-ing* use, and it increases the generalizability of our findings beyond one particular synthetic voice.

There were four *-ing* use conditions: exclusively standard *-ing* use (100% *-ing*), mostly standard *-ing* use (70% *-ing*), mostly nonstandard *-in'* use (30% *-ing*) and only nonstandard *-in'* use (0% *-ing*). In the 70% and 30% *-ing* conditions, the alternate pronunciation always occurred on the 3rd, 6th, and 9th present progressive verb in the passage to control for primacy and recency effects. For example, in the 30% *-ing* condition, all present progressives were produced with nonstandard *-in'* except for the 3rd, 6th, and 9th present progressives, which were produced as standard *-ing*. I counterbalanced the rate of *-ing* use by each voice across participants.

Materials

In the current study, I used four Google WaveNet neural synthetic voices synthesized from the most current version of the software in 2023. I chose the most current versions of Voice D, Voice F, Voice H, and Voice I for use in this study. Voice D and Voice I were designed by Google Voice to sound like masculine voices, while Voices F and H were designed to sound feminine. All stimuli were amplitude normalized to 65 dB.

I developed two broadcast passages for use as stimuli in Experiment 2; one was a news broadcast, and the other was a radio DJ broadcast. Participants were randomly sorted into two groups and listened to either a news broadcast or a radio DJ broadcast. Both passages were designed to be roughly one minute long and contained ten present progressive verbs. The passages can be found in Appendix Figure C.

Participants and Procedure

Participants (n = 168; 131 F, 37 M; mean age 19, max age 33) were recruited from the University of California, Davis undergraduate population and were given partial course credit for

their participation in the study. Participants were excluded from the current study if they had participated in Experiment 1. All participants reported that English was their strongest language and had no known hearing difficulties. Before the experiment, participants listened to a short audio clip and completed a word identification task to ensure they could hear and understand the audio.

Participants were instructed that I wanted their input on which of four voices would be best for a new app I was developing. I instructed them that they would hear four voices and evaluate them on nine different rating metrics. Participants were not informed that the voices they would be evaluating were synthetic but were also not under the guise that the voices were human.

Participants were randomly sorted into two groups. One group listened to a news broadcast while the other group listened to a radio DJ broadcast read by each of the four voices. Within each group, listeners heard the script produced by the four voices one at a time and answered evaluation questions about each voice immediately after hearing the script produced by that voice. The evaluation questions were on a sliding scale from 0–100 and used the same question frames as Experiment 1. There was an optional comment box where participants could write evaluations they had about the voice they just heard.

After participants had completed the evaluation portion for the four voices, they were presented with a seven-second clip of the passage read by each voice and were asked to select whether they thought that voice was a human, computer, or unsure. They were able to replay the clip an unlimited number of times to make their decision. Next, participants responded whether they thought human or synthetic voices were better for news broadcasting/a radio DJ overall.

Participants then filled out a demographic questionnaire and reported how often they use voice AI devices and what they use voice AI for.

Results

Cognitive Organization of Social Evaluations

As in Experiment 1, I used principal component analysis (PCA) to investigate whether there was collinearity between some of the social evaluation questions I asked in the study.

The first two principal components account for 72.55% of the variance in the data. Figure 4.1 below illustrates that the directions and loadings of PC 1 and 2 appears to group the rating metrics into three categories: status, credibility, and humanlikeness. Status refers to the social prestige that a speaker carries, which is characteristic of the well-known relationship between intelligence and professionalism in listener evaluations of (ING) from previous literature described in Chapter 1. In the current study, I find that trustworthiness and appropriateness for a news broadcasting position are closely covariant. Pycha and Zellou (2024) found that listeners vary in how credible they consider a synthetic voice based on its regional accent, and here I propose trustworthiness and appropriateness are congruent to listeners' perceptions of credibility in the Pycha and Zellou study. Lastly, PC 1 and 2 closely group naturalness, friendliness, authenticity, and the two social presence questions. More generally, the PCA loadings and the resulting grouping of the social evaluation questions are consistent with previous research on social categorization of synthetic voices (Lilley et al., 2025; Nass & Brave, 2005; Pycha & Zellou, 2024).

Figure 4.1

Principal Components 1 and 2 of the Nine Rating Metrics Presented to Participants

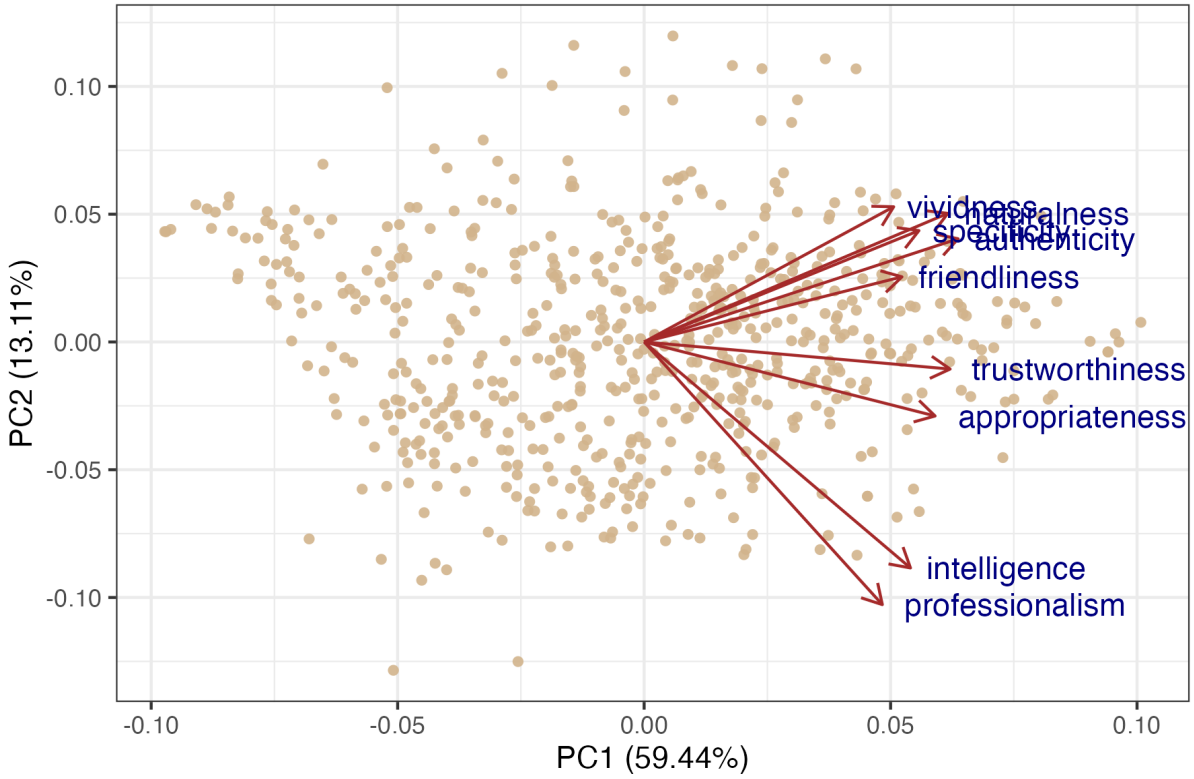


Table 4.1 represents the rating metrics that I reduced into a smaller set of indexical dimensions. To reduce the rating metrics into a smaller set of evaluation categories, each participant's responses to the grouped rating metric were averaged into the reduced dimension. For example, if a participant rated a voice as 60% professional and 65% intelligent, that would be averaged to a score of 62.5% on social status.

Table 4.1

Reduction of Rating Metrics into Social Evaluation Categories

Rating Metric	Reduced Dimension
Professionalism, Intelligence	Social Status
Appropriateness, Trustworthiness	Credibility
Naturalness, Authenticity, Vividness, Specificity, Friendliness	Humanlikeness

Note. The nine rating metric questions participants answered in the study, reduced into three evaluation categories.

Bayesian Regression of Effects on Listener Evaluations of Synthetic Voices

To estimate what effects had a meaningful influence on participant evaluations, I used Bayesian zero-one inflated (ZOIB) regression modeling, an architecture that is explained in detail in Chapter 2. The model formula in this study was developed based on my research questions, findings from Experiment 1, and k-fold cross-validation. I modeled the effect of (ING), participant and speaker gender, speaking context, participants' belief that a voice was a computer, and participants' voice AI use habits. I also included a random effect by-subject.

The model included two-way interactions for computer response and broadcast context, (ING) variation and computer response, (ING) variation and speaker gender, (ING) variation and participant gender, speaker gender and participant gender, and computer response and participant voice AI use.

I also modeled two different three-way interactions. I modeled the three-way interaction of (ING) variation, belief that a voice was a human, and broadcast context. I also modeled the three-way interaction between (ING) variation, belief about whether a voice was human, and the speaker's gender. While it was possible to model more three-way interactions, cross-validation illustrated that including these model terms would likely result in overfitting.

Each of the evaluation categories were fit to separate models based on the grouped categories from the PCA analysis, and all models used the same formula. I used the same model formula for the mu, alpha, and gamma parameters. The phi parameter is estimated only as the intercept without any additional parameter estimates to maximize interpretability of the model. I chose to model the data with priors sampled from the normal distribution with a standard deviation of two. The model syntax is as follows:

participant rating ~ ing use*computer response*broadcast context +
ing use*computer response*speaker gender +
participant voice-AI use*computer response +
ing use:participant gender +
participant gender*speaker gender +
talker + (computer response |id)

Because of the small number of nonbinary participants (n = 2) in my study, I did not include them in the model. I also did not model "not sure" responses to the computer response question because very few responses (n = 28, 4% of total responses) were chosen by listeners. I also could not model a three-way interaction between listener gender, speaker gender, and (ING) use, because each speaker gender did not produce all four levels of (ING) variation.

There were four estimated effects (the effects of speaker and the effect of female voice) that for every social evaluation model had such a high degree of statistical uncertainty that their estimates likely entirely came from sampling the prior distribution. Each of these effects had confidence intervals from -2 to +2, the boundaries of my specified prior distribution, suggesting they are not meaningfully predictive of the data. I chose not to include these four effects on the effect size plots to improve the readability of the graphs.

Interpretation of Regression Coefficients

Below, I interpret the effects estimated by the model that appear to have a meaningfully large influence on participant rating behavior based on Krushcke's (2018) theory of region of practical equivalence (ROPE), described in Chapter 2. First, I interpret the effects that were shown to have a statistically meaningful influence on listener evaluations across all three evaluation categories (social status, credibility, and humanlikeness). I explain these effects together rather than interpreting them within each social evaluation category to decrease redundancy and clarify what effects apply across all social evaluations.

In the subsequent section, I analyze the meaningfully large effects specific to each evaluation category one at a time. I included data visualizations of the trends for effects that have a meaningfully large influence on participants' rating behavior to better illustrate how those effects influence listeners' rating behavior. At the end of the analysis of each evaluation category, I present effect size plots that include all the effects and interactions estimated in the regression model. The 32 model coefficient estimates are plotted, along with error bars for the 95% upper and lower confidence intervals.

Influential Effects Across All Evaluation Categories

1. Participant Belief About Whether a Voice was Human or Computer

Unlike Experiment 1, where participants showed a near-ceiling degree of accuracy identifying the voices as human or computer, participants in this study varied more in their responses and showed a lower degree of confidence on whether voices were human or computer. Table 4.2 counts the percentage of "computer," "human," and "unsure" responses for the synthetic voices in the study. Given that each participant heard four synthetic voices, there are four responses per participant.

Table 4.2
Proportion of Responses for Participants' Belief about Whether a Synthetic Voice Was a Human or Computer

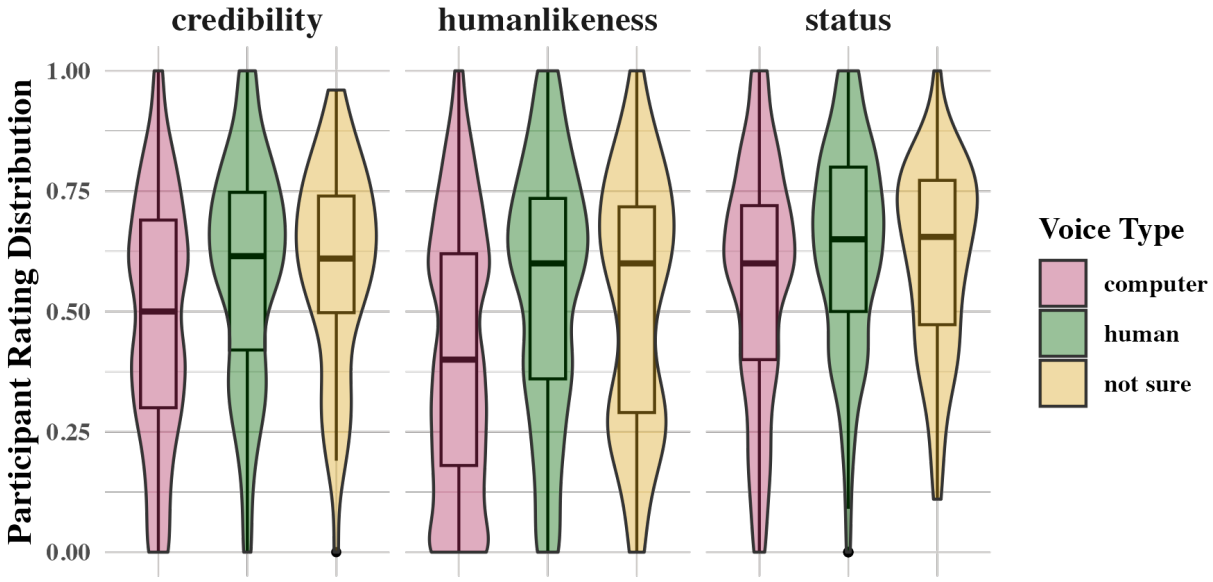
Perception of synthetic voice	Count	Percent of total observations
Computer	467	69%
Human	185	27%
Not sure	28	4%

I found an effect of participants' belief about whether a voice was a human or computer in regression models for every evaluation category. Figure 4.2 illustrates that voices that participants perceived as computers were evaluated substantially lower than voices that

participants thought were human or were uncertain about. The magnitude of this effect varies based on the evaluation category. For example, in evaluations of credibility, voices perceived as computers had a median evaluation score ten points lower than the other two response categories. For humanlikeness, perceived computer voices were not only rated much lower overall but received far more zero ratings than any other evaluation category.

Figure 4.2

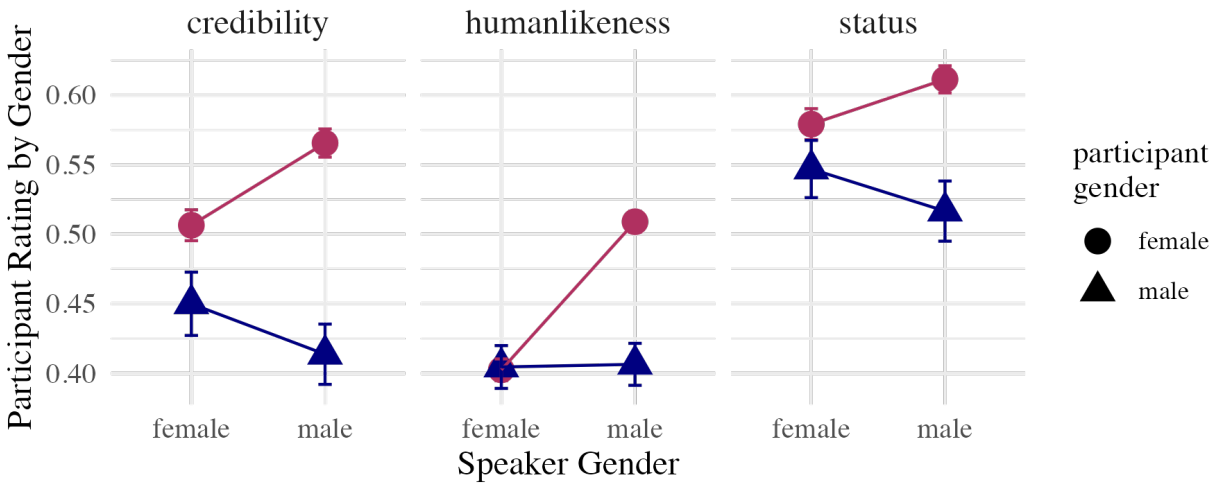
Listener Evaluations of the Three Rating Metrics, Subset by Belief of Human or Computer Voice



2. Gender Effects

In general, female participants gave higher ratings to voices in the study. The effect of the participant gender was one of the largest in the models of credibility and social status. In the humanlikeness model, there is a trend towards an effect of female participants giving higher ratings of humanlikeness.

Figure 4.3
Effect of Listener Gender on Evaluations of Male and Female Speakers



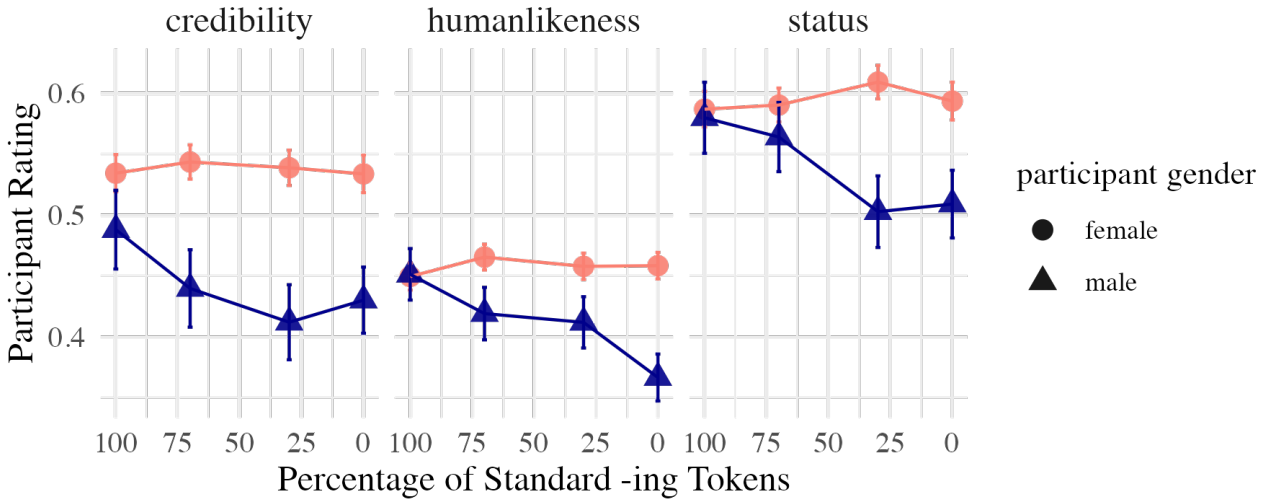
There is a substantial interaction between listener and speaker gender across all the evaluation categories. Figure 4.3 illustrates that across rating metrics, female participants gave higher ratings to voices whose gender was different from their own. This was also true for male participants, except in the humanlikeness evaluation category, where men evaluated both genders as similarly humanlike.

3. Participant Gender and (ING) Variation Judgments

There was also an interaction between participant gender and evaluation of (ING) in all three evaluation categories. Male listeners evaluated voices that used exclusively standard *-ing* much more positively than voices that used any amount of nonstandard *-in'*, while female listeners' evaluations of voices changed only slightly, as shown in Figure 4.4.

Figure 4.4

Differences in (ING) Variable Evaluation Across Participant Genders



4. Broadcast Context

Across the board, participants gave higher evaluations of social status, credibility, and humanlikeness for voices who produced the news broadcast script. However, the effect of (ING) use interacts with the rating metrics differently, which is explored in the individual rating categories below. Next, I explore the regression effects that appeared meaningfully large within social evaluation categories.

Evaluation Category-Specific Influential Effects

Humanlikeness

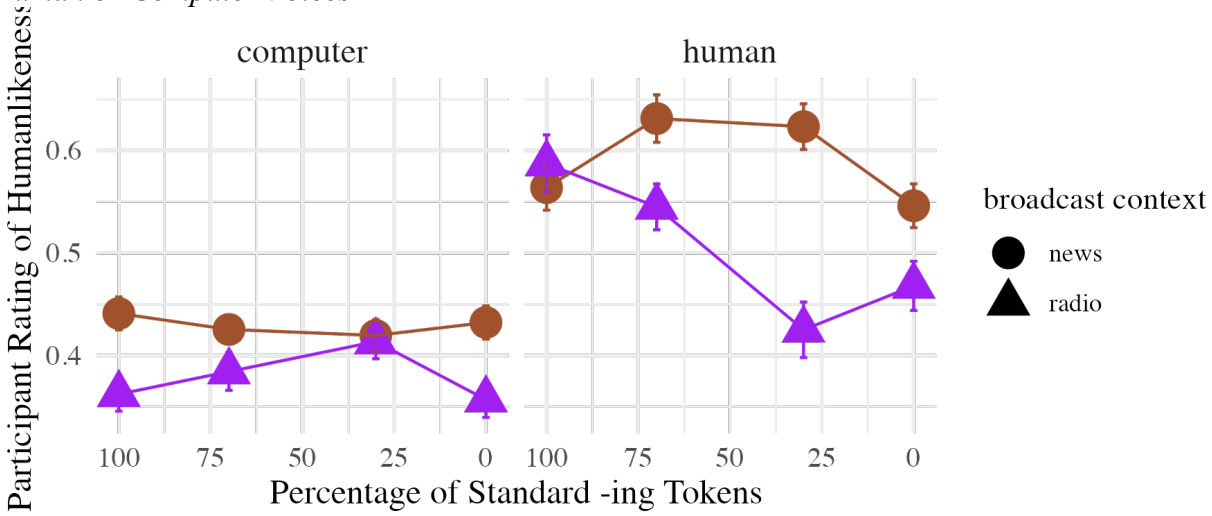
Humanlikeness is the only evaluation category where there was a meaningful effect of (ING) variation alone on participant responses. Voices that used only nonstandard *-in'* were evaluated as less humanlike overall.

I found a three-way interaction between (ING), broadcast, and perception of a voice as human, illustrated in Figure 4.5. When voices were perceived as human in the news broadcast, they were evaluated as less humanlike when they used either exclusively standard or exclusively nonstandard language. But, when they used standard and nonstandard (ING) variably, they were

evaluated as most humanlike compared to the other voices. There was also an interaction between perceived human voice, radio broadcast, and 30% *-ing* usage, where human voices that used mostly nonstandard variation in the radio broadcast were evaluated negatively compared to when they used any other rate of (ING). This is also reflected in the more positive evaluation of 30% *-ing* use in perceived computer voices in the radio broadcast.

Figure 4.5

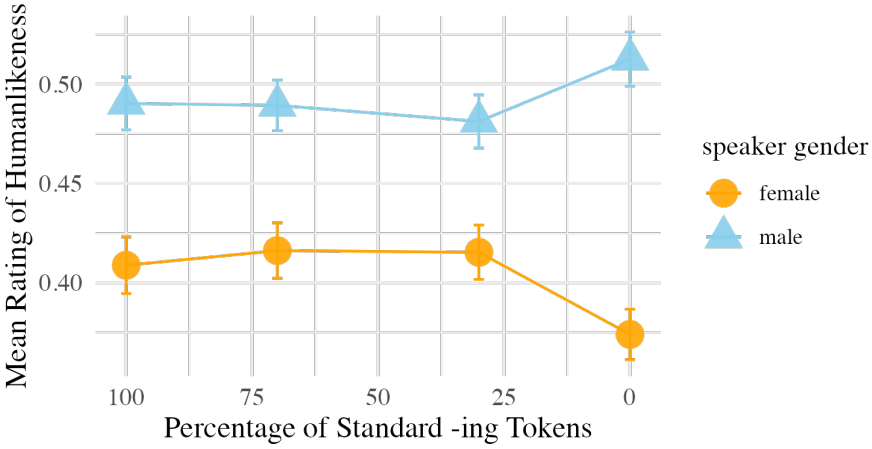
Mean Participant Evaluations of Humanlikeness of the Four Synthetic Voices, Subset by Belief of Human or Computer Voices



I also found an interaction where female voices that used only nonstandard *-in'* were evaluated as less humanlike compared to male voices, shown in Figure 4.6. Meanwhile, male voices that used exclusively the nonstandard variant were evaluated as more humanlike.

Figure 4.6

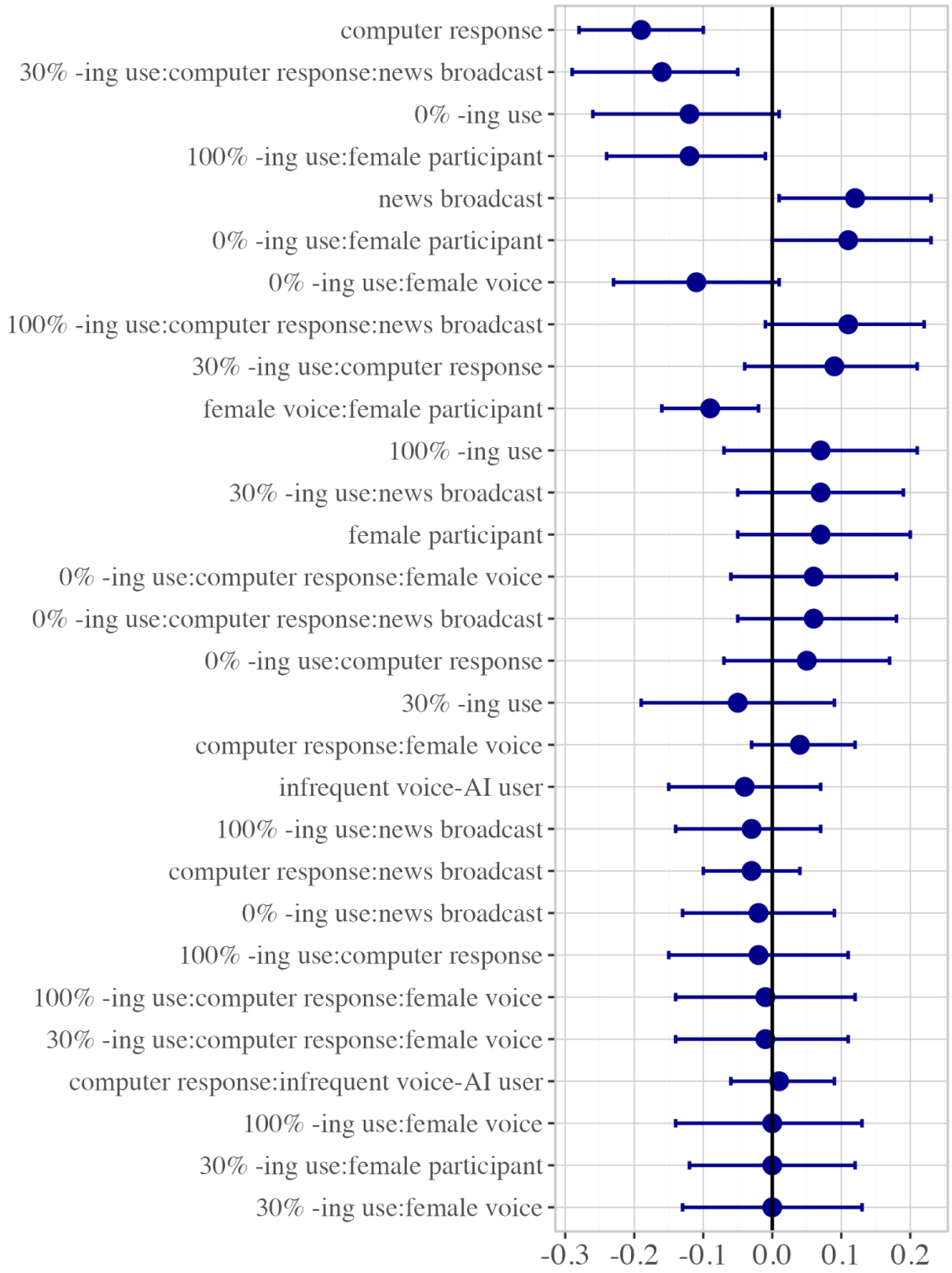
The Influence of (ING) Variation and Male and Female Speakers on Evaluations of Humanlikeness



In Figure 4.7 below, all the effects and interactions estimated in the regression model are plotted, with error bars at the 95% lower and upper confidence intervals.

Figure 4.7

Estimated Effects and Possible Interactions for Participant Ratings of Humanlikeness

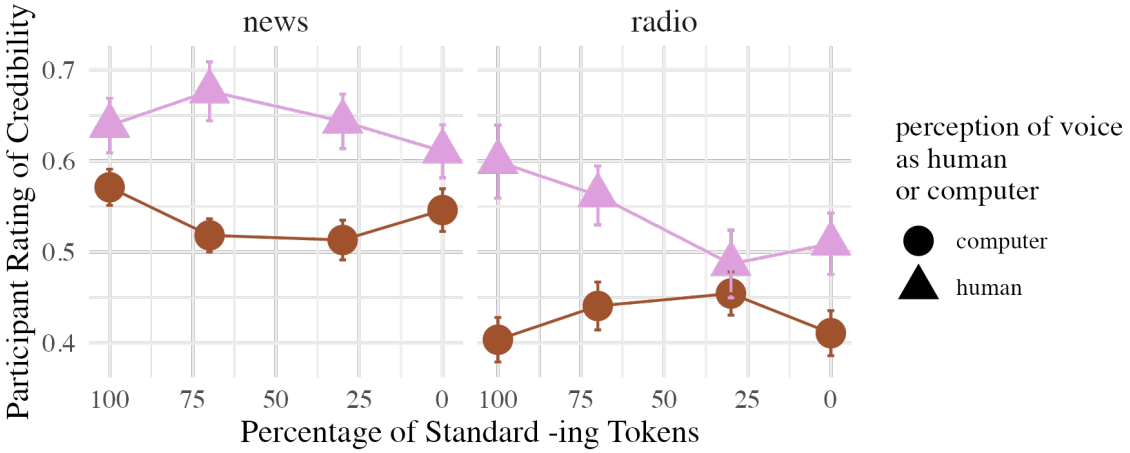


Note. The plot does not include the effect of individual talker or the effect of female voice. Values are plotted in descending order of the absolute value of the effect size.

Credibility

Similar to humanlikeness evaluations, I found a three-way interaction between (ING), broadcast context, and perception of the voice as a human. The regression model reflects a positive effect of 100% *-ing* use in the radio broadcast condition when the participant perceived the voice as a human. On the other hand, there is a three-way interaction between 30% *-ing* use, computer response, and radio broadcast, where voices that were perceived as computers got a "boost" in credibility while perceived human voices were rated as less credible. As shown in the Figure 4.8 below, there is a drop in credibility from completely standard (ING) use to nonstandard (ING) use in perceived computer voices, but perceived human voices received roughly similar ratings of completely standard (ING) use versus 30% nonstandard *-in'* use.

Figure 4.8
Mean Participant Evaluations of Credibility of the Synthetic Voices, Subset by Whether the Participant Believed the Voice to be a Human or Computer

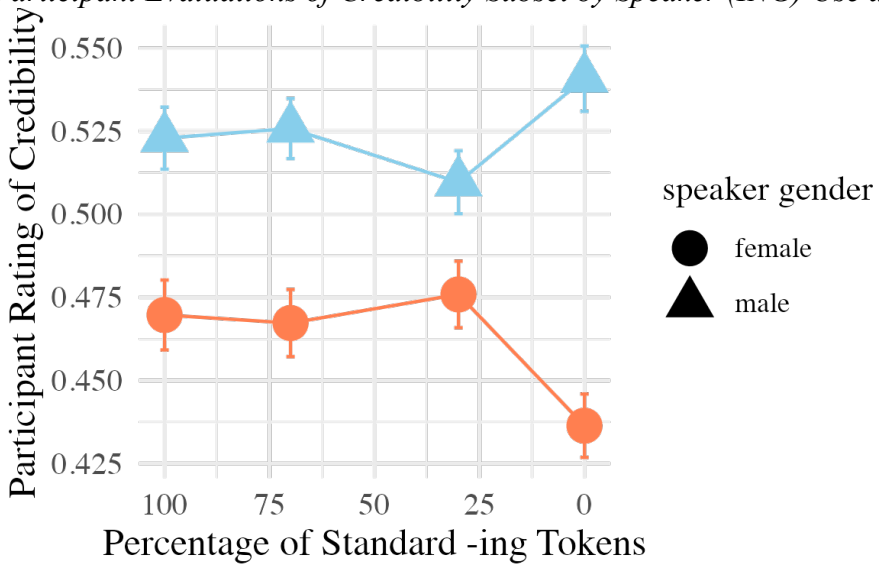


I also found an interaction effect of (ING) use and speaker gender on listener evaluations of credibility, as illustrated in Figure 4.9 below. When female voices used exclusively nonstandard *-in'*, they were rated as less credible than male voices that used only nonstandard *-in'*. On all other rates of (ING) variation, participant evaluations of credibility varied only

slightly between male and female voices, though the model output also shows a possible trend between 30% *-ing* use and speaker gender.

Figure 4.9

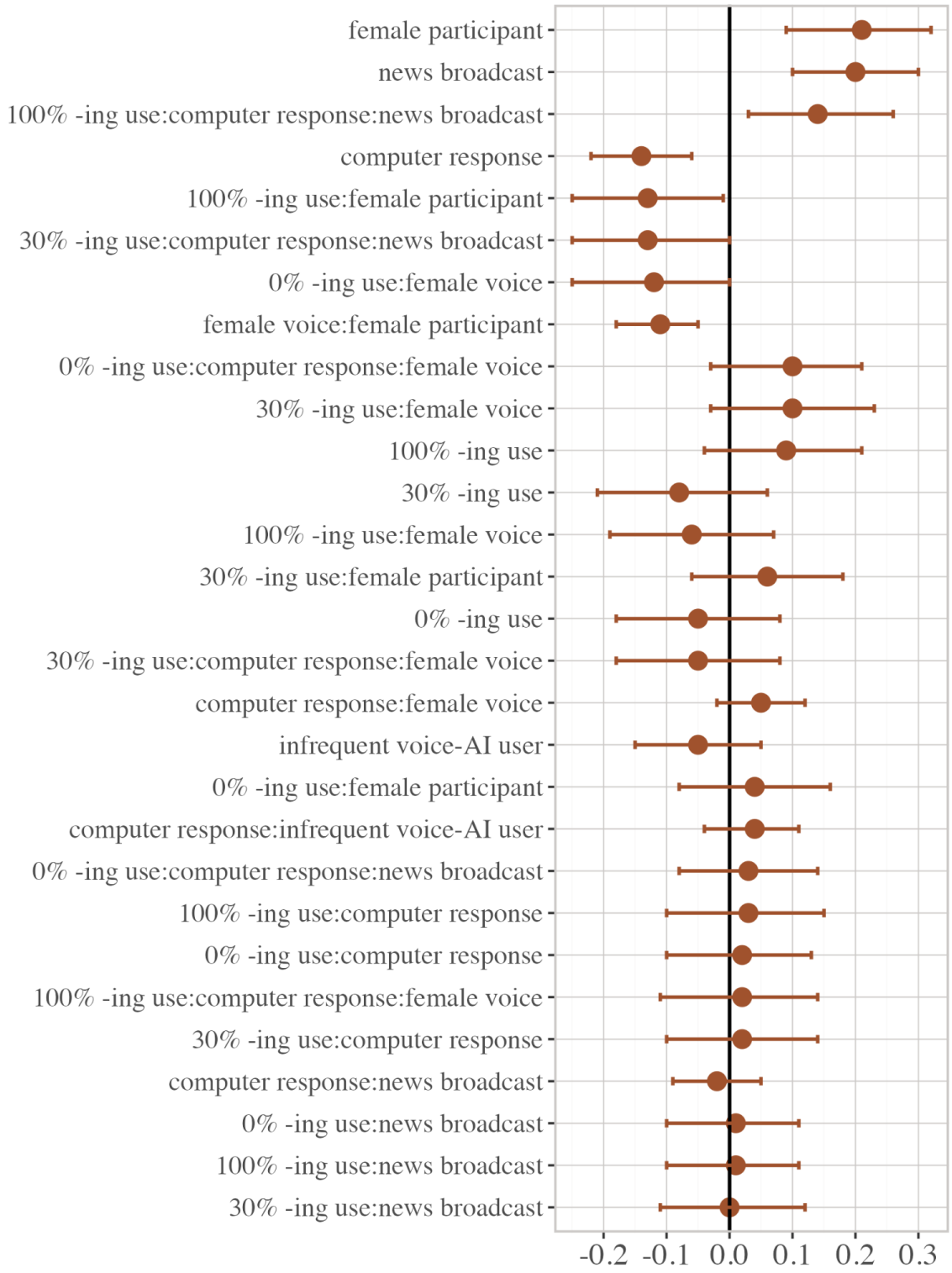
Participant Evaluations of Credibility Subset by Speaker (ING) Use and Speaker Gender



Next, Figure 4.10 presents the estimated effects and all possible interactions for participant ratings of credibility.

Figure 4.10

Estimated Effects and Possible Interactions for Participant Ratings of Credibility

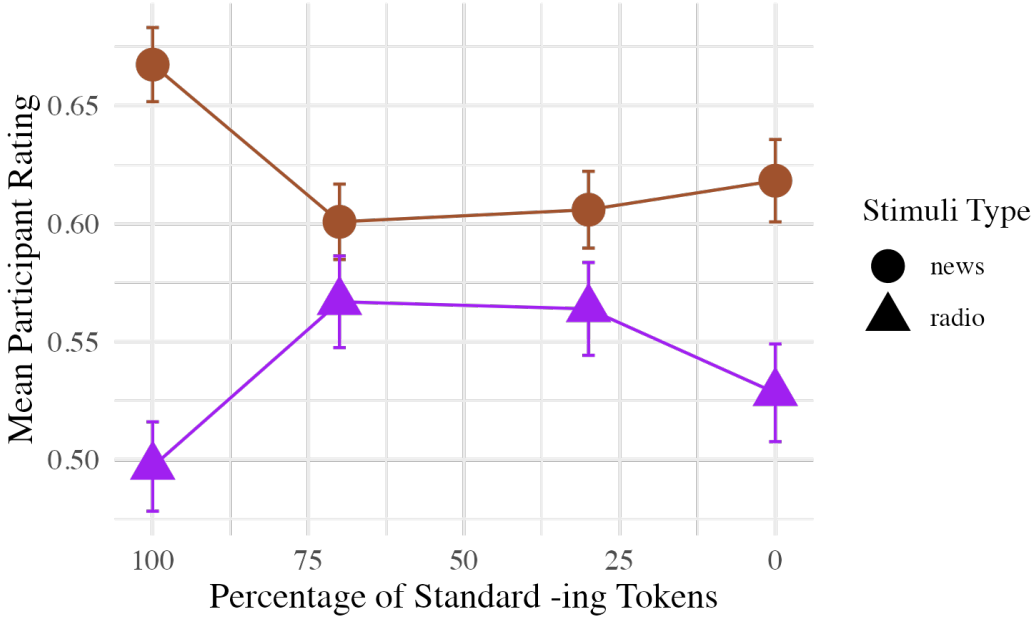


Note. The plot does not include the effect of individual talker or the effect of female voice. Values are plotted in descending order of the absolute value of the effect size.

Social Status

Unlike humanlikeness and credibility, I did not find a three-way interaction between (ING), broadcast context, and perception of a voice as human. For social status, I did not find an interaction between belief that a voice was a computer and (ING) variation.

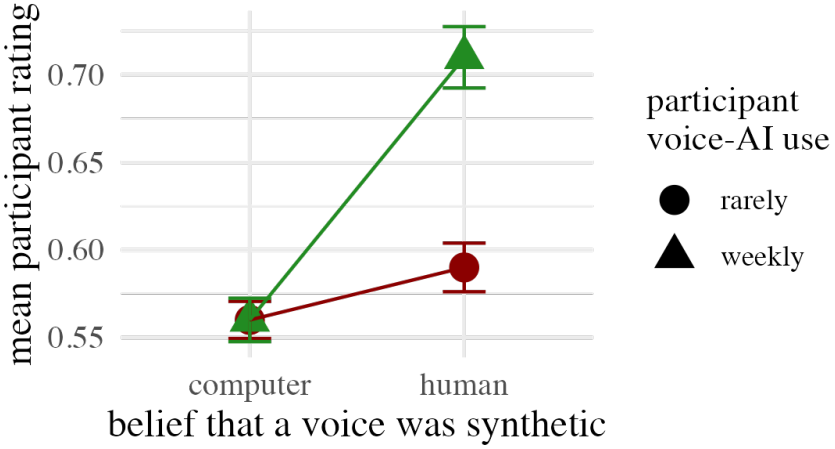
Figure 4.11
Mean Participant Evaluations of Social Status Plotted Against Speakers' (ING) Use



Instead, the primary effect on listener evaluations of social status was a two-way interaction effect between (ING) use and broadcast context. As illustrated in Figure 4.11, evaluations of (ING) variation are nearly inverse based on whether the participants heard the radio or news broadcast. For participants in the news broadcast condition, speakers that used exclusively standard *-ing* were evaluated as having greater social status than speakers who used any amount of nonstandard *-in'* use. Yet, participants in the radio DJ condition considered 100% *-ing* users to have the lowest social status.

Figure 4.12

Participant Evaluations of Social Status Based on Their Voice AI Use Frequency and Belief Whether a Voice Was Human or Computer



Interestingly, I found that participant's voice AI use habits influenced how participants evaluated the social status of the voices in this study, shown in Figure 4.12. In general, infrequent voice-AI users considered the voices in the study to have lower social status. But, there is a large interaction effect between participant voice AI use and participants' belief about whether the voice was a computer. Participants who use voice AI often evaluated speakers that they perceived as human much more positively than those they perceived as computers. Figure 4.13 plots the effects and interactions estimated in the social status model.

Figure 4.13

Estimated Effects and Possible Interactions for Participant Ratings of Social Status

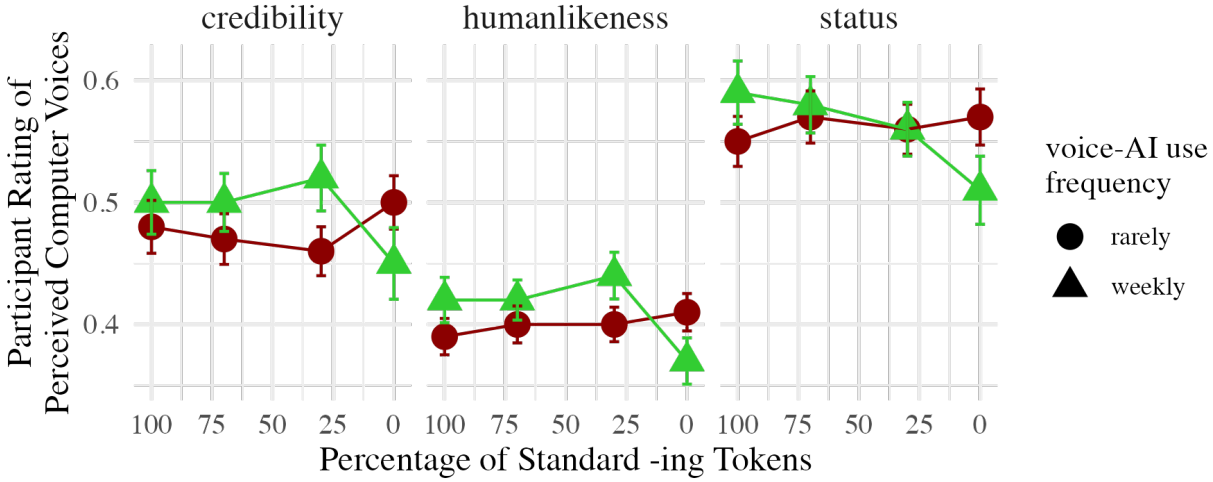
Note. The plot does not include the effect of individual talker or the effect of female voice. Values are plotted in descending order of the absolute value of the effect size.

Listener Voice AI Use Habits, (ING), and Computer Voice Evaluations

Considering that participants' voice AI use habits and belief about whether a voice was a computer interacted in their perceptions of social status, a natural question is whether there could be a three-way interaction between voice AI use frequency, perception of a voice as a computer, and (ING) on participants' social evaluations. Do participants that use voice AI assistants frequently evaluate sociolinguistic variation in synthetic voices more positively than participants who rarely interact with voice AI?

Unfortunately, this three-way interaction cannot be estimated in the model because some participants in the study evaluated all the voices as computers or all the voices as human. Figure 4.14 plots mean participant ratings of voices they perceived as computers in the three evaluation categories, subset by how often they use voice AI.

Figure 4.14
Participant Evaluations of Voices Perceived as Computers Only, Plotted Against Participants' Voice AI Use Frequency



Note. These graphical trends could not be statistically evaluated in the regression model.

A visual inspection of the graph trends in Figure 4.14 suggest that frequent voice AI users perceive nonstandard *-in'* use in computer voices negatively, while infrequent users or nonusers of voice AI rated voices that used nonstandard variation relatively neutrally or perhaps even

more positively in voices that used *-in'*. Importantly, these trends are based only on voices that the participants accurately perceived as computers, ruling out the possibility that infrequent voice AI users had incorrectly believed them to be human talkers. In fact, there is no difference in accuracy of identification of computer voices between frequent and rare voice AI users in this study (frequent users' accuracy = 0.69 ± 0.029; rare users' accuracy = 0.68 ± 0.022). A future study could investigate the potential impact of this three-way interaction in a regression model.

Discussion

In the current study, I examined the influence of listener and speaker traits, (ING) variation, and perception of voices as a human or computer. I find that each of these variables are influential on listeners' evaluations of synthetic voices. Women gave higher ratings to the voices than men did, and participants of both genders evaluated the voices more negatively when the voice was the same gender as their own. This gender interaction effect is consistent with studies in sociolinguistics that listeners evaluate speakers of their same gender more critically than they do for other genders (Gordon, 1997; Wollum, 2019). Speakers who read the news broadcast were rated as significantly more credible and had higher social status, likely because participants associate news broadcasts with knowledge and prestige, compared to more colloquial speech typical of radio broadcasts (Bayley & Zapata, 1994).

(ING) Variation

A primary question of this research study was whether (ING) variation influences participant evaluations of synthetic voices. The only evaluation metric where the regression model estimated a simple effect of (ING) use was humanlikeness, where 0% *-ing* use was perceived as less humanlike across the board. Listeners' perception that 0% *-ing* is less

humanlike regardless of any other factors may be because human speakers very rarely exclusively use *-in'* (Houston, 1985).

Instead, differences in perception of (ING) primarily interacted with other judgments participants made about the speaker or the participant's personal characteristics. The two-way interaction between participant gender and (ING) use was one of the largest effects on every social evaluation in this study. I found that women not only gave substantially more positive ratings overall but also only slightly varied their perception of different levels of (ING) variation. In contrast, sociolinguistic studies that suggest women evaluate nonstandard variation more negatively (Trudgill, 1972). But female speakers that used nonstandard *-in'* were evaluated as less trustworthy, less appropriate for broadcasting, and less humanlike than male speakers. This finding is consistent with findings from sociolinguistics that female speakers are evaluated more negatively than male speakers when they use nonstandard speech (Gordon, 1997; Mechler, 2025).

For humanlikeness and credibility, differences in (ING) variation were closely tied to the perception of a voice as a computer. The three-way interaction between (ING) use, computer response, and news broadcast affected listener evaluations of credibility and humanlikeness very similarly in their respective models. When participants believed a voice was a computer producing a news broadcast and used only the standard *-ing* variant, they evaluated that voice as substantially more credible and more humanlike compared to other voices. On the other hand, voices that used nonstandard *-in*', read the news, and were perceived as computers were evaluated negatively.

But, I did not find a three-way interaction of broadcast context, (ING), and belief that the voice was a computer for listener evaluations of social status. This is surprising given the

interactions of computer response and (ING) in the regression models for credibility and humanlikeness. Instead, there was a two-way interaction between (ING) and broadcast context—consistent with previous sociolinguistic studies on (ING) that investigated professionalism and intelligence (Campbell-Kibler 2007; Labov et al., 2011). In particular, the largest effect on listener evaluations of social status was the interaction of 100% standard *-ing* use and the news broadcast condition, very similar to the finding by Labov et al. (2011). In other words, once participants heard any amount of nonstandard variation in the news broadcast condition, they evaluated those voices negatively, but increasing rates of *-in*' did not lead listeners to give even more negative social evaluations. In contrast to the news broadcast condition, I find that 100% standard *-ing* use in the radio DJ condition was rated the lowest on social status. I propose this is because listeners expect some amount of nonstandard variation in radio DJ speech, which is typically colloquial (Bayley & Zapata, 1994).

Perception of a Voice as a Computer

A substantial influence on how listeners socially evaluated voices in this study was whether they believed them to be human or computer. In each regression model, I found a simple effect of perception of a voice as a computer. This effect varied in intensity across social evaluation categories. As shown in Figure 4.2, there was a ten-point difference for median evaluations of credibility, and a twenty-five-point difference for median evaluations of humanlikeness; additionally, participants were far more likely to rate perceived computer voices as a zero on humanlikeness than in the social status or credibility categories. In fact, belief that a voice was a computer was the most influential effect on participant evaluations of humanlikeness.

Yet, neither the interaction between computer response and voice-AI use frequency nor the simple effect of voice-AI use influenced participant evaluations of humanlikeness. This is surprising since people who have more experience with voice-AI in their daily life might have stronger expectations about what makes a synthetic voice sound realistic or humanlike. Instead, social status is the only evaluation category where a participants' voice AI use habits influenced their perceptions of the voices. I found an interaction between a participant's belief about whether a voice was a computer and their personal voice AI use habits. Participants use voice AI rarely considered perceived computer voices to have similar social prestige to perceived human ones. On the other hand, frequent voice AI users considered voices that they perceived to be human as substantially more professional and intelligent than voices they perceived to be computers. People who use voice AI frequently have more experience with its limited social uses, which might explain why they evaluated perceived computer voices as having lower social status (Gambino & Liu, 2022).

Figure 4.14 illustrates that there might be a connection between how often people use voice AI technology and their perception of (ING) variation in computer voices. While rare users of voice AI gave increasingly positive evaluations of nonstandard *-in'* use, the opposite was found for frequent users of AI; in fact, their ratings dropped sharply for the 0% *-ing* condition. This suggests that people who use voice AI often have entrenched expectations for what voice AI should sound like and penalize voices that differ from the standard heavily. Although this three-way interaction could not be estimated in the regression model, the two-way interaction of voice AI use frequency and belief that a voice was a computer in the social status model suggests that the graphical trend illustrated in Figure 4.14 could be statistically reliable if I had sufficient observations to test it.

Implications for Theory of Human-Computer Interaction

The findings of the current study illustrate that social evaluations are mediated by listeners' expectations for computer voices in unique ways. Participant evaluations of credibility and humanlikeness were most influenced by the belief about whether the voice was a computer, while participant ratings of social status were influenced more by (ING) variation.

I propose that the interaction between (ING) use, computer voice, and broadcast context is due to participants' expectations for social behavior by computers in everyday life. Similar to Experiment 1, participants overwhelmingly responded they preferred human voices overall; only 4% of participants responded that they preferred synthetic voices for reading the news/radio. I suggest the credibility and humanlikeness evaluation patterns by participants stem from their top-down preference for human voices in real-life applications like the "app" I was developing in this study.

Meanwhile, evaluations of social status were not influenced by their beliefs about the role of computers in society and instead were an extension of the social meaning of (ING) variation in human speech. Statistical analysis clearly reflects that there is not an interaction between computer response and (ING) variation on participants' evaluations of social status. In other words, participants might consider both human and synthetic talkers to sound similarly intelligent and professional when they vary their use of (ING), but participants do not consider computer voices trustworthy, appropriate, or humanlike when they imitate sociophonetic variation that is typically exclusive to human voices. Moreover, social status might be an evaluation metric where social evaluations from human voices are directly transferred to computer ones, but for credibility and humanlikeness, participant judgments are mediated by their expectations of a "social script" for computer voices.

Finally, although this effect could not be statistically estimated in the model, the negative evaluations by frequent voice AI users of voices they perceive as computers that use nonstandard *-in'* could have meaningful implications for theory of human-computer interaction and sociolinguistics. For example, frequent voice AI users consider social computers less intelligent than humans based on their experiences with its limited social uses (Gambino et al., 2020; Gambino & Liu, 2022). Figure 4.14 reflects that frequent voice AI users have started to apply the negative associations of nonstandard *-in'* that have been found in previous sociolinguistic studies on (ING). People who use voice AI rarely or never do not appear to transfer the social meaning of (ING) to computer voices. The sudden sharp decrease in evaluations of nonstandard *-in'* by frequent users suggests they have developed strong social expectations for computer voices. Through frequent use, people have unconsciously begun to interact with computers as a new social group, and "branched off" into a new social script for human-computer interactions. Moreover, the findings of the current study are similar to the theory of routinization of social scripts for social computers.

Implications for Theory of Sociolinguistics

In theories of sociolinguistics for human speech, it is widely accepted that social evaluations of a person's speech are interconnected to multiple speaker traits, or indexical fields (Eckert, 2008). In the current study, the absence of a simple effect of female speaker on any of the social evaluation categories provides further evidence that participant evaluations of speaker gender are closely intertwined with other variables like conversational context, listener gender, and socio-stylistic language change. This perspective is supported by the strong interaction effect I found between participant and speaker gender and the interaction between participant gender and evaluation of the (ING) variable.

Another important source of social meaning in language comes from a speakers' *social group*—in this experiment, I examined whether a listener's perception of a voice as a human or computer constitutes a new social group. Listeners notice patterns in language use by a particular group and assign social meaning to them (Eckert, 2012). Social groups that listeners have strong stereotypical associations of are more strongly evaluated on their variant use (Preston, 1996). In the current study, the interaction effects between (ING) and belief that a voice is a human provide evidence that social stereotypes are more strongly evaluated when listeners have a strong stereotype for that particular group—in today's world, computer voices rarely use socio-stylistic language variation, so people have not developed social stereotypes for computer voices. But in the current study, the interaction effect between frequent voice AI use and variable perception of social status provide evidence that a unique social group for computer voices might be forming among frequent users of voice AI. Similar to the findings by Preston (1996) for human speakers, the population most familiar with the social group being evaluated have the strictest expectations for that group's socio-stylistic language use.

Limitations

A limitation of this study was the small number of nonbinary participants who participated in the study, and thus I could not model the effect of gender-nonconforming participants' evaluations of the voices in this study. A future study can specifically aim to recruit nonbinary participants to a balanced level with other genders.

I was not able to model the three-way interaction between participant voice AI use, (ING) variation, and belief about whether a voice was a computer because some participants entirely inaccurately believed that all the voices in the study were human, while other participants

correctly identified all as computers. A larger study with more participant observations might be sufficient to reliably estimate this three-way interaction in a regression model.

Another challenge for statistical modeling in the current study was the small number of "not sure" responses to the computer response question; only about 4% of respondents chose that they were not sure if a voice was human or computer. Based on the graphical trends from Figure 4.2, it appears that participants' "not sure" responses pattern similarly to voices that were perceived as human. In future studies, I would aim to either use stimuli that are humanlike enough to elicit a greater proportion of "not sure" responses, or recruit more participants so that reliable statistical analysis of that group could be evaluated.

As a final consideration, this study only examined young adult listeners. Significant generational differences are found between younger and older adult users of technology, and so I would expect to find that older adults socially evaluate the synthetic voices in this study differently. In the next chapter, Experiment 3, I address this research question.

Chapter 5: Older Adults' Evaluations of Sociophonetic Variation in Synthetic Speech

Introduction

The main findings of Experiment 2 are relevant to both the fields of sociolinguistics and human-computer interaction. First, social evaluations are influenced by gender, and these factors interact with how listeners perceive the (ING) variable. Consistent with other sociolinguistic studies, I find that feminine voices are penalized more when they use the nonstandard *-in'* variant when compared to masculine voices, and participants evaluate voices more positively overall when the voice is a gender different from their own. Listeners are more sensitive to (ING) variation in formal speaking contexts and penalize voices that use nonstandard *-in'* more when it is in a more formal context compared to an informal one. I also find that men judge nonstandard *-in'* variation more negatively than women. (ING) variation influenced participants' evaluations more for social status, while belief about whether a voice was human or computer was most influential on participant responses for credibility and humanlikeness.

Social Evaluations Across the Lifespan

In light of these findings, I next asked whether these evaluation patterns would extend to listeners of other age groups, namely older adults. Previous research has shown that older adults evaluate sociolinguistic variation differently from younger adults, and these different evaluation patterns can be specific to a particular sociolinguistic variant. (ING) variation has been shown to be evaluated more negatively by older listeners than younger ones (Labov et al., 2011). While Labov et al. (2011) found evidence of age grading between the high school and college-aged groups, their study does not address social evaluations of (ING) variation by middle-aged or older adults. The inclusion of middle-aged and older adults in sociolinguistic studies is important because there is evidence that people return to a more vernacular style of speaking as they move

into middle and older age (Paunonen, 1994). This change is hypothesized to occur as older adults feel less pressure to conform to formal speech in professional workplace contexts, and by extension evaluate vernacular speech less critically when used by other talkers. Labov (1972) found that older adult men used more vernacular forms than the immediately younger age group and suggested that in their old age, men become less concerned with power relationships and relax their standard language use. Eckert (2017) hypothesizes that the workplace is clearly associated with standard language norms, but in retirement the status of standard language becomes more complex. However, Eckert (2017) emphasizes that this perception is an assumption, because the lack of sociolinguistic studies on older adults is "a vast wasteland in the study of variation" (p. 165). The nature of sociolinguistic perception in adulthood is a very open question that the current study hopes to address.

Mechler (2025) investigated how middle-aged and older adults perceive (ING) variation across the lifespan and found that middle-aged and older men evaluated vernacular speakers the most positively compared to the other age and gender groups, while young women were perceived as the least professional by listeners of all age groups and genders. She found (ING) variation had only a small effect on listener evaluations; instead, listeners' evaluations seemed to be primarily based on speaker age and gender, which Mechler hypothesized might be due to cues of age and gender outcompeting the possible influence of (ING). In the current study, I consider the influence of gender, age, (ING), and the possible interactions among these variables.

Computer and Voice AI Use Across the Lifespan

Social perception of (ING) variation by middle-aged and older adults is an open topic for sociolinguistic research, and there are further questions in human-computer interaction related to how older adults use technology. Answering these questions can be challenging because older

adults have been shown to use the internet and voice AI less than younger adults, particularly for social purposes (Auxier & Anderson, 2021). This suggests that older adults' social evaluations of computer voices could be influenced by both social factors like (ING) variation and gender as well as perception of computer voices.

Intersection of Sociolinguistic and Social-Technological Ideologies

Another factor that further increases the complexity of sociolinguistic perception studies in computer speech is the possible intersection between age-graded sociolinguistic perception and the social and political life events that we collectively think of as generational divides. Eckert (2017) gives the example that the fiscal conservatism of middle-aged people who were children during the Great Depression is attributed to *both* their age and their life experience during the Depression. In this sense, Eckert argues, chronological age and cohort membership can overlap in ways that are more socially complex than a person's age alone.

The interaction between age and generational ideology can be extended to perceptions of smart computers in today's world It is especially interesting that a key generational divide in today's world, the widespread use of AI and "smart" computers, is most often used by adults younger than 30 (Auxier & Anderson, 2021) —the age at which people are more likely to use standard language more often (Labov et al., 1972). Thus, younger adults are currently experiencing the social pressure to conform to standard language norms but are also more accepting of "social" computers that mimic human behavior and therefore might be expected to react more positively to sociolinguistic variation in computer voices, a new social behavior. Yet, older adults are hypothesized to be more permissive of nonstandard language use but are ideologically at odds with the proliferation of and highly distrustful of computers that mimic social traits (Thormundsson, 2024). The competing influence of the generational cohort who is

characteristically distrustful of AI against the same older adults' tendency to return to more vernacular forms could have complex influences on how they socially evaluate nonstandard variant use when spoken by a computer.

Technology for the Aging Population

Despite older adults' negative perceptions of AI, there are also potentially immense benefits to its use for aging and disabled populations. Social isolation in elderly populations has been shown to be at least as dangerous to an elderly adult's life expectancy as obesity or physical inactivity and even has a predictive effect on whether an elderly adult will develop cardiovascular disease (Holt-Lunstad et al., 2010; Leigh-Hunt et al., 2017). Access to technology for elderly adults has been shown to improve their health outcomes in multiple ways. First, Heo et al. (2015) found that elderly adults who often use the internet experience fewer feelings of social isolation and a greater sense of social support. A small study by Griol and Callejas (2016) suggested that voice AI chatbots designed to complete cognitive exercises with patients with Alzheimer's disease might help slow the progression of the disease. Healthcare systems are moving towards greater use of agentic AI for both social support and cognitive decline monitoring (Khalil et al., 2025), which is especially beneficial for older adults who do not have social support networks in family or friends. But additional research is necessary to understand what aspects of voice AI interactions are truly beneficial to the aging population to safeguard against overreliance on agentic AI for healthcare purposes (Khalil et al., 2025).

However, this technology is only useful to older adults if they use it regularly, which may be challenging given their negative attitudes toward AI and lower rates of technology use. One practical solution could be the creation of conversational voice AI tailored to their needs. A study by Hu et al. (2024) worked collaboratively with older adults to create voice assistants that not

only serve functional purposes but also can converse about social topics related to exercise and health education. They found participants were more likely to use voice AI if they were able to customize it. My current study investigates how sociolinguistic traits are perceived, gathering information that could improve the social-functional approach described by Hu et al. (2024). Language is inherently pro-social, and sociolinguistic variation, when used by human talkers, is socially meaningful (Eckert, 2012).

Current Study

The current study, Experiment 3, aims to address an interdisciplinary set of research questions from sociolinguistics, human-computer interaction, and research on healthy aging. First, I am interested in how social evaluation of (ING) evolves across the lifespan, and if social evaluations of (ING) variation are similar for voices that listeners perceive as human versus voices that listeners perceive as synthetic. Next, I ask if social evaluation of computer voices is different for older adults, and whether this interacts with older adults' perception of (ING) variation in voices they perceive as computers. Finally, I ask whether older adults' accuracy in perceiving the same voices in Experiment 2 as computers is comparable to that of younger adults.

I expected to find that older adults perceive (ING) variation differently than the younger adults in Experiment 2. Like Experiment 2, I expected that some participants would mistake computer voices for human ones and would give more positive social evaluations for voices they believe are humans. I also expected to find that older adults perceive voice AI more negatively, particularly on metrics of trustworthiness and solidarity. I expected to find that, like Experiment 2, younger female voices would be perceived more negatively when they use the nonstandard variant; I expected this effect to be even greater for older adult listeners. I anticipated that the

interaction of nonstandard *-in'* variant use with the perception of a voice as a computer would be different than what I found in Experiment 2. To this end, the current study is important not only to understand the social perception of (ING) across the lifespan, but also in a new social context also in which older adults are currently understudied.

Methods

Experiment Design

Experiment 3 follows the same experimental design as Experiment 2. Participants listened to four recordings of the same one-minute passage that contained ten instances of present progressive verbs read by four different Google synthetic voices. The two feminine synthetic voices used in this study were Voice F and Voice H, and the two masculine voices were Voice D and Voice I, identical to Experiment 2. Both stimuli passages can be found in Appendix Figure C.

Participants and Procedure

Participants (n = 247; age range 40 to 82, mean age 54; 119F, 95M, 2 nonbinary) were recruited through the web-based survey platform Prolific and received $6.00 USD for their participation in the 20-minute study. Participants were eligible to take the survey if they were between 30–100 years old and resided in the state of California at the time of survey completion. No participants in the current study had participated in Experiment 1 or 2. All participants reported no known hearing difficulty and stated English was their strongest language. At the start of the survey, participants completed a short word-identification task to ensure they could hear and understand audio.

Participants were instructed that I was developing a new app and wanted feedback on which voice was best to use for the app. Before the start of the experiment, participants were

informed they would rate four voices on nine rating metrics (professionalism, intelligence, appropriateness for broadcasting, friendliness, trustworthiness, naturalness, authenticity, specificity, and vividness). They were not informed that the voices in the study were synthetic voices.

Participants were sorted into either a news broadcast or radio DJ broadcast group. Participants in each group heard all four synthetic voices (two feminine, two masculine) read the passages one at a time, each with different rates of *-in'* (100%, 70%, 30%, or 0% nonstandard *-in'* use). After listening to each voice, participants used a sliding scale from 0–100 to provide ratings on each of the same survey questions as in Experiments 1 and 2. There was an optional free-response text entry box where participants could provide additional comments about the voice.

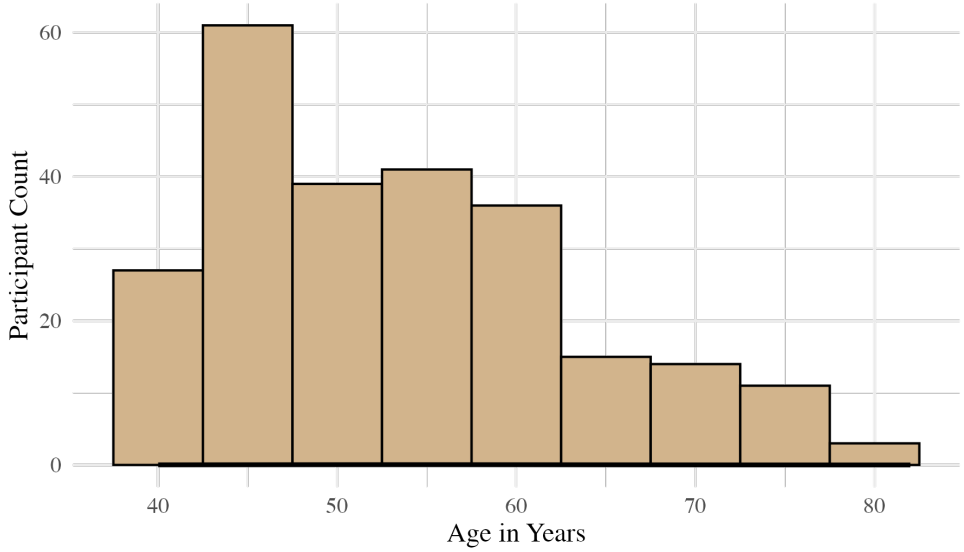
After participants evaluated all four voices, they were presented with a seven-second clip of each voice and asked to select whether they thought the voice was human, computer, or unsure. They also responded whether they thought synthetic or human voices were better for broadcasting overall. Next, participants were asked what they thought we were testing in our study and were required to enter a response into a free-response question box. Finally, participants completed a demographic questionnaire and reported how often they use voice AI devices in their everyday life and what, if anything, they use voice AI for.

Results

When collecting data for Experiment 3, I aimed to recruit a relatively balanced spread of participant ages from 40 to 75, with the expectation that participants were unlikely to be older than 75 (but were still eligible to complete the study, up to age 100). While I specifically targeted the Prolific survey to collect a balanced number of participants over 75, few participants

of this page were willing to take the survey. Barnard et. al (2013) found that elderly adults are more likely to struggle with technology and have lower rates of use, which may explain the difficulty in recruiting elderly participants for this study. Figure 5.1 below illustrates the distribution of ages of participants in Experiment 3.

Figure 5.1
Distribution of Participant Ages



Cognitive Organization of Social Evaluations

I used principal component analysis (PCA) to examine whether participants' evaluations of the nine rating categories were covariant, and if they covaried similarly to the evaluation patterns I found in Chapters 2 and 3. The first two principal components account for 80.54% of the variance in the data.

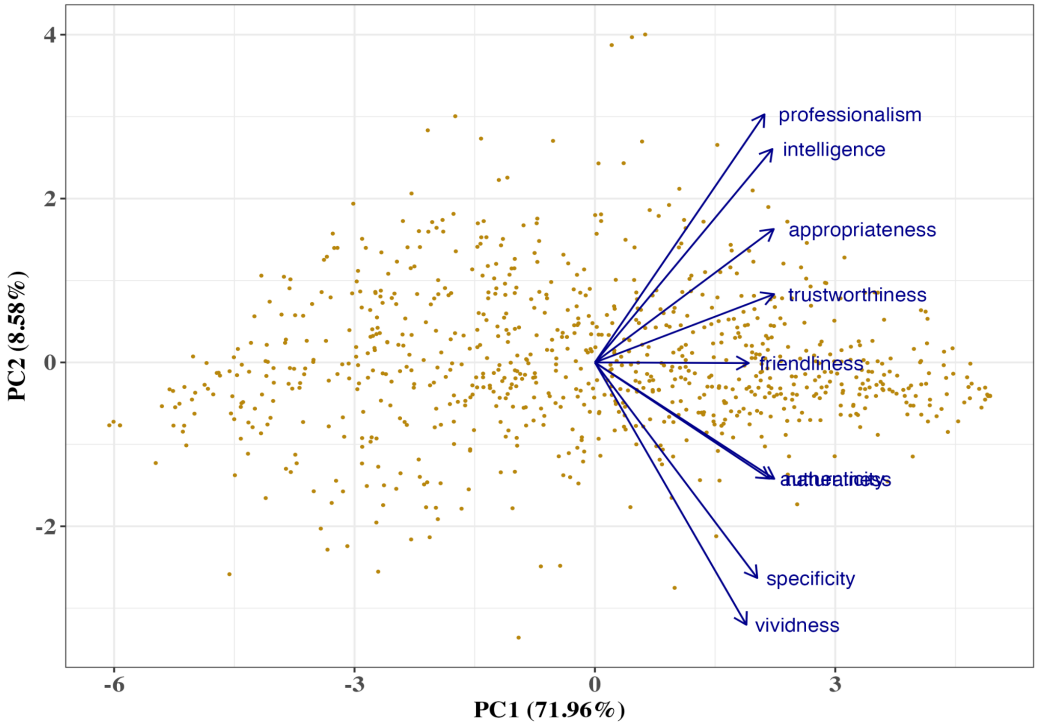
Broadly speaking, older adults had finer-grained interpretations of what each of the nine rating metrics was designed to measure. In Experiments 1 and 2 with younger adults, PCA reflected that participants' social evaluations could be collapsed into four and three dimensions respectively. For example, the PCA of the younger adult data from Experiment 2 reflects a very high degree of covariation between ratings of naturalness, authenticity, friendliness, specificity, and vividness, which I suggested reflects evaluations of "humanlikeness." However, in the older

adults' evaluations, I find very close covariation between naturalness and authenticity, as well as close covariation between specificity and vividness, but not covariation between those two clusters. This suggests that the older adults considered these two kinds of measures of how "real" a synthetic voice sounds to have distinctly different meanings, while younger adults appear to broadly lump those rating metrics together into a holistic judgment of humanlikeness.

Like Experiments 1 and 2, I find that evaluations of professionalism and intelligence covary. Also consistent with the two previous experiments I find that appropriateness and trustworthiness are situated in between evaluations of status and humanlikeness. Experiment 3, the current study, is the only study where friendliness is a distinct evaluation category; in the other studies on younger adults, friendliness was closely covariant with measures of humanlikeness. Figure 5.2 below plots the directions and eigenvector loadings for principal components 1 and 2 on this data.

Figure 5.2

Principal Components 1 and 2 of the Nine Rating Metrics Presented to Participants



Given the differences in participant rating behavior on the nine rating categories, I collapse the rating categories into a different set of dimensions than in Experiment 1 and Experiment 2, shown below in Table 5.1. Although there are more evaluation categories than Experiments 1 and 2, the evaluation categories remain thematically similar across studies as well as within literature that reports the relationship between these social evaluations, described in Chapters 1, 3, and 4. This suggests that while different participant populations interpret the same rating metrics differently, they are broadly consistent in what they consider the social evaluations to mean.

Table 5.1
The Nine Rating Categories in This Study Collapsed into Six Social Evaluation Dimensions

Clustered Trait Ratings	Reduced Dimension
Professionalism, Intelligence	Status
Appropriateness	Appropriateness
Trustworthiness	Trustworthiness
Friendliness	Friendliness
Naturalness and Authenticity	Naturalness
Specificity, Vividness	Social Presence

Bayesian Regression of Effects on Listener Evaluations of Synthetic Voices

I used Bayesian zero-one inflated beta (ZOIB) regression to model the data and designed the model based on my research questions, the findings from Experiments 1 and 2, and k-fold cross-validation. I modeled the simple effects of (ING) variation, belief that a voice was a computer, broadcast context, participant gender, participants' voice AI use, speaker gender, participant gender, and talker.

I included two-way interactions of participant and speaker gender, participant voice AI use and belief that a voice was a computer, participant gender and (ING) variation, speaker gender and (ING) variation, and broadcast context and (ING) variation.

I also included two three-way interactions in the model. I modeled the interaction of (ING), belief that a voice was a computer, and broadcast context; I also modeled the interaction of (ING), belief that a voice was a computer, and speaker gender.

I encountered some indices of convergence problems when I applied the same model architecture from Experiment 2 to the current experiment. Upon inspection of the bulk and tail effective sample sizes and $\hat{R}$ (R-hats) for each estimated effect, I found the convergence issues were due to an overly complex model structure for the current experiment in the alpha and gamma parameters. As shown in Figure 5.3 below, compared to Experiment 2, there were relatively few participant responses in this study at the [0,1] boundary; in other words, participants in this study were less likely than participants in Experiment 2 to evaluate a voice as 0 or 100 on a rating metric. The alpha parameter estimates the likelihood of [0,1] responses in a given dataset, so a lack of participant responses at 0 (as shown in friendliness, status, and trustworthiness) leads to convergence problems in estimating the alpha parameter. The gamma parameter estimates the likelihood that when a response was at the [0,1] boundary the response was 1. Thus, the gamma parameter is predicated on convergence of the alpha parameter and cannot be accurately estimated if alpha is not. To repair this issue in the regression model, I simplified the effect estimates for alpha to the intercept and a by-subject random effect of (1|id). The gamma parameter was estimated as the intercept only.

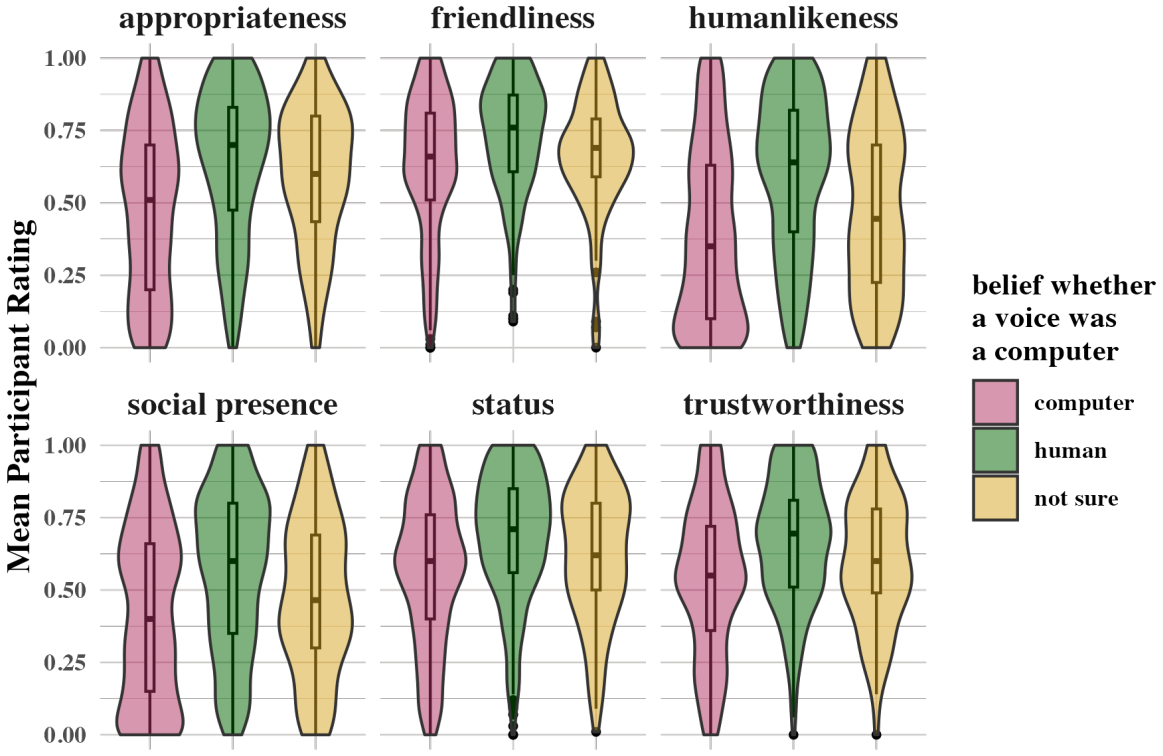
This model formula is identical to Experiment 2 for the mu and phi parameters and differs in only the alpha and gamma parameters. Priors were sampled from the normal distribution with a standard deviation of two. The model syntax is as follows:

```
participant rating ~
    ing use*computer response*broadcast context +
    ing use*computer response*speaker gender +
    voice AI use*computer response +
    ing use*participant gender +
    participant gender*speaker gender +
    talker + (computer response |id),
  phi ~ 1,
  zoi ~ 1 + (1 |id),
  coi ~ 1
```

Separate models were fit for each of the reduced dimensions in Table 5.1 for a total of six models. The clustered trait ratings are averaged to form the response variable. For example, the clustered trait ratings of vividness and specificity are combined into the reduced dimension of "social presence," and participant rating of social presence is the response variable for that model. Like Experiment 2, all evaluation categories in this experiment were fit to separate models using the same model formula.

Figure 5.3

Participant Responses Subset by Belief that a Voice was a Computer



Because of the small number of nonbinary participants in this study (n = 2), I did not include them in the regression model. I also did not include "not sure" responses to the computer response question in the model (n = 77, 8% of participant responses).

Like Experiment 2, there were four estimated effects (individual talker and the effect of female voice) that for every social evaluation had a high degree of statistical uncertainty that they were likely sampled entirely from the prior distribution. These four effects are not included on the effect size plots below to improve readability.

To interpret the estimated effect sizes on participant evaluations for each of the evaluation categories, I evaluate them based on Krushke's (2018) theory of ROPE. Below, I first present the regression model effects that had a meaningful influence on all or most (≥4) of the evaluation categories. I explain these effects together to decrease redundancy in my discussion of the estimated effects for the individual evaluation categories that follows. For every evaluation category, I include an effect size plot that shows all estimated effects in the regression model, ordered from largest to smallest absolute value, with error bars at the 95% upper and lower confidence intervals for each effect.

Influential Effects Across Evaluation Categories

1. Participant Belief About Whether a Voice was Human or Computer

In the regression model, I found that one of the largest effects on participants' evaluations for every evaluation category was belief about whether a voice was a human or computer. Voices that were perceived as computers were rated lower across the board, as illustrated in Figure 5.3 above. Additionally, compared to the young adult participants in Experiment 2, older adults were less accurate at identifying whether a voice was human or computer. In the current study, 60% of

participant responses accurately identified the voices as synthetic, while younger adults had 69% identification accuracy when they listened to the exact same talker stimuli.

Table 5.2
Proportion of Responses to the Post-survey Question Asking Whether Each Voice Was Human, Computer, or Unsure

Participant Response	Count	Percent of total observations
Computer	559	60%
Human	296	32%
Not sure	81	8%

Note. Each participant answered the question four times, once for each voice they listened to.

2. Voice AI Use Frequency

Another primary influential effect was participants' voice AI use habits. Participants who rarely or never use voice AI technology rated the voices lower. This effect was one of the largest for every social evaluation category except for naturalness. There might be an effect of voice AI use on participants' social evaluations of naturalness, but this effect was not estimated with confidence in the model. For every evaluation category in this study, infrequent voice AI users gave lower ratings to voices in general.

3. Broadcast Context

For four of the six evaluation categories in this study there was a large effect of broadcast context on participant ratings. Voices that read the news broadcast were rated as more natural, more trustworthy, more appropriate for broadcasting, and as higher social status. There was not a simple effect of broadcast context for friendliness and social presence; instead, there was a two-way interaction between belief that a voice was a computer and broadcast context, where voices that were perceived as computers were rated as less friendly and had lower social presence when they read the news.

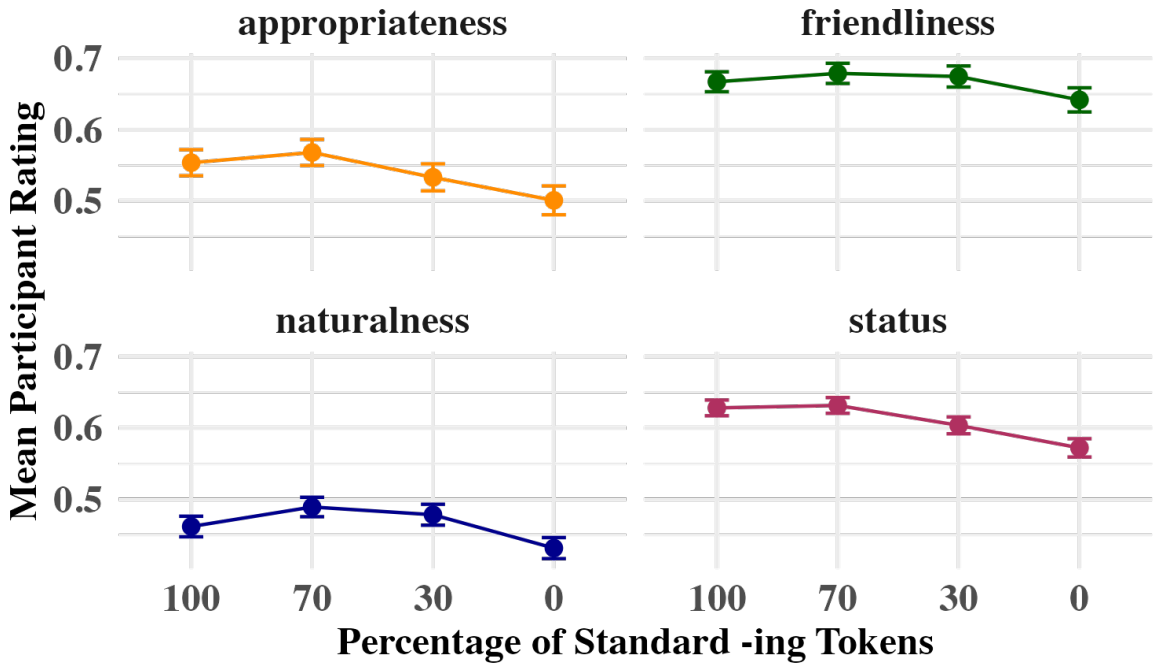
3a. Two-Way Interaction Between Perception of a Voice as a Computer and Broadcast Context.

For all evaluation categories I found a small but statistically reliable effect of a two-way interaction between perception of a voice as a computer and the news broadcast condition. Participants' evaluations of voices they perceived as computers were evaluated lower in the news broadcast condition compared to the radio DJ condition, where there was a slightly smaller difference between perceived human and perceived computer voices.

4. (ING) Variation

There was also a simple effect of (ING) variation on social status, appropriateness, naturalness, and friendliness evaluations. But within those categories, not every amount of (ING) variation was influential. Figure 5.4 illustrates the graphical trends of (ING) variation on the evaluation categories. In the next section, (ING) variation will be discussed in more detail for each of the evaluation categories where it was relevant.

Figure 5.4
Evaluation Metrics Where the Simple Effect of (ING) Was Influential on Participant Ratings



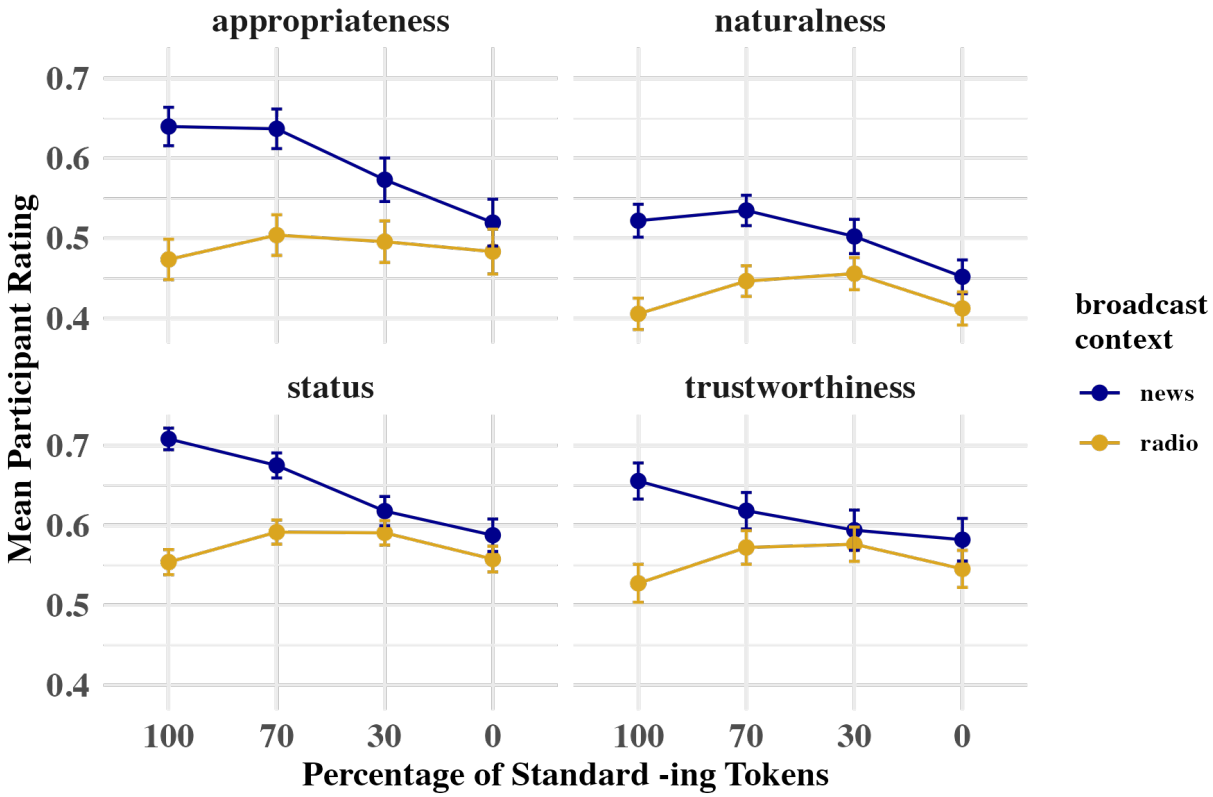
I also found a large effect of a two-way interaction between (ING) variation and broadcast context for the evaluation categories of appropriateness, social status, naturalness, and trustworthiness, as shown in Figure 5.5. There was also a statistically reliable effect of this two-way interaction on social presence and naturalness, but this effect was comparatively small.

5. Three-Way Interaction Between (ING), Belief that a Speaker was a Computer, and Female Voice

For every evaluation category except naturalness, there was a three-way interaction effect between 100% standard *-ing* use, belief that a voice was a computer, and speaker gender. When participants believed a female voice was a human and used exclusively the standard *-ing* variant, they evaluated that voice as less socially present, less friendly, less trustworthy, less appropriate for broadcasting, and of lower social status than male voices that were perceived as human and used exclusively *-ing*. Meanwhile, when participants perceived the voice as a computer, their social evaluations of (ING) variation were not mediated by gender. In the section below I explore effects on individual evaluation categories and provide mean participant ratings plots that illustrate the interaction of gender, perception that a voice was a computer, and (ING) use.

Figure 5.5

Interaction Effect of Broadcast Context on Participants' Evaluations of (ING)



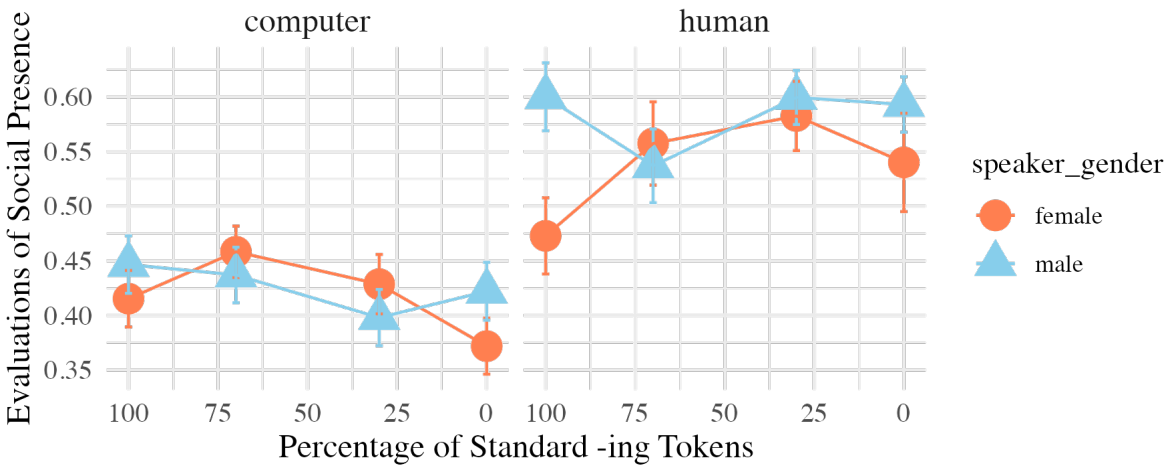
Evaluation Category-Specific Influential Effects

Social Presence

The largest influence on participant evaluations of a voice's social presence (that is, how much they felt the voice was speaking to them specifically and how vividly they could imagine the person talking) was participants' voice AI use habits. In general, participants who use voice AI rarely or never evaluated the voices as less socially present overall. Belief that a voice was a computer was the second most influential effect. I also found a two-way interaction between speaker gender and exclusively standard *-ing* use. (ING) variation was influential on participants' ratings of female voices, but male voices who used different rates of *-ing* were not rated substantially differently from each other. But, participants' evaluations of social presence also had an interaction between (ING), gender, and belief that a voice was a computer. When

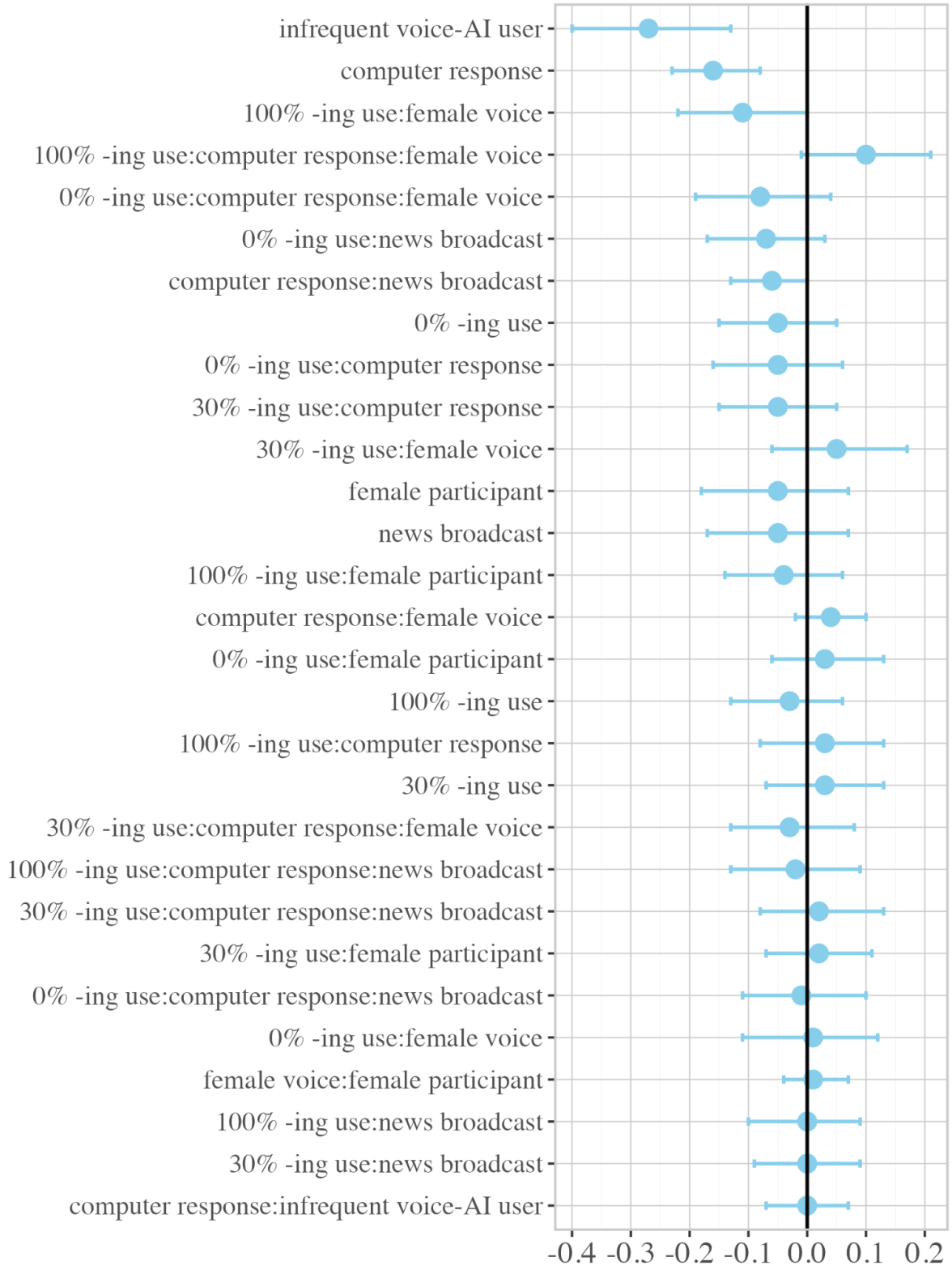
participants believed the female speakers to be human, they rated them as less socially present when they used either only *-ing* or only *-in* compared to male voices. Figure 5.6 below plots the three-way interaction between (ING) variation, speaker gender, and participants' belief about whether a voice was a human or computer.

Figure 5.6
Three-Way Interaction Between (ING) Use, Speaker Gender, and Belief a Voice Was a Computer



I also found a small two-way interaction between broadcast type and belief that a voice was a computer, where voices that were perceived as computers were evaluated as less socially present when reading the news compared to perceived computer voices that read the radio DJ broadcast script. On the next page, Figure 5.7 plots the effect sizes of all estimated effects and interactions in the social presence model.

Figure 5.7
Estimated Effects and Interactions on Evaluations of Social Presence



Naturalness

The most influential effect on participant evaluations of naturalness was participants' perception that a voice was a computer. Voices that read the news were perceived as more natural than voices that read the radio DJ script. I also found a simple effect of nonstandard *-in* variation on participant ratings of naturalness, where voices in the 0% *-ing* condition were evaluated as the least natural, and voices that used 30% *-ing* were evaluated as comparatively more natural. There was not a simple effect of 100% *-ing* use on participant evaluations of naturalness, but it appears to be evaluated more negatively than variable use of *-ing* and *-in'*. In other words, participants considered a mix of *-ing* and *-in'* to sound more natural than exclusively standard or exclusively nonstandard variable use. But, when voices that read the news used exclusively the *-ing* variant, they were evaluated as comparatively more natural than other levels of (ING) use.

I also found a three-way interaction between (ING) variation, computer response, and broadcast context, plotted in Figure 5.8. While perceived human voices are evaluated as similarly natural regardless of their amount of (ING) use in the news broadcast condition, perceived computer voices are considered most natural when they use higher rates of standard *-ing*. In the radio DJ condition, perceived human voices are evaluated as the most natural when they use mostly the nonstandard *-in'* variant due to a combined influence of the simple effect of 30% *-ing* use and a two-way interaction between 30% *-ing* use and radio DJ broadcast. Like social presence, I find a small interaction effect between belief that a voice was a computer and broadcast type. Figure 5.9 below plots all the effects and interactions estimated in the regression model.

Figure 5.8

Three-way Interaction of Broadcast Context, Computer Response, and (ING) Variation

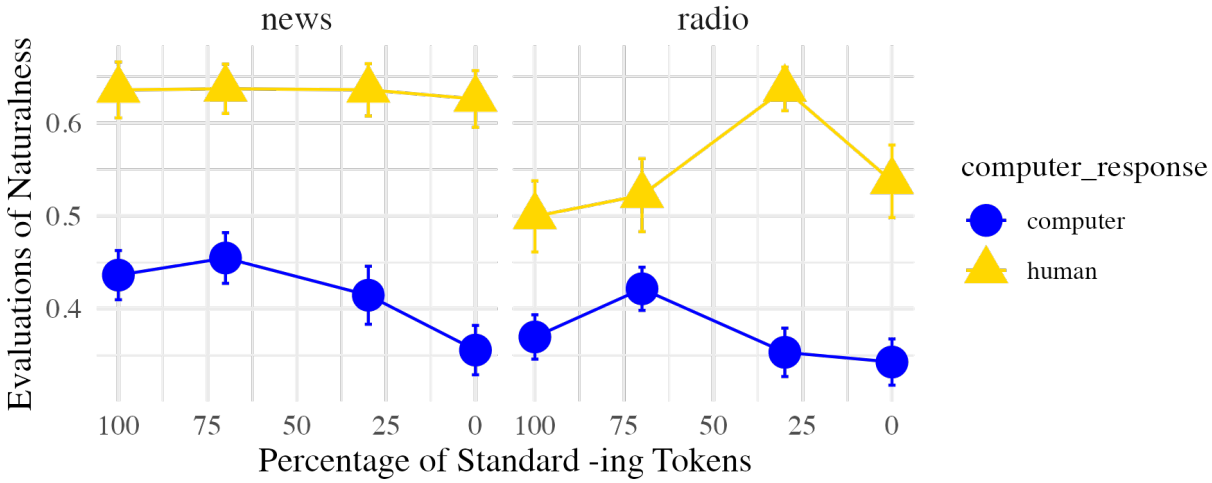
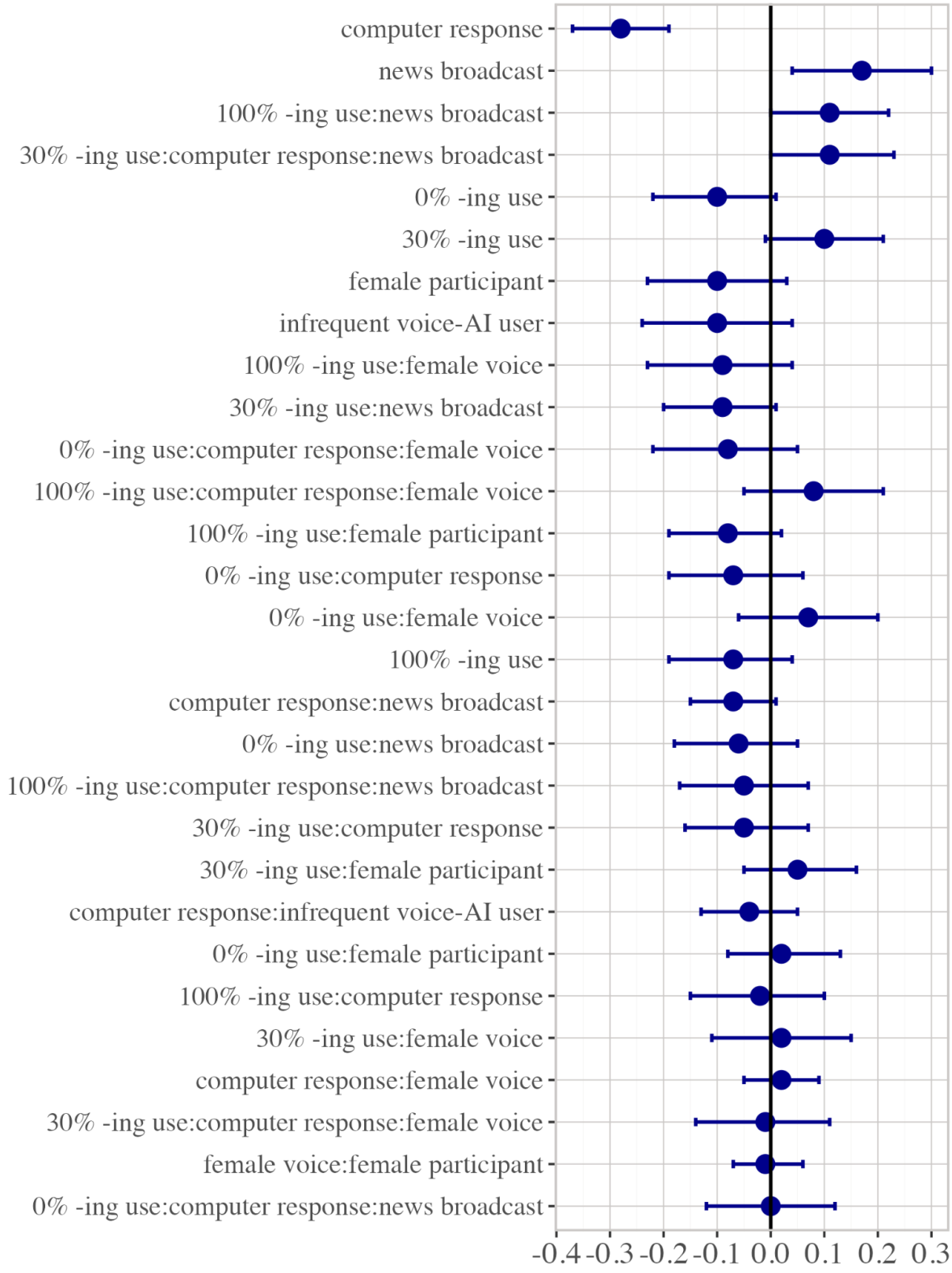


Figure 5.9

Estimated Effects and Interactions for Participant Evaluations of Naturalness



Friendliness

The two largest effects on friendliness evaluations were perception of a voice as a computer and exclusively nonstandard *-in'* use, which was evaluated by listeners as less friendly overall. I also found that infrequent voice AI users evaluated voices as less friendly, and a small two-way interaction between belief that a voice was a computer and the news broadcast condition.

As mentioned above, there was a three-way interaction effect between speaker gender, perception of a voice as a computer, and (ING) variation, shown in Figure 5.10. 100% *-ing* use was positively evaluated by listeners, while the 30% *-ing* three-way interaction was evaluated comparatively negatively by listeners. Friendliness was the only evaluation category where two different rates of (ING) variation were statistically reliably influential on participant ratings in this three-way interaction; for the other five evaluation categories in this study, only the 100% *-ing* condition was estimated to be statistically reliable. Below, Figure 5.11 plots all the estimated effects and interactions in the regression model and their effect sizes, sorted in descending order by absolute value of the effect.

Figure 5.10
Three-Way Interaction Between (ING) Use, Speaker Gender, and Belief a Voice Was a Computer

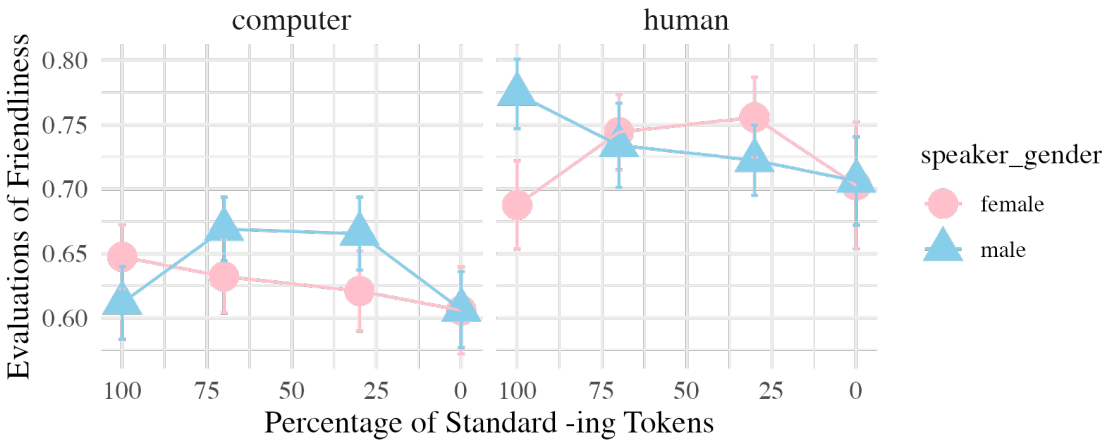
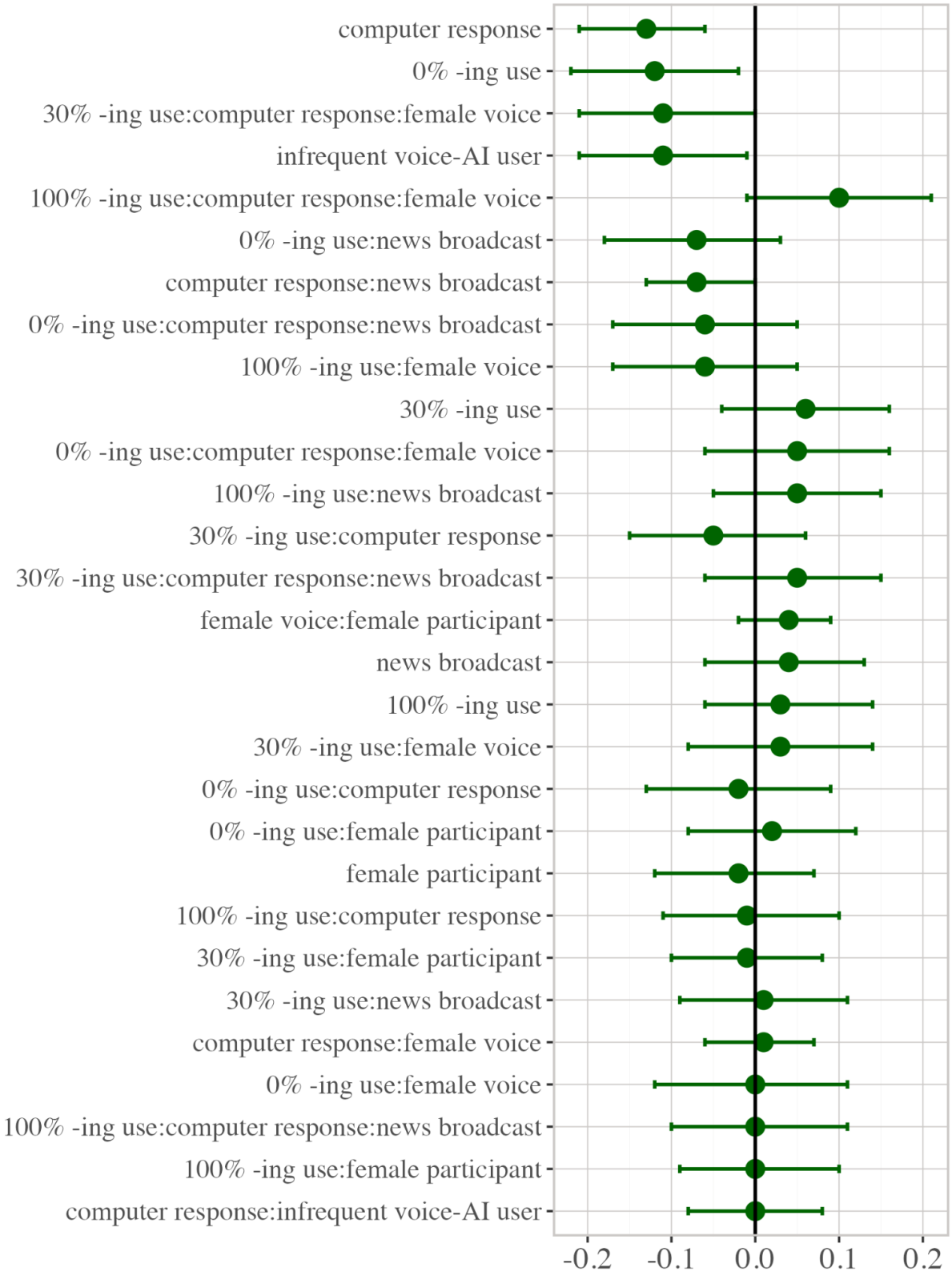


Figure 5.11

Estimated Effects and Interactions on Participant Evaluations of Friendliness



Trustworthiness

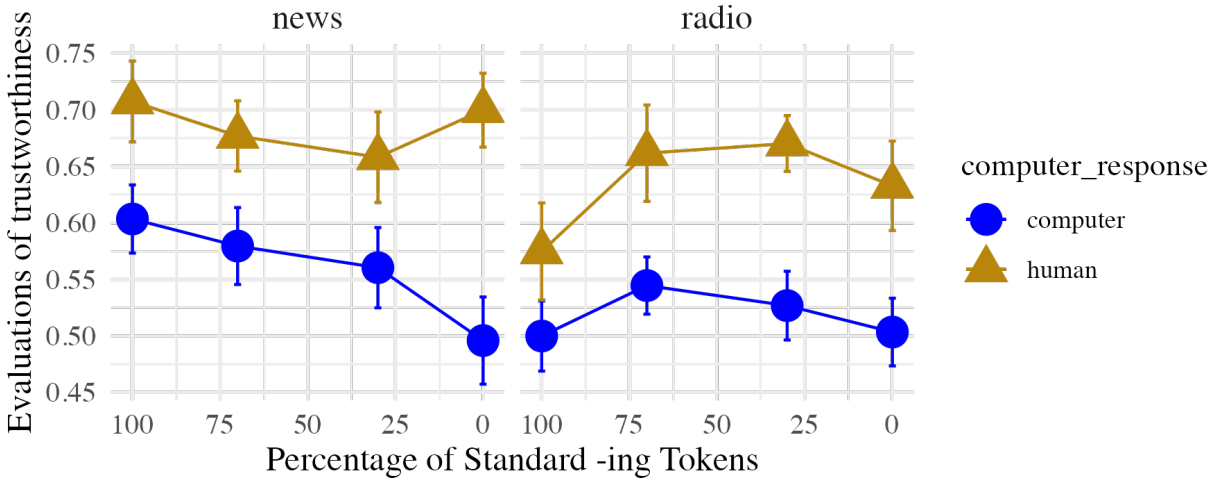
The largest effect on trustworthiness was the perception that a voice was a computer. Participant infrequent voice AI use also had a negative effect on evaluations of trustworthiness. There was a simple effect of news broadcast, where voices were rated as more trustworthy if they were reading the news.

In the trustworthiness social evaluation category, I find many influential effects of *-ing*. There was a large two-way interaction between perception 100% standard *-ing* use and the news broadcast condition, where voices were rated as more trustworthy if they used 100% *-ing* while reading the news. There was also a two-way interaction between exclusively standard *-ing* use and female voice, where female voices were rated as relatively more trustworthy when they used only *-ing* compared to other conditions. I also found a two-way interaction where participants rated 0% *-ing* use as even less trustworthy when they believed that a voice was a computer.

There were two statistically reliable three-way interactions between belief that a voice was a computer, (ING) variation, and broadcast type that influenced participant ratings of trustworthiness, shown in Figure 5.12 below. In the news broadcast condition, participants' evaluations of voices they perceived as humans diverged from computer voices when they used exclusively nonstandard *-in'* variation. Perceived computer voices were evaluated more negatively for every increase in nonstandard *-in'* use in the news broadcast, but perceived human voices were evaluated relatively similarly in the news broadcast condition based on their amount of -i*n'* use. Yet, in the radio DJ broadcast condition, participant ratings of perceived computer and human voices followed similar trends for (ING) variation evaluations, where 100% standard *-ing* use was evaluated as less trustworthy than other conditions.

Figure 5.12

Three-Way Interaction of Perception of a Voice as a Computer, Broadcast Context, and (ING) Variation



Note. The 100% *-ing* condition and 30% *-ing* condition were estimated as statistically reliable by the model.

I found another three-way interaction on social evaluations of trustworthiness speaker gender, between perception of a voice as a computer, and exclusively standard *-ing* use, as described above. This interaction is illustrated in Figure 5.13 below. Finally, Figure 5.14 illustrates all estimated effects and interactions in the trustworthiness model.

Figure 5.13

Three-way Interaction of Gender, Perception of a Voice as a Computer, and (ING) Variation

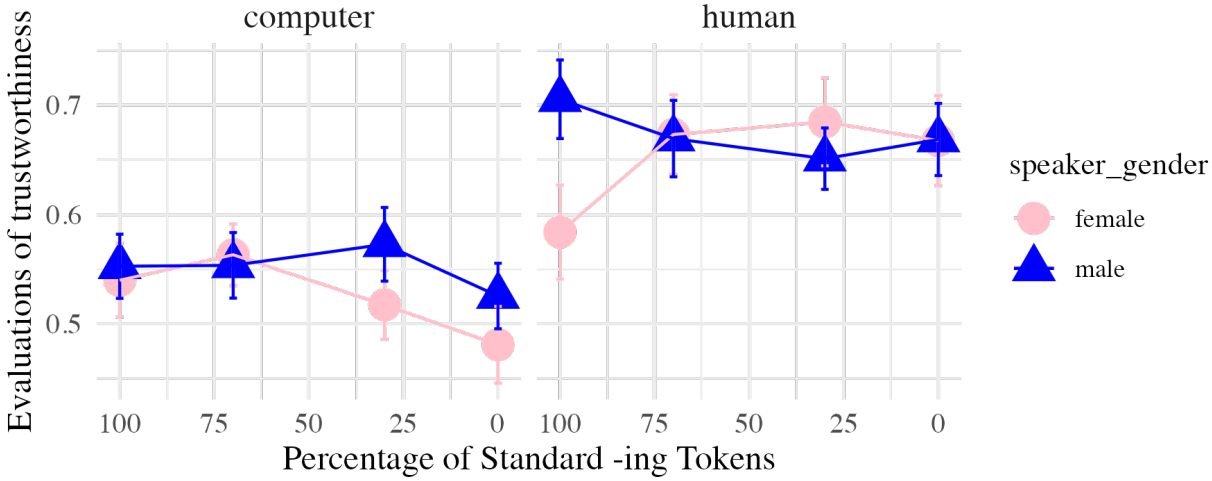
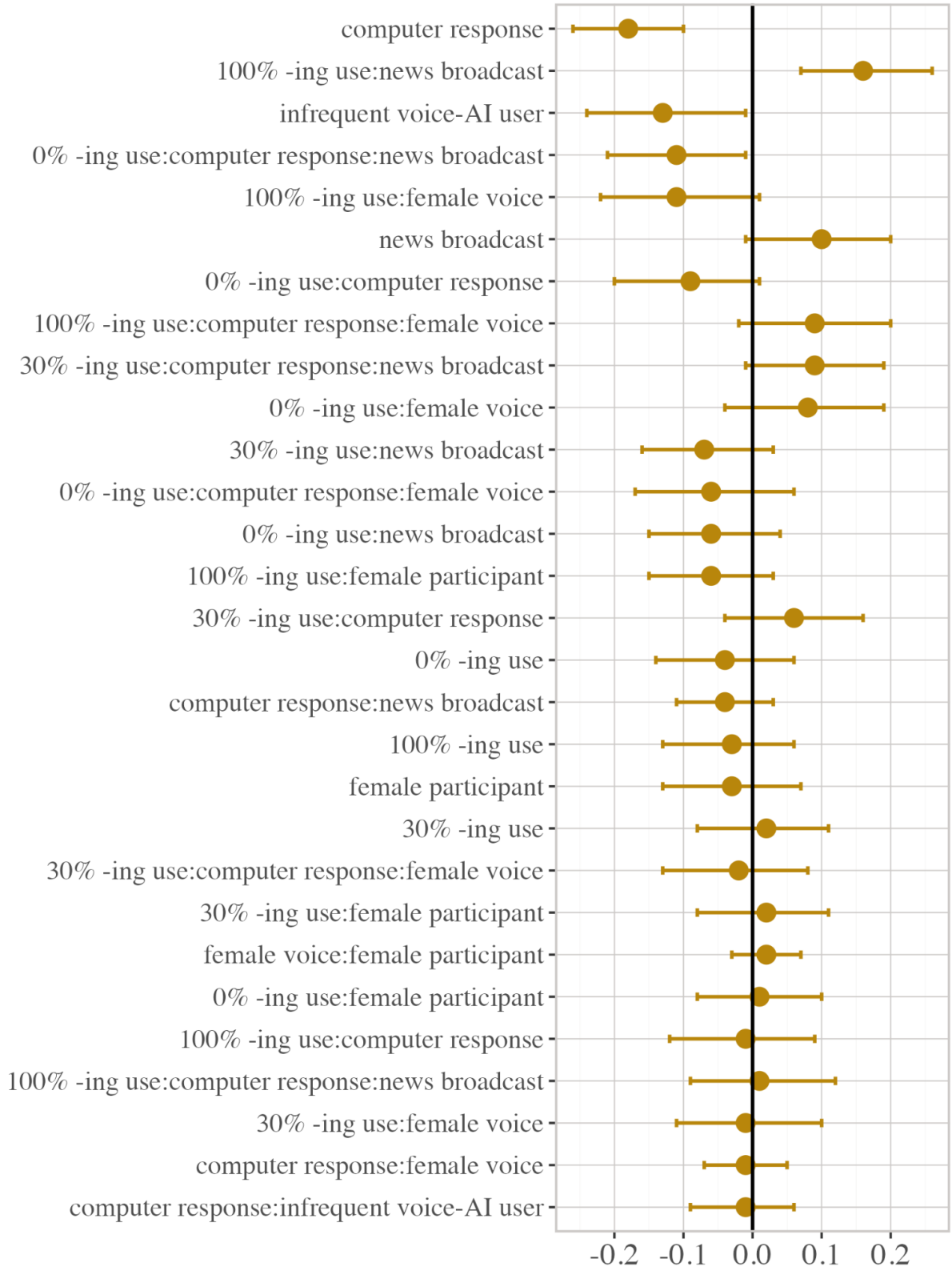


Figure 5.14

Estimated Effects and Interactions on Participant Evaluations of Trustworthiness



Appropriateness

The most influential effect on participant evaluations of appropriateness was broadcast condition. Voices that read the news were considered more appropriate. This is interesting because participants who were in the news broadcast condition answered the question "Is this voice good for news broadcasting?" while participants in the radio DJ condition were asked "Is this voice good for a radio DJ broadcast?," so this effect was likely not caused by participants' interpretation of the meaning of "broadcasting" since it was explicitly contextually defined in the question they answered. The next most influential effect was perception of a voice as a computer, followed by the effect of participants' voice AI use frequency. Both simple effects led to lower ratings of appropriateness.

Evaluations of appropriateness were clearly affected by (ING) variation. I find a negative effect of exclusively nonstandard *-in'* use on participant evaluations of appropriateness. There is also a very clear influence of an interaction between (ING) variation and broadcast context on participant evaluations of appropriateness. Voices that used exclusively *-ing* were rated the highest, and with any additional use of nonstandard *-in'*, those voices were rated lower. This effect applied regardless of whether a voice was perceived as a human or computer and can be seen in the "appropriateness" portion of Figure 5.5.

Finally, I find the three-way interaction between speaker gender, perception of a voice as a computer, and (ING) variation, as described above and pictured here in Figure 5.15. As with the other social evaluations, it appears that female voices perceived as human were rated negatively compared to male voices perceived as humans; but this effect did not apply to computer voices, which received similar evaluations of appropriateness regardless of their

gender but instead primarily based on their varying use of (ING). All estimated effects and interactions are plotted in Figure 5.16 below.

Figure 5.15

Three-Way Interaction of Gender, Perception of a Voice as a Computer, and (ING) Variation

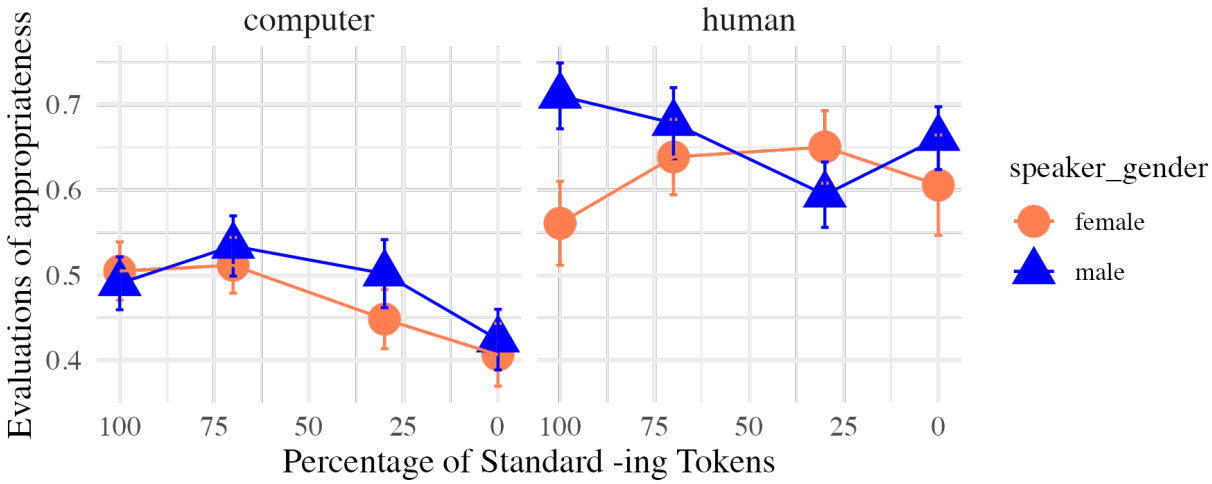
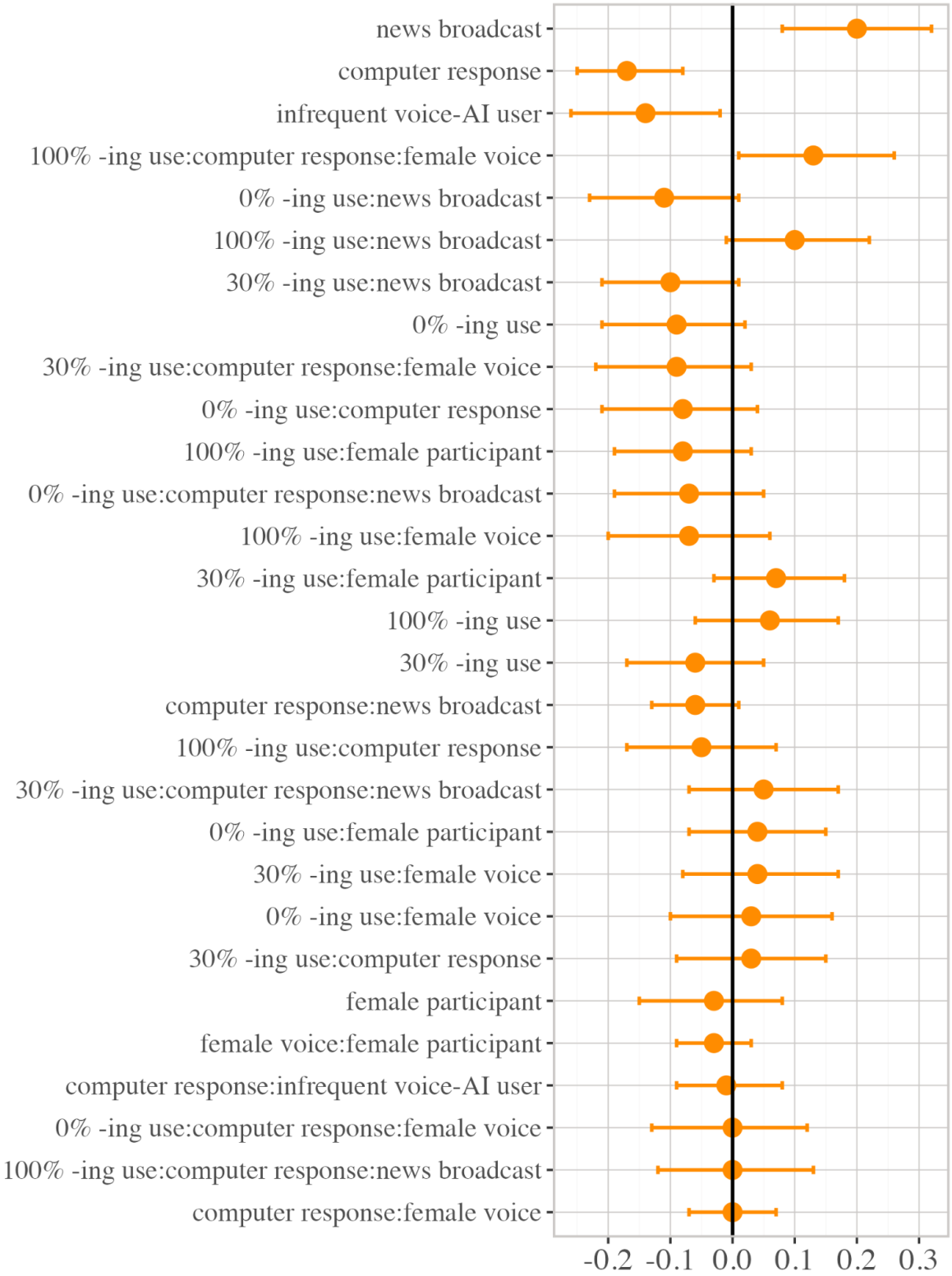


Figure 5.16

Estimated Effects and Interactions on Participant Evaluations of Appropriateness



Social Status

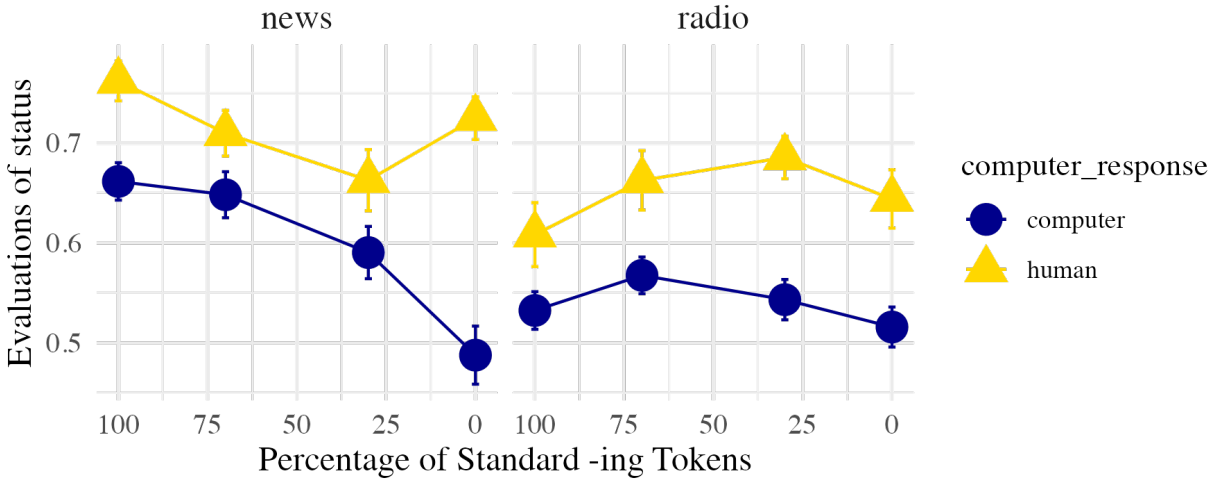
Like other evaluation categories, voices perceived as computers and participants who use voice AI infrequently evaluated the voices in this study more negatively. Similar to the appropriateness evaluations, I find a simple effect of news broadcast condition, where voices were rated as having higher social status when they read the news.

(ING) variation played an important role in participants' evaluations of social status. I found that a simple effect of exclusively *-ing* use had a positive effect on listener evaluations, while exclusively *-in'* use had a negative effect. This effect was amplified in the news broadcast condition. I found a statistically reliable effect where voices that used exclusively *-ing* in the news broadcast received a further boost in evaluations of social status, as pictured in Figure 5.5 above. On the other hand, voices that used mostly nonstandard *-in'* were penalized more in the news broadcast condition than in the radio DJ one. Further, there was a two-way interaction between exclusively *-in'* use and perception that a voice was a computer. When participants believed a voice was a computer, they evaluated 0% *-ing* variant use more negatively than when they believed that voice was a human.

I found two three-way interactions in participants' social evaluations of status. First, when a voice used exclusively the nonstandard *-in'* variant while reading the news and participants believed a voice was a computer, it was evaluated especially negatively, as shown in Figure 5.17. This interaction, combined with the two-way interaction between 0% *-ing* and computer response as well as the simple effect of 0% *-ing* use, led to substantially more negative evaluations of social status for perceived computer voices.

Figure 5.17

Three-way Interaction Between (ING) Variation, Broadcast Context, and Belief That a Voice Was a Computer



I also found a three-way interaction between speaker gender, perception of a voice as a computer, and (ING) variation, illustrated in Figure 5.18. The pattern shown here is similar to what I found for the same three-way interaction on evaluations of appropriateness described above. Lastly, the effects and interactions estimated in the social status regression model are plotted in Figure 5.19.

Figure 5.18

Three-way Interaction of Gender, Perception of a Voice as a Computer, and (ING) Variation

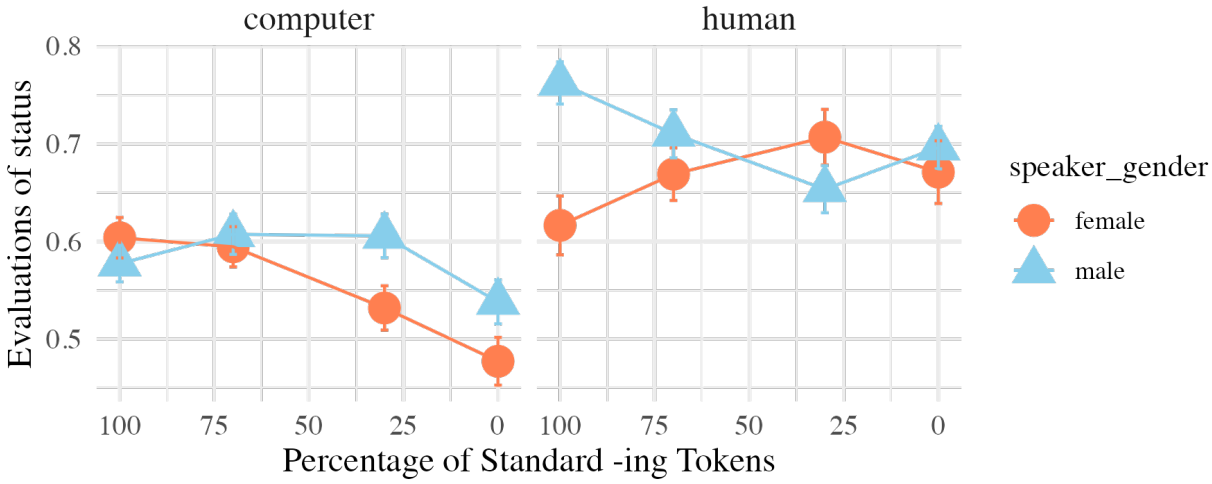


Figure 5.19

Estimated effects and Interactions on Participant Ratings of Social Status

Discussion

In the current study, I found that (ING), belief about whether a voice was a computer, broadcast context, speaker gender, and participants' own voice AI use habits played a role in participants' evaluations of the four synthetic voices they heard.

I found a simple effect where speakers who read the news broadcast were rated more positively on most evaluation metrics in this study. There was not a simple effect of broadcast on the friendliness and social presence metrics, but there was a two-way interaction between broadcast context and perception of a voice as a human. I suggest that, like Experiment 2, the simple effect of broadcast context could be because participants associate news broadcasts with prestige (Bayley & Zapata, 1994). As mentioned in Chapter 1, positive evaluations of a speaker's social prestige are also tied to other positive traits like attractiveness, moral goodness, and intelligence (Gordon, 1997; Milroy, 2002), which may explain why the "news broadcasters" in this study were evaluated more positively overall.

Participants' voice AI use habits also influenced how they rated voices in this study. Interestingly, I did not find an interaction between participants' voice AI use frequency and their social evaluations of voices they perceived as computers. Instead, participants who use voice AI rarely gave lower ratings in this study in general. One interpretation for this finding might be that participants, particularly older adults, who rarely use voice AI technology also rarely use other kinds of digital media, and might not have strong opinions about social evaluations of digital media more generally. In fact, the influence of voice AI use frequency was largest for the social presence rating category—the questions that asked how well the participant could imagine the person speaking. In a virtual experiment completed online, participants who use digital media rarely might simply be unaccustomed to making social judgments through digital media.

Another influential effect in this study was when participants believed a voice was a computer, they evaluated that voice more negatively in every social evaluation category. Regardless of any other social characteristic of the voice like gender or (ING) use, speakers were evaluated negatively simply by being perceived as computers. But listeners' perception of voices as computers also interacted with the broadcast context for some evaluation categories. Participants rated voices as even less natural, less friendly, less socially present, and less appropriate when they perceived them as computers and heard them in the news broadcast condition.

(ING) Variation

The primary question in this experiment aims to understand how the sociolinguistic variable (ING) is perceived by older adults when spoken by a computer voice. I found a simple effect of (ING) variation on participant ratings of social status, appropriateness for broadcasting, friendliness, and naturalness. But, not every level of (ING) variation had a statistically reliable influence on participant ratings in these social categories. Instead, exclusively nonstandard *-in'* use was the most predictive of lower participant ratings on naturalness, friendliness, appropriateness for news broadcasting, and social status. Exclusively standard *-ing* use also had a predictive effect on participants' evaluations of social status, but this was the only evaluation category where I found a simple effect of 100% *-ing* use was influential on participant ratings.

Beyond the simple effects of exclusively standard and exclusively nonstandard (ING) variation, there were also many interactions between (ING) variation and other variables in this study. (ING) variation and broadcast context interacted in participants' evaluations of status, appropriateness, trustworthiness, and naturalness. For each of these evaluation categories, the statistically reliable interaction was between exclusively standard *-ing* use and news broadcast. I

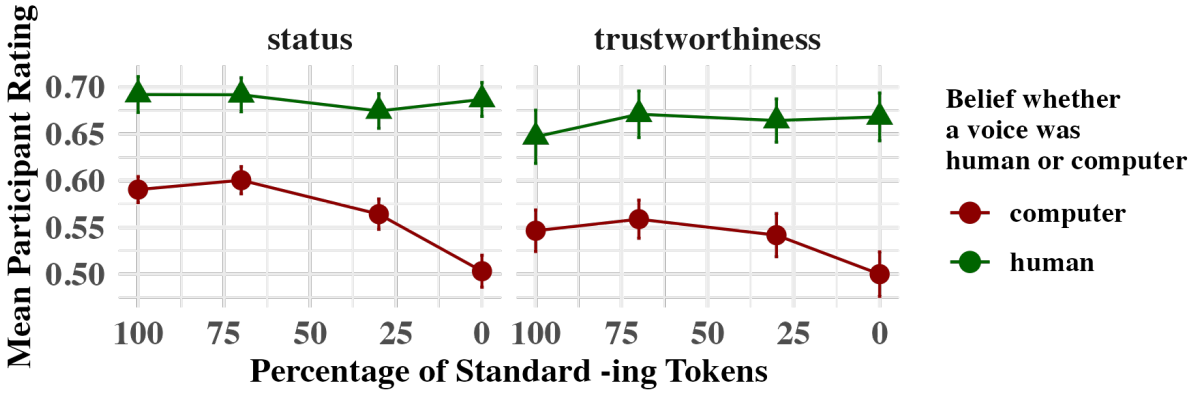
suggest this interaction is present for naturalness because exclusively standard language fit listeners' expectations for a news broadcasting speaking style. For appropriateness and status, nonstandard *-in'* use was also predictive of participants' lower ratings of voices in the news broadcast. Combined with the simple effect of (ING) above, I suggest that listeners' expectations for (ING) variation are strongest in their ratings of appropriateness and social status.

I only found a statistically reliable two-way interaction between (ING) variation and belief that a voice was a computer for evaluations of trustworthiness and status, where exclusively nonstandard *-in'* use was evaluated as less trustworthy and as less socially prestigious, as shown in Figure 5.20 below. It appears that participants applied their expectations for standard language norms to computer voices more than to human voices. Participant evaluations of nonstandard *-in'* use led to increasingly negative social evaluations of computer voices, but perceived human voices appear less affected by (ING) variation.

I expected there would be more two-way interactions between (ING) and computer response. Instead, (ING) and belief that a voice was a computer more often interacted with a third interaction term—either speaker gender or news broadcast.

Figure 5.20

Evaluations of status and trustworthiness subset by belief that the voice was a computer



(ING) Variation, Belief That a Voice Was a Computer, and Broadcast Context

I found a three-way interaction between (ING) variation, belief that a voice was a computer, and broadcast condition for the social evaluation categories of naturalness, trustworthiness, and social status. For naturalness, voices perceived as human were considered more natural when they used mostly nonstandard *-in'* use in the radio DJ broadcast. This suggests the influence of the colloquial style of radio broadcasts as described by Bayley and Zapata (1994) leads participants to evaluate nonstandard use as more natural, but only when they think the nonstandard usage was by a human voice. Nonstandard *-in'* use also interacted with participants' evaluations of voices in the news broadcast condition, where voices that were perceived as computers were considered markedly less trustworthy than voices that were perceived as human. I found a similar pattern for participant evaluations of social status. When perceived human voices used nonstandard *-in'* in the news broadcast condition, they were evaluated as having much higher social status than computer voices that used nonstandard *-in'* while reading the news. I suggest that these three-way interaction effects are primarily reflective of whether participants consider it acceptable for computer voices to nonstandard speech variants, and that they are especially negatively perceived when they use nonstandard speech in formal speech contexts compared to perceived human talkers.

Speaker Gender, Perceived Humanness of a Voice, and (ING)

Like Experiment 2, I do not find a simple effect of speaker gender on any evaluation category. In other words, it is not that being a female speaker results in lower social evaluations, but how female speakers' social traits—including sociolinguistic variant use and apparent humanness—influence participants' social evaluations.

In all the social evaluation categories except naturalness, I find a three-way interaction between speaker gender, belief that a voice was a computer, and 100% *-ing* use. Though the patterns vary slightly across evaluation categories, this effect appears to stem from different evaluations of male and female voices that were perceived as human. When perceived human voices were female, they were rated substantially lower than men when they used exclusively standard *-ing* use. Meanwhile, participant evaluations of voices that were perceived as computers were not mediated by judgements of speaker gender.

In sociolinguistic studies, it has been shown that women are evaluated more negatively than men when they use the nonstandard variant (Gordon, 1997), so it is surprising to find in the current study that standard variant use is evaluated lower in women than it is in men. I did not include a four-way interaction in the model that would evaluate if there was an interaction between participant and speaker gender, computer response and (ING) because I considered that overfitting for this data considering the size of the participant pool. However, I did include a two-way interaction for participant and speaker gender in the model, and this effect was estimated near zero for every social evaluation category. I also included a two-way interaction between participant gender and (ING) use evaluation. It is possible that an interaction between participant gender and 100% *-ing* use might exist for naturalness, friendliness, trustworthiness, and appropriateness, but these effects were not estimated with confidence by the model. Moreover, a four-way interaction between participant and speaker gender, computer response, and standard (ING) might exist and in fact could be predictive in the model, but another study with a larger sample size would be more appropriate to estimate this effect.

Implications for Theories of Sociolinguistics

My findings of a simple effect of exclusively nonstandard *-in'* use are consistent with studies in sociolinguistics that listeners have increasingly strong evaluations of a speaker's personal qualities as their rate of nonstandard language use increases. The effect of exclusively standard *-ing* use in the social status and appropriateness evaluation categories is additionally consistent with the findings by Labov et al. (2011). Further, the two-way interaction between (ING) and broadcast context is consistent with studies from sociolinguistics that find that participant evaluations of (ING) are context specific (Campbell-Kibler, 2008).

The fact that exclusively nonstandard *-in'* use was evaluated negatively in the friendliness social evaluation category is situated between two different interpretations of nonstandard *-in'* in sociolinguistics studies. Increased levels of *-in'* variation have frequently been reported as related to increased friendliness (Campbell-Kibler 2007; Labov et al., 2011). But, Campbell-Kibler (2008) found that participant evaluations of nonstandard *-in'* varied with their perception of other indexical characteristics of the speaker; in her study, she found that nonstandard *-in'* use was evaluated by listeners as inauthentic and patronizing. It might be the case that the voices in this study had other voice characteristics that led to 0% *-ing* use being perceived as inauthentic and therefore less friendly.

The lack of an effect of a two-way interaction between participant gender and (ING) variation is interesting because this was one of the most influential effects on participant evaluations in Experiment 2. I also did not find an interaction between participant and speaker gender. These two effects would be expected to be influential in a human sociolinguistic study. I propose that the lack of these effects could be because older adults are less sensitive to computers as a social category and therefore do not have gender-based expectations for computer

voices. Another interpretation might be that older adults are less critical of nonstandard variation as they age (Labov, 1972; Eckert, 2017). But, a future study directly comparing the effect of participant age and (ING) variation would provide important insight into whether this effect is also generalizable to younger adults.

Implications for Theories of Human-Computer Interaction

In the current study, I found that broadcast context and its interactions with computer response or (ING) were more influential on participants' ratings than a simple effect of (ING) or a two-way interaction between (ING) and speaker gender. These findings contrast with the sociolinguistic perceptions by younger adults in Experiment 2. I propose that older adults are more sensitive to speaking style changes in perceived computer voices than younger adults, while younger adults have started to make evaluations about the social characteristics of a computer voice—in other words, increased exposure to computer voices across the lifespan leads listeners to transfer social evaluations of gender and (ING) from human voices to computer ones.

From the perspective of theories of human-computer interaction, these findings provide support for the idea that social scripts for interactions with social computers become entrenched as listeners have more exposure to social computers (Gambino et al., 2020). Since older adults have not been immersed in social language use by computer voices compared to younger adults and interact with it more rarely, their evaluations are based more on expectations for speaking style and whether they thought a voice was a computer than for social-indexical traits used by computer voices.

Limitations and Future Directions

One limitation of completing sociolinguistic studies on older adults is that it can be challenging to contact older adult participants, particularly through digital media (Barnard et al.,

2013). The findings of the present study might be skewed by the fact that the older adults recruited in this study use the internet to complete online surveys and therefore might be more proficient users of other kinds of technology, including voice AI. Future work can recruit participants for a traditional in-person study and specifically target older adults who are not frequent internet users.

A future study can more closely examine how digital media use frequency by older adults, particularly social technologies like social media and chatbots designed for social-emotional support, might influence their evaluations of sociolinguistic variants when used by computer voices. As pointed out by Eckert (2017), there is a dearth of sociolinguistic studies on elderly adults in human interactions; this suggests a greater need for sociolinguistic studies on elderly adults with computer interactants. One way to better understand older adults' social attitudes towards computer voices might be through a qualitative study that probes participants' beliefs about the social role of computers in a more open-ended way. This can be explored in future research.

Chapter 6: Conclusion

In this dissertation, I explored how synthetic voices that use the sociolinguistic variant (ING) are perceived by listeners. In three experiments, I investigated how different amounts of nonstandard *-in'* use, participant characteristics of age and gender, and perception of a voice as a computer influenced participants' evaluations of synthetic voices. I examined whether sociolinguistic evaluation patterns for (ING) variation are similar across human and synthetic voices and what social cues are necessary for linguistic variation to be considered socially meaningful. I found that listeners were sensitive to (ING) variation in synthetic speech, and in some categories evaluated it similarly to its social meaning in human speech. However, I never found that sociolinguistic perception of (ING) in computer voices was completely opposite to its social evaluations in human speech. In other words, depending on the social evaluation, speaking context, and interlocutor traits, the social meaning of (ING) was either directly transferred to computer voices or was not influential; for example, I never found that exclusively standard *-ing* use was evaluated as having lower social status in computer voices. Another primary influence on listeners' perceptions was whether they believed the voice to be a computer. (ING) variation and perception of a voice as a human or computer had strong influential effects across all three studies. Within the three studies, I also found evidence that participant and speaker gender, participant age, and speaking style can interact with how participants perceive (ING) and its social meaning within computer speech.

General Discussion

Understanding the Relationship Between Social Evaluation Survey Questions

One innovation of this dissertation research was the inclusion of principal component analysis (PCA) to better understand how participants interpreted the meaning of the survey

questions. In all three experiments, PCA reflected a close relationship between authenticity, naturalness, and the two social presence questions. The social presence questions were developed by Lee and Nass (2005) to test how "real" a computer voice sounded, so it is reasonable that they would cluster together with naturalness and authenticity. In theory, the "authenticity" of a voice could be interpreted to mean whether the voice is "real" (as in humanlike) or "truthful," but across all three studies, I find authenticity is evaluated synonymously to naturalness.

I also found that the importance of (ING) variation in a particular evaluation category was relevant to the eigenvector loading and direction of Principal Component (PC) 2. Across studies, status and appropriateness remained closest to each other on the y axis, followed by trustworthiness. In the regression models across studies, I found that (ING) variation was most influential on participant evaluations of status, followed by appropriateness and to a lesser extent trustworthiness. This suggests that PC 2 might have related evaluation categories based on the impact of (ING) variation on participant ratings.

The set of reduced dimensions from PCA in Experiment 1 and Experiment 2 are largely similar. The main difference was that participants in Experiment 2 closely related trustworthiness and appropriateness for broadcasting whereas participants in Experiment 1 had considered them separate evaluation categories. One potential explanation for the different grouping is that participants in Experiment 1 were directly comparing human and synthetic voices, and when asked if they preferred human or synthetic voices for broadcasting overall at the end of the experiment, participants overwhelmingly preferred human voices. Participants in Experiment 1 may have taken the question “Is this voice good for news broadcasting?” to mean “Is this [human] voice good for news broadcasting?” and this effect might have been magnified by the blocked design that directly compared synthetic voices against human ones. Participants in

Experiment 2 did not have human voices to directly compare against and also believed some of the synthetic voices were human. This might explain why trustworthiness and appropriateness were grouped together in Experiment 2— since there were no human voices to compare against, participants' general belief that human voices were better for news broadcasting was not influential in their selections of whether a voice was appropriate versus trustworthy.

In Experiment 3, I found that the relationship between rating metrics was finer-grained than in Experiments 1 and 2. Given that the main difference between the participants in Experiments 1 and 2 versus Experiment 3 is participant age, it is possible that older adults interpreted the social evaluation questions differently from younger adults. Another difference between Experiment 3 and the others is how the data was collected. Prolific, the online survey platform I used to recruit older adult participants, has stringent criteria for participants to be allowed to enroll in studies, so participants might be more likely to think closely about the differences in meaning between seemingly similar survey questions. In future work, the underlying cause for the difference in PCA results in Experiment 3 compared to 1 and 2 could be investigated by including survey collection method as an effect in a regression model or by collecting young adult participants through Prolific as well.

Because PCA is exploratory, it requires some interpretation on the researcher's part as to what the relationships between each principal component might represent. It should be noted that the particular words I chose to describe the PCA relationships were my interpretation, but PCA relationships are estimated by a machine learning algorithm that is not influenced by any hypothesis that some characteristics are expected to be related. In other words, it is mathematically clear that some rating metrics covary, and I can mathematically determine which rating metrics covary, but the description of the covariant relationship between those groups is

my interpretation. PCA cannot be used to support a "truth claim" about how listeners interpret the meaning of different social evaluation categories, but this exploratory analysis technique provides insight into how different listener groups might interpret social evaluation questions when applied to computer voices.

Influence of Social Traits on Listener Evaluations

Across studies, (ING) variation was particularly influential in participants' evaluations of social status and appropriateness for broadcasting, similar to what has been found in previous sociolinguistic studies on (ING) (Campbell-Kibler 2009; Labov et al., 2011). Across studies, nonstandard *-in'* was evaluated lower, but the amount of *-in'* use that resulted in the most negative evaluations was not always 0% *-ing*. Instead, I found that either 30% or 0% *-ing* use was rated most negatively, depending on the speaking context and the social evaluation being made. This is interesting compared to previous research on (ING) variation, which found that 0% standard variant use is typically rated much lower than conditions that have at least some instances of the standard variant (Labov et al., 2011; Mechler, 2025). This result suggests that voice naturalness interacts with other sociolinguistic traits. Campbell-Kibler (2008) proposed a similar explanation for why speakers were evaluated as less friendly based on their use of *-in'*, where participants evaluated speakers more negatively when they perceived their use of the nonstandard variant as inauthentic.

In Experiment 1, I found that listeners transferred their social perceptions of (ING) variation onto computer voices, but also that (ING) variation affected listeners' perceptions more when it was used by human voices. I also found this effect in Experiments 2 and 3 when participants believed the voice was a human. In other words, (ING) perception is mediated by participants' *belief* about whether the voice was a human. Depending on the social evaluation,

speaking context, and interlocutor traits, the social meaning of (ING) was either directly transferred to computer voices from human voices or was not influential.

However, this effect was also mediated by age. Younger participants were more likely to consider humans and computers to belong to two different social categories, while older adults' judgments of computer speech were more influenced by style and broadcast context.

A very large difference between my studies on younger adults and older adults was the influence of voice AI use habits. I found one single interaction of voice AI use in Experiments 1 and 2, but voice AI use was one of the largest effects for nearly every social evaluation category in Experiment 3. Even though I did not find an interaction effect of voice AI use, belief that a voice was a computer, and social evaluations in this study, I believe it is likely that there is a connection between voice AI use, or at least exposure to synthetic voices in everyday life, and listeners' perceptions of synthetic voices. However, this must be explored in future research.

In the dissertation studies, I find varying effects of speaker and participant gender on their social evaluations of computer voices. While sociolinguistic studies have historically reported that older women are the most critical of nonstandard variation (Chappell, 2016; Wollum, 2019), I find a different pattern in the present research. In Experiments 1 and 2, participant gender was strongly influential on participants' rating behavior. In Experiment 1, female participants were more likely to give negative social evaluations, which would be considered typical in sociolinguistic studies. In Experiment 2, female participants gave relatively higher evaluations than male participants. Both studies had participant pools of young adults, so participant age was likely not a mediating factor in this effect. I propose two alternative explanations for this finding. First, Experiment 1 used exclusively female voice stimuli. Women have been shown to evaluate other women more negatively when they use nonstandard variation

(Eckert, 2008; Gordon, 1997). In Experiment 2, which includes voices of both genders, I found an interaction between speaker and participant gender where participants gave higher ratings to voices whose gender was different from their own, which likely accounts for the finding that women gave lower evaluations in Experiment 1. Surprisingly, however, female participants' perceptions of different levels of (ING) variation in Experiment 2 did not change how they evaluated the voice. Instead, male listeners followed the pattern predicted by Labov et al. (2011), where every additional instance of nonstandard *-in'* was evaluated more negatively by the listener. Finally, in Experiment 3, I find no effect of participant gender or any interactions of participant gender with other variables. The underlying reason for why I found participant gender effects for undergraduate participants but not older adults is unclear, but one possibility might be that older adults do not judge nonstandard variation as harshly once they have established their social status in the workplace or retirement (Eckert, 2017; Labov, 1972). More research is necessary to understand the influence of gender on (ING) perception and how it relates to the findings of this dissertation.

Limitations

Both a strength and a limitation of this dissertation is the use of cross-splicing for sociolinguistic variation studies. Cross-splicing is a widely accepted sociolinguistic method and is useful in understanding the social role of one particular linguistic variant. The main benefit of cross-splicing is that it produces a highly controlled study but is also somewhat artificial for the same reason. Mechler (2025) argues that cross-splicing studies like Labov et al. (2011) might artificially increase the actual salience of (ING) variation, and Vaughn (2022) showed that cross-splicing itself is perceptible to listeners compared to naturalistic speech, even when executed with great care. These limitations might apply to this dissertation research as well. However,

synthetic speech is inherently artificial and characteristically invariant. Synthetic speech is highly invariant compared to any human speaker's language productions, and as a result this limitation of cross-splicing might not be influential on linguistic studies of sociolinguistic variant use in synthetic speech.

Another limitation of the dissertation studies is that different voice stimuli were used in Experiment 1 compared to Experiments 2 and 3. Because voice AI technology is advancing rapidly, synthetic voice development improves significantly from year to year. For the current research, I chose to use the most cutting-edge, naturalistic voice synthesis methods available at the time that the experiment was designed because one of my main research questions is how the naturalness of synthetic voices affects listener social evaluations. But the change in voice stimuli limits my ability to make one-to-one comparisons with the synthetic voices used in Experiment 1. Future studies should consider carefully whether reusing the same synthetic voices for replicability versus using the most current versions of synthetic voices is more appropriate for their particular experimental approach.

Another limitation of this dissertation research is that I did not examine participants' level of education or type of profession. While the participants in Experiments 1 and 2 are enrolled in a university, the older adults in Experiment 3 were not selected based on their level of education. A future study that has a broader sample of younger adults from college and working-class backgrounds as well as a study that specifically examines the profession of older adults are areas for future research.

Additionally, an inherent fact of the older adults surveyed in Experiment 3 is that they have sufficient technological competence to complete a study online and of their own volition made an account on an online survey-completion platform. Therefore, Experiment 3 does not

include older adults who have very limited technological competency or have limited technological use. This is a frequent issue in studies on older adults and technology and by extension a systemic issue in understanding how older adults socially perceive voice AI technology (Barnard et al., 2013). In order to gain the fullest picture of older adults' social perceptions of computer voices, future research can recruit older adults that have lower technological competency than the ones in the current research and examine how differences in technological competency might affect their social evaluations of synthetic voices and their ability to differentiate computer voices from synthetic ones.

Broader Impacts and Future Directions

As mentioned in the limitations section above, relatively little work has examined how older adults socially perceive voice AI or even anthropomorphized computer interactants in general. Voice AI interactants and social computers have many possible beneficial uses for the aging population, from decreasing experiences of loneliness, improving health outcomes, and even staving off the development of dementia (Griol & Callejas, 2016; Khalil et al., 2025; Yan et al., 2024). But, to ensure that older adults actually use technology made to support healthy aging, they have to want to use it. The findings of this dissertation research inform how voice AI can be developed to increase uptake by elderly adults for social and cognitive engagement purposes.

On the other hand, older adults are a particularly vulnerable population for malevolent uses of voice AI. As shown in the current studies, there is a sizable difference in accuracy in identifying synthetic voices as computers between younger and older adults. This has concerning implications for older adults' exposure to synthetic speech designed for malicious purposes, namely "voice deepfakes" that are used to create realistic scams where a synthetic voice is generated to sound very similar to a real human. It has already been shown that older adults are

more vulnerable to voice deepfake attacks than younger adults (Müller et al., 2022). As the sophistication of voice AI technology continues to develop the possible uses of it for scams and misinformation are also proliferating. Thus, future scholarly work should investigate how people perceive voice AI and what cues they can use to differentiate it from real human speech.

Theoretical Contributions

The studies on human-computer interaction in the last 20 years suggest that human relationships with computers are evolving. The media equivalence theory developed by Nass et al. (1994), which hypothesized that people directly transfer social behavior from human interactions, may have explained human social behavior with computers at that point in time. More recent theories in human computer interaction hypothesize that increased user exposure to social computers has led to the development of a separate social category from humans (Gambino et al., 2020). But, up to this point, research in human computer interaction has not considered how this effect might interact with participant age. The findings of this dissertation suggest that older adults transfer sociolinguistic evaluations to computer voices differently from younger adults.

Taken together, the findings of this dissertation are consistent with research on social perception of (ING) variation. But, listener characteristics, particularly age, have a very large effect on how participants perceive voices that they believe to be computers. A main theoretical stake of sociolinguistic studies is that listeners bring their own values, beliefs, and personality traits to interactions with other talkers, and in the dissertation studies I find this is very clearly the case for synthetic voices as well. Belief that a voice was a computer was only one of many interrelated social perceptions that changed listeners' rating behavior, suggesting that computer voices have their own "social class" in the mind of listeners.

Moreover, I have found support for the hypothesis that computer voices are evaluated using sociolinguistic perceptions from human interactions, in line with theories in sociolinguistics that language variation has social meaning. Yet, a key ingredient of social meaning is still missing from interactions with computers compared to humans— in today's world, only the human interactant in a human-computer conversation has the ability to change their social positioning by expressing politeness, annoyance, or even abusive language. Computer interlocutors do not have agency to dynamically respond to these changing social conditions with nearly the same capability as human talkers, which is a crucial difference between human and computer interlocutors (Gambino & Liu, 2022). With future development of technology and the ever-increasing sophistication of chatbot design, this might not be true in the next few years. It will be important to understand how stance taking and social positioning between human and computer interlocutors unfolds once synthetic voices have the sophistication necessary to dynamically update their stylistic choices and social positioning.

References

Agha, A. (2003). The social life of cultural value. *Language & Communication*, *23*(3-4), 231–273.

Aoki, N. B., Cohn, M., and Zellou, G. (2022). The clear speech intelligibility benefit for text-to-speech voices: effects of speaking style and visual guise. JASA Express Lett. 2:045204. doi: 10.1121/10.0010274

Aoki, N. B., & Zellou, G. (2024). Being clear about clear speech: Intelligibility of hard-of-hearing-directed, non-native-directed, and casual speech for L1-and L2-English listeners. *Journal of Phonetics*, *104*, 101328.

Auxier, B. & Anderson, M. (2021). *Social media use in 2021*. Pew Research Center. https://www.pewresearch.org/wp-content/uploads/sites/20/2021/04/PI_2021.04.07_Social-Media-Use_FINAL.pdf

Ball, P. (1983). Stereotypes of Anglo-Saxon and non-Anglo-Saxon accents: Some exploratory Australian studies with the matched guise technique. *Language Sciences, 5*(2), 163–183.

Banks, J. (2020). Optimus primed: Media cultivation of robot mental models and social judgments. *Frontiers in Robotics and AI*, 7, 62.

Barnard, Y., Bradley, M. D., Hodgson, F., & Lloyd, A. D. (2013). Learning to use new technologies by older adults: Perceived difficulties, experimentation behaviour and usability. *Computers in Human Behavior*, *29*(4), 1715–1724.

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

Barreda, S., & Silbert, N. (2023). *Bayesian Multilevel Models for Repeated Measures Data.* Routledge. https://santiagobarreda.com/bmmrmd/variation-in-parameters-random-effects-and-model-comparison.html#sec-c6-out-sample-crossval

Baugh, J. (2003). Linguistic profiling. In A. Ball, S. Makoni, G. Smitherman, & A. Spears (Eds.), *Black linguistics* (pp. 167-180). Routledge.

Baugh, J. (2016). Linguistic profiling and discrimination. In *The Oxford handbook of language and society*, 349–368.

Bayley, R., & Zapata, J. (1994). Alternancia de códigos y normas de lenguaje en el sur de Texas. *Discurso Teoría y Análisis,*16, 51–71.

Boersma, P. & Weenink, D. (2023). Praat: Doing phonetics by computer [Computer program]. Version 6.3.16. Retrieved August 29, 2023 from http://www.praat.org/

Bucholtz, M. (1999a). "Why be normal?": Language and identity practices in a community of nerd girls. *Language in Society*, *28*(2), 203–223.

Bucholtz, M. (1999b). You da man: Narrating the racial other in the production of white masculinity. *Journal of Sociolinguistics, 3*(4), 443–460.

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

Campbell-Kibler, K. (2006). Variation and the listener: The contextual meanings of (ING). *University of Pennsylvania Working Papers in Linguistics, 12*(2), 6. https://core.ac.uk/download/pdf/76382035.pdf

Campbell-Kibler, K. (2007). Accent, (ING), and the social logic of listener perceptions. *American Speech, 82*(1), 32–64.

Campbell-Kibler, K. (2008). I'll be the judge of that: Diversity in social perceptions of (ING). *Language in Society, 37*(5), 637–659. https://doi.org/10.1017/S0047404508080974

Campbell-Kibler, K. (2009). The nature of sociolinguistic perception. *Language Variation and Change, 21*(1), 135–156.
Campbell-Kibler, K. (2010). Sociolinguistics and perception. *Language and Linguistics Compass*, *4*(6), 377–389.
Campbell-Kibler, K. (2012). The implicit association test and sociolinguistic meaning. *Lingua, 122*(7), 753–763.
Chappell, W. (2016). On the social perception of intervocalic /s/ voicing in Costa Rican Spanish. *Language Variation and Change*, *28*(3), 357-378.
Chen, Y. H., Keng, C. J., & Chen, Y. L. (2022). How interaction experience enhances customer engagement in smart speaker devices? The moderation of gendered voice and product smartness. *Journal of Research in Interactive Marketing*, *16*(3), 403–419.
Cheshire J. (2009). *Variation in an English Dialect*. Cambridge University Press.
Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., Aylett, M., Cabral, J., Munteanu, C., Edwards, J., & Cowan, B. (2019). The state of speech in HCI: Trends, themes and challenges. *Interacting with Computers*, *31*(4), 349–371. https://doi.org/10.1093/iwc/iwz016
Clark, L., Ofemile, A., & Cowan, B. R. (2021). Exploring verbal uncanny valley effects with vague language in computer speech. In B. Weiss, J. Trouvain, M. Barkat-Defradas, & J. Ohala (Eds.), *Voice attractiveness: Studies on sexy, likable, and charismatic speakers*, (pp. 317-330). Springer.
Cohn, M., Keaton, A., Beskow, J., & Zellou, G. (2023). Vocal accommodation to technology: the role of physical form. *Language Sciences*, *99*, 101567.
Cohn, M., Predeck, K., Sarian, M., & Zellou, G. (2021). Prosodic alignment toward emotionally expressive speech: Comparing human and Alexa model talkers. *Speech Communication, 135,* 66–75.
Cohn, M., & Zellou, G. (2020, October). Perception of concatenative vs. neural text-to-speech (TTS): Differences in intelligibility in noise and language attitudes. *Proceedings of Interspeech. China,* https://doi.org/10.21437/Interspeech.2020
Corretge, R. (2023). *Praat Vocal Toolkit.* https://www.praatvocaltoolkit.com
Coupland, N., & Bishop, H. (2007). Ideologised values for British accents. J*ournal of Sociolinguistics, 11*(1), 74-93.
Danielescu, A. (2020, July). Eschewing gender stereotypes in voice assistants to promote inclusion. *Proceedings of the 2nd Conference on Conversational User Interfaces, Spain,* (pp. 1–3).
Drager, K. (2010). Sociophonetic variation in speech perception. *Language and Linguistics Compass, 4*(7), 473–480.
Dunbar, A., King, S., & Vaughn, C. (2024). Dialect on trial: An experimental examination of raciolinguistic ideologies and character judgments. *Race and Justice.* https://doi.org/10.1177/21533687241258772
Eckert, P. (1989). *Jocks and burnouts: Social categories and identity in the high school.* Teachers College Press.
Eckert, P. (2000). *Linguistic variation as social practice: The linguistic construction of identity in Belten High*. Blackwell.
Eckert, P. (2008). Variation and the indexical field. *Journal of Sociolinguistics, 12*(4), 453–476. https://doi.org/10.1111/j.1467-9841.2008.00374.x

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology*, 41, 87–100.
Eckert, P. (2017). Age as a sociolinguistic variable. In F. Coulmas (Ed.), *The handbook of sociolinguistics* (pp. 151-167). Blackwell.
Endrass, B., Rehm, M., & André, E. (2011). Planning small talk behavior with cultural influences for multiagent systems. *Computer Speech & Language*, 25(2), 158–174.
Fischer, J. L. (1958). Social influences on the choice of a linguistic variant. *Word*, *14*(1), 47–56. https://doi.org/10.1080/00437956.1958.11659655
Foulkes, P., & Docherty, G. (2006). The social life of phonetics and phonology. *Journal of Phonetics, 34*(4), 409–438.
Foulkes, P., Docherty, G. J., & Watt, D. (2005). Phonological variation in child-directed speech. *Language*, *81*(1), 177-206.
Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication*, *1*, 71–85.
Gambino, A., & Liu, B. (2022). Considering the context to build theory in HCI, HRI, and HMC: Explicating differences in processes of communication and socialization with social technologies. *Human-Machine Communication*, *4*, 111-130.
Gessinger, J. (2010). Language variation, language change and perceptual dialectology.
Giles, H., & Ogay, T. (2013). Communication accommodation theory. In B. Whaley, & W. Samter (Eds.), *Explaining communication* (pp. 325–344). [eBook]. Routledge.
Gong, L., & Nass, C. (2007). When a talking-face computer agent is half-human and half-humanoid: Human identity and consistency preference. *Human Communication Research*, *33*(2), 163–193.
Gordon, E. (1997). Sex, speech, and stereotypes: Why women use prestige speech forms more than men. *Language in Society, 26*(1), 47–63.
GovInfo. (2001, June 7). *PL 107-16 Economic Growth and Tax Relief Reconciliation Act of 2001*. https://www.govinfo.gov/app/details/PLAW-107publ16
Griol, D., & Callejas, Z. (2016). Mobile conversational agents for context-aware care applications. *Cognitive Computation*, *8*(2), 336–356.
Heiss, A. (2021, November 8). A guide to modeling proportions with Bayesian beta and zero-inflated beta regression models. [Blog post] https://www.andrewheiss.com/blog/2021/11/08/beta-regression-guide/
Heo, J., Chun, S., Lee, S., Lee, K. H., & Kim, J. (2015). Internet use and well-being in older adults. *Cyberpsychology, Behavior, and Social Networking*, *18*(5), 268–272.
Holliday, N. (2021). Prosody and sociolinguistic variation in American Englishes. *Annual review of linguistics, 7*(1), 55–68.
Holliday, N. (2023). Siri, you've changed! Acoustic properties and racialized judgments of voice assistants. *Frontiers in Communication*, *8*.
Houston, A. C. (1985). *Continuity and change in English morphology: The variable (ING)*. [Doctoral dissertation, University of Pennsylvania]. University of Pennsylvania ScholarlyCommons. https://core.ac.uk/download/pdf/76388804.pdf
Hunt, A., & Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings, USA,* 1, 373–376.
Hu, X., Desai, S., Lundy, M., & Chin, J. (2024). *Beyond functionality: Co-designing voice user interfaces for older adults' well-being*. arXiv preprint arXiv:2409.08449.

Kerswill, P., & Trudgill, P. (2005). The birth of new dialects. In P. Auer, F. Hinskens, & P. Kerswill (Eds.), *Dialect change: Convergence and divergence in European languages* (pp. 196-220). Cambridge University Press.

Khalil, R. A., Ahmad, K., & Ali, H. (2025). *Redefining elderly care with agentic AI: Challenges and opportunities*. arXiv preprint arXiv:2507.14912.

Kiesling, S. F. (1998). Men's identities and sociolinguistic variation: The case of fraternity men. *Journal of Sociolinguistics, 2*(1), 69-99. https://doi.org/10.1111/1467-9481.00031

Kiesling, S. F., Coupland, N., & Jaworski, A. (2009). Fraternity men: Variation and discourses of masculinity. In N. Coupland & A. Jaworski (Eds.), *The new sociolinguistics reader* (pp. 187–200). Bloomsbury.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280. https://doi.org/10.1177/2515245918771304

Kuligowska, K., Kisielewicz, P., & Włodarz, A. (2018). Speech synthesis systems: disadvantages and limitations. *International Journal of Engineering & Technology, 7,* 234–239.

Labov, W. (1963). The social motivation of a sound change. *Word, 19*(3), 273-309.

Labov, W. (1964). Phonological correlates of social stratification. *American Anthropologist*, *66*(6), 164-176.

Labov, W. (1969). A study of non-standard English. *ERIC Clearinghouse for Linguistics.* https://eric.ed.gov/?id=ED024053

Labov, W. (1972). Some principles of linguistic methodology. *Language in Society, 1*(1), (pp. 97-120).

Labov, W. (1986). The social stratification of (r) in New York City department stores. In H. Allen & M. (Eds.), *Dialect and language variation* (pp. 304–329). Academic Press.

Labov, W., Ash, S., Ravindranath, M., Weldon, T., Baranowski, M., & Nagy, N. (2011). Properties of the sociolinguistic monitor. *Journal of Sociolinguistics, 15*(4), 431–463.

Lambert, W. E., Hodgson, R. C., Gardner, R. C., & Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *The Journal of Abnormal and Social Psychology, 60*(1), 44.

Lee, K. M., & Nass, C. (2005). Social-psychological origins of feelings of presence: Creating social presence with machine-generated voices. *Media Psychology*, *7*(1), 31–45.

Leigh-Hunt, N., Bagguley, D., Bash, K., Turner, V., Turnbull, S., Valtorta, N., & Caan, W. (2017). An overview of systematic reviews on the public health consequences of social isolation and loneliness. *Public Health*, 152, 157–171.

Levon, E., Sharma, D., Watt, D. J., Cardoso, A., & Ye, Y. (2021). Accent bias and perceptions of professional competence in England. *Journal of English Linguistics*, *49*(4), 355-388.

Levon, E., Sharma, D., & Ye, Y. (2022). Dynamic sociolinguistic processing: Real-time changes in judgments of speaker competence. *Language 98*(4), 749–774. https://doi.org/10.1353/lan.2022.0020.

Levon, E., & Ye, Y. (2020). Language, indexicality and gender ideologies: contextual effects on the perceived credibility of women. *Gender and Language*, *14*(2), 123–151. https://doi.org/10.1558/genl.39235

Lilley, K. D., Dossey, E., Cohn, M., Clopper, C. G., Wagner, L., & Zellou, G. (2025). Social evaluation of text-to-speech voices by adults and children. *Speech Communication, 166,* 103163.

Lippi-Green, R. (1997). What we talk about when we talk about Ebonics: Why definitions matter. *The Black Scholar*, *27*(2), 7–11.
Loureiro-Rodriguez, V., Boggess, M. M., & Goldsmith, A. (2012). Language attitudes in Galicia: using the matched-guise test among high school students. *Journal of Multilingual and Multicultural Development, 34*(2), 136–153.
Martín Rojo, L., & Molina, C. (2017). Cosmopolitan stance negotiation in multicultural academic settings. *Journal of Sociolinguistics*, *21*(5), 672–695.
Mechler, J. (2025). Speaker evaluation across the adult lifespan: Examining age and gender effects in the perception of Tyneside voices. *Journal of Language and Aging Research*, *3*(1), 48-75.
Mendoza-Denton, N. C. (1997). *Chicana/Mexicana identity and linguistic variation: An ethnographic and sociolinguistic study of gang affiliation in an urban high school.* [Doctoral dissertation, Stanford University]. ProQuest Dissertations and Theses.
Mendoza-Denton, N. (2011). The semiotic hitchhiker's guide to creaky voice: Circulation and gendered hardcore in a Chicana/o gang persona. *Journal of Linguistic Anthropology*, *21*(2), 261-280.
Milroy, L. (2002). Standard English and language ideology in Britain and the United States. In T. Bex, & R. J. Watts (Eds.), *Standard English* (pp. 173–206). [eBook]. Routledge.
Müller, N. M., Pizzi, K., & Williams, J. (2022, October). Human perception of audio deepfakes. *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia,* (pp. 85–91).
Nass, C. I., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship.* MIT Press.
Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, *7*(3), 171.
Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of Applied Social Psychology*, *27*(10), 864–876.
Nass, C., & Steuer, J. (1993). Voices, boxes, and sources of messages: Computers and social actors. *Human Communication Research*, *19*(4), 504–527.
Nass, C., Steuer, J., & Tauber, E. R. (1994, April). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, USA,* (pp. 72–78).
Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, *18*(1), 62–85.
Ning, Y., He, S., Wu, Z., Xing, C., & Zhang, L. J. (2019). A review of deep learning-based speech synthesis. *Applied Sciences, 9*(19), 4050.
Olmstead, K. (2017, December 12). *Nearly half of Americans use digital voice assistants, mostly on their smartphones.* Pew Research Center. https://www.pewresearch.org/short-reads/2017/12/12/nearly-half-of-americans-use-digital-voice-assistants-mostly-on-their-smartphones/
Ospina, R., & Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, *56*(6), 1609–1623.
Paunonen, H. (1994). The Finnish language in Helsinki. In B. Nordberg (Ed.), *The sociolinguistics of urbanization: The case of the Nordic countries* (pp. 223–246). Walter de Gruyter.

Perera, N. N., & Ganegoda, G. U. (2023, February). A comprehensive review on speech synthesis using neural-network based approaches. *Proceedings of the 3rd International Conference on Advanced Research in Computing,* (pp. 214-219).

Podesva, R. J. (2007). Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics*, *11*(4), 478–504.

Podesva, R. J., & Callier, P. (2015). Voice quality and identity. *Annual Review of Applied Linguistics*, *35*, 173–194.

Pradhan, A., & Lazar, A. (2021). Hey Google, do you have a personality? Designing personality and personas for conversational agents. *Proceedings of the 3rd Conference on Conversational User Interfaces* (pp. 1–4).

Preston, D. R. (1996). Where the worst English is spoken. *Focus on the USA*, *16,* 297–360.

Preston, D. R. (2011). *Perceptual dialectology: Nonlinguists' views of areal linguistics* (Vol. 7). Walter de Gruyter.

Preston, D. R. (2013). Language with an attitude. In J. K. Chambers, & N. Schilling (Eds.), *The handbook of language variation and* change (pp. 157–182). Wiley.

Purnell, T., Idsardi, W., & Baugh, J. (1999). Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology, 18*(1), 10–30.

Pycha, A., & Zellou, G. (2024). The influence of accent and device usage on perceived credibility during interactions with voice-AI assistants. *Frontiers in Computer Science*, *6*. https://doi.org/10.3389/fcomp.2024.1411414

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people*. Cambridge University Press.

Robinson, D (2014). *Understanding the beta distribution (using baseball statistics)*. Variance explained. http://varianceexplained.org/statistics/beta_distribution_and_baseball/

Scarborough, R. (2005). *Splicing*. [Course Supplemental Handout]. https://web.stanford.edu/class/linguist205/index_files/Suppl%20handout%204%20-%20Splicing.pdf

Schleef, E., & Flynn, N. (2015). Ageing meanings of (ing): Age and indexicality in Manchester, England. *English World-Wide, 36*(1), 48–90.

Schleef, E., Flynn, N., & Ramsammy, M. (2015). Production and perception of (ing) in Manchester English. In E. Torgersen, S. Hårstad, B. Mæhlum, U. Røyneland (Eds.), *Language Variation-European Perspectives V,* (pp. 197–210). John Benjamins Publishing Company.

Schleef, E., Meyerhoff, M., & Clark, L. (2011). Teenagers' acquisition of variation: A comparison of locally-born and migrant teens' realisation of English (ing) in Edinburgh and London. *English World-Eide, 32*(2), 206–236. https://doi.org/10.1075/eww.32.2.04sch

Sharma, D., Levon, E., & Ye, Y. (2022). 50 years of British accent bias: Stability and lifespan change in attitudes to accents. *English World-Wide, 43*(2), 135–166.

Schwaferts, P., & Augustin, T. (2020). *Bayesian decisions using regions of practical equivalence (ROPE): Foundations.* Department of Statistics, University of Munich. 10.5282/ubm/epub.74222

Shulevitz, J. (2018). Alexa, should we trust you? *The Atlantic*. https://www.theatlantic.com/magazine/archive/2018/11/alexa-how-will-you-change-us/570844/

Stylianou, Y., & Syrdal, A. (2001, May). Perceptual and objective detection of discontinuities in concatenative speech synthesis. *2001 IEEE international conference on acoustics, speech, and signal processing proceedings, USA,* (Vol. 2), 837–840.

Sundar, S. S., & Kim, J. (2019, May). Machine heuristic: When we trust computers more than humans with our personal information. *Proceedings of the 2019 CHI Conference on human factors in computing systems* (pp. 1-9).

Sutton, S.J., Foulkes, P., Kirk, D., & Lawson, S. (2019). Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction. In *Proceedings of the 2019 CHI conference on human factors in computing systems, Scotland,* 1–14.

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., DeCero, E., & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: thematic analysis. *Journal of Medical Internet Research*, *22*(3), e16235.

Tang, Y., Horikoshi, M., & Li, W. (2016). ggfortify: Unified interface to visualize statistical results of popular R packages. *The R Journal*, *8*(2), 474.

Thormundsson, B. (2024, May 21). U.S. adults on impact of AI on their lives, 2024 by age. *Statista*. https://www.statista.com/statistics/1466969/opinion-ai-impact-on-their-lives-us-by-age/

Tinwell, A. (2013). Applying psychological plausibility to the uncanny valley phenomenon. In M. Grimshaw-Asgaard (Ed.), *The Oxford Handbook of Virtuality,* (pp. 173-186). Oxford Handbooks.

Tobing, P. L., Wu, Y. C., Hayashi, T., Kobayashi, K., & Toda, T. (2019, May). Voice conversion with cyclic recurrent neural network and fine-tuned WaveNet vocoder. *International Conference on Acoustics, Speech and Signal Processing, Brighton,* 6815–6819.

Trudgill, P. (1972). Sex, covert prestige and linguistic change in the urban British English of Norwich. *Language in Society, 1*(2), 179–195.

Trudgill, P. (1974). Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in Society, 3*(2), 215–246. https://doi.org/10.1017/S0047404500004358

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131.

Vaughn, C. (2022). What carries greater social weight, a linguistic variant's local use or its typical use? *Proceedings of the Annual Meeting of the Cognitive Science Society, 44*(44).

Vaughn, C. (2022). The role of internal constraints and stylistic congruence on a variant's social impact. *Language Variation and Change, 34*(3), 331–354.

Vaughn, C., & Kendall, T. (2018). Listener sensitivity to probabilistic conditioning of sociolinguistic variables: The case of (ING). *Journal of Memory and Language*, 103, 58–73.

Villarreal, D. (2018). The construction of social meaning: A matched-guise investigation of the California Vowel Shift. *Journal of English Linguistics, 46*(1), 52–78.

Vuorre, M. (2019, February 18). How to analyze visual analog (slider) scale data. *Matti's website.* https://vuorre.com/posts/2019-02-18-analyze-analog-scale-ratings-with-zero-one-inflated-beta-models.

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

Wollum, E. (2019). *The uptalk downgrade: Comparing age-and gender-based perceptions of uptalk in four highly-skilled professions* [Doctoral dissertation, Te Herenga Waka-Victoria University of Wellington]. Open Access.

Yu, D., Cohn, M., Yang, Y.M, Chen, C.Y., Wen, W., Zhang, J., Zhou, M., Jesse, K., Chau, A., Bhowmick, A. Iyer, S., Sreenivasulu, G., Davidson, S., Bhandare, A., & Yu, Z. (2019). Gunrock: A social bot for complex and engaging long conversations. *arXiv preprint arXiv:1910.03042*.

Yan, C., Johnson, K., & Jones, V. K. (2024). The impact of interaction time and verbal engagement with personal voice assistants on alleviating loneliness among older adults: an exploratory study. *International Journal of Environmental Research and Public Health*, *21*(1), 100.

Zellou, G. (2022). *Coarticulation in phonology.* Cambridge University Press.

Zellou, G., Cohn, M., & Pycha, A. (2023). Listener beliefs and perceptual learning: differences between device and human guises. *Language*, *99*(4), 692–725.

Zellou, G., & Holliday, N. (2024). Linguistic analysis of human-computer interaction. *Frontiers in Computer Science, 6,* 1384252.

Appendix

Figure A. Experiment 1 Prompt

We are developing a newscasting app!

We want our app to have a newscaster who reads the news to users. We want to choose the best voice to read the news, and we have four voices we can choose from to read the broadcast.

You will rate each voice on these eight qualities:
1. Appropriateness for news broadcasting
2. Professionalism
3. Intelligence
4. Friendliness
5. Trustworthiness
6. Authenticity
7. Naturalness
8. Perceived Age

Then you'll answer two questions about your experience listening to the broadcast.
1. While you were hearing the broadcast, how much did you feel as if the voice was talking to you specifically?
2. While you were hearing the broadcast, how vividly were you able to mentally imagine the person who was speaking?

There is also a comment box if you want to tell us your thoughts about each voice.

Note. After this page, participants advanced to the following page before starting the task.

It's the same one-minute news report each time, so you will only need to focus on an individual voice one-at-a-time.

The listening portion is only four minutes in total for all voices, but please *listen closely each time*.

Figure B. Experiment 1 Stimuli

President Bush announced tonight that he was putting all available White House resources into support for the new tax cut bill. Democratic leaders of the House and Senate are preparing compromise legislation. Republican spokespersons predicted that record numbers of working-class Americans would be receiving tax refund checks before the end of the year. Senator Edward Kennedy's staff announced that the tax cuts are creating a new elite who are excused from paying their fair share of the cost of government. At the Office of Management of the Budget, officials are trying to estimate the size of the deficit that will be produced by the new legislation. Federal Reserve Board chairman Alan Greenspan stated that he was not confirming that tax cuts would lead to a change in prime interest rates, nor was he denying it.

The Washington Post is publishing today a list of all members of Congress who will receive tax refunds greater than $1,000 as a result of the proposed tax cuts.

Figure C. Prompt for Experiments 2 and 3

News Broadcast Prompt	**Radio DJ Prompt**
We are developing a newscasting app! We want our app to have a newscaster who reads the news to users. We have four voices we can choose from to read the broadcast**.**	We are developing a music recommendation app! We want our app to have a voice who makes playlist recommendations for listeners. We have four voices we can choose from to read the recommendations.

Note. Participants in Experiments 2 and 3 were sorted into the news broadcast group or radio DJ broadcast group before the start of the experiment. Participants read the prompt that corresponded to the group they were sorted into, then clicked a "next" button that took them to the following page.

You will hear four voices read an audition script. **You will rate each voice on these eight qualities:** 1. Appropriateness for a [news broadcast OR music] app 2. Professionalism 3. Intelligence 4. Friendliness 5. Age 6. Trustworthiness 7. Authenticity 8. Naturalness **Then you'll answer two questions about your experience listening to the audition script.** 1. While you were hearing the audition script, how much did you feel as if the voice was talking to you specifically? 2. While you were hearing the audition script, how vividly were you able to mentally imagine the person who was speaking? There is also a comment box if you want to tell us your thoughts about each voice.

Note. After this instruction, participants advanced to the following page before starting the task.

You'll hear the same one-minute news report each time and then answer the rating questions. **The listening portion is only four minutes in total for all voices. We know this will be a little repetitive, but the task is very short. Please *listen closely each time*.**

Figure D. Stimuli for Experiments 2 and 3

News Broadcast

The National Merit Foundation announced today that they will be awarding high school senior Eric Smith with the prestigious national merit scholarship. Eric was competing against a pool of over fifteen thousand other applicants across the United States. A spokesperson from the foundation stated that the scholarship review board was seeking high school seniors with excellent academic performance as well as an active role in community service. Eric had been working hard all year to maintain straight A's. He also spent over nine months preparing his scholarship application essay. Aside from his academics, Eric was participating in community service as often as he could. The animal shelter where Eric completed his community service said Eric was helping at the shelter almost 30 hours every week. After graduation, Eric is planning to attend the University of Virginia and will be majoring in psychology. Eric will be receiving the scholarship at tonight's award ceremony. All members of the community are welcome to attend this event.

Radio DJ Broadcast

I'm putting a playlist together for you of some songs that I've been really into this summer. I love picking out new music for people and making recommendations. There's so many tracks to choose from across so many different genres, but I'm sharing the ones I love the most. Lately, I'm playing a lot of indie pop and R&B. I like adding some Latin pop and new wave stuff to the mix as well. I'm always searching for new music to add to the playlist. I'm even following a lot of my favorite artists on social media so I can listen to their songs as soon as they come out. I'm always updating the music on this playlist, so you can check it out whenever you want. Anyway, I think you'll like listening to my playlist but you have to let me know what you think. I tried to find stuff you'd like, and I hope you enjoy it.