The Influence of Conversational Role on Phonetic Alignment toward Voice-AI and Human Interlocutors

Running header: **Role and interlocutor affects alignment**

Georgia Zellou, Michelle Cohn, and Tyler Kline

Phonetics Laboratory, Linguistics Department, University of California, Davis
469 Kerr Hall, One Shields Ave.,
Davis, CA 95616, USA

Corresponding author email: gzellou@ucdavis.edu

Abstract

Two studies investigated the influence of conversational role on phonetic imitation toward human and voice-AI interlocutors. In a Word List Task, the giver instructed the receiver on which of two lists to place a word; this dialogue task is similar to simple spoken interactions users have with voice-AI systems. In a Map Task, participants completed a fill-in-the-blank worksheet with the interlocutors, a more complex interactive task. Participants completed the task twice with both interlocutors, once as giver-of-information and once as receiver-of-information. Phonetic alignment was assessed through similarity rating, analyzed using mixed-effects logistic regressions. In the Word List Task, participants aligned to a greater extent toward the human interlocutor only. In the Map Task, participants as giver only aligned more toward the human interlocutor. Results indicate that phonetic alignment is mediated by the type of interlocutor and that the influence of conversational role varies across tasks and interlocutors.

Keywords: phonetic alignment, voice-AI, human-device interaction, conversational role

1. INTRODUCTION

Human speech contains a huge amount of variation, some of which is shaped by multiple factors present in a speaker's environment. For one, there is much work showing that talkers adopt the speech and language patterns of another talker (or even a voice presented over headphones) in subtle and subconscious ways, a process known as 'alignment' (or 'imitation' or 'entrainment') (Goldinger, 1998). Alignment can occur at multiple levels of linguistic structure (phonetic, lexical, syntactic) (e.g., Pickering & Garrod, 2004) and even gesture (Oben & Brône, 2015) and posture (Shockley et al., 2007). Furthermore, there is evidence that alignment is not constrained to interaction between humans: people align toward computers, as well (Bell, 2003; Branigan et al., 2011; Cowan et al., 2015; Fandrianto & Eskenazi, 2012; Gessinger et al., 2021, 2019; Thomason

et al., 2013). Yet, there are many questions about the dynamics of these patterns of alignment toward computer interlocutors and the extent to which behaviors from human-human interaction 'transfer' to interactions with technology.

The current paper tests whether the conversational role of the participant — as either the giver or receiver of information — influences phonetic alignment toward computer and human interlocutors. We examine phonetic alignment, where a speaker imitates the acoustic-phonetic properties of another talker's speech (e.g., Babel, 2012; Namy et al., 2002). Phonetic alignment has been shown across studies to vary based on conversational role (e.g., Pardo, 2006; Pardo et al., 2010; Pardo et al., 2013), as well as serve as an index of social harmony: interactions with more phonetic alignment are associated with improved rapport between interacting humans (e.g., Levitan et al., 2012) or with a computer dialogue system (e.g., Thomason et al., 2013). Furthermore, phonetic alignment is relevant given that people regularly *talk* to technology; increasingly, people interact with voice-activated artificially intelligent (voice-AI) systems on phones and devices (e.g., Apple's Siri and Amazon's Alexa) (Bentley et al., 2018) to complete a variety of tasks in interactive, yet, goal-oriented and utilitarian ways. Indeed, there is a growing body of research aimed at understanding alignment toward voice-AI systems (Cohn & Zellou, 2019; Raveh et al., 2019; Zellou et al., 2020; Zellou & Cohn, 2020).

As mentioned earlier, phonetic alignment has been shown across studies to vary based on conversational role (Pardo, 2006; Pardo et al., 2010, 2013, 2018). For example, Pardo (2006) had participant dyads complete a map task with one participant assigned either the role of 'information-giver' or the role of 'information-receiver'. In this task, the giver had a complete map with detailed directions marking the pathway along various landmarks from the start point to the location of the finish position, while the receiver had a map with just the start point and the landmarks. The task was for the giver to provide instructions to the receiver on how to navigate around the landmarks to get from start to finish. They found that both conversational role and gender of the participant had an effect on degree of phonetic alignment: female givers aligned more than female receivers, while male receivers aligned more than male givers. But, in other work, gender does not mediate the pattern (Pardo et al., 2013).

Will there be greater phonetic alignment from 'givers' of information when they talk to voice-AI systems? On the one hand, we might predict differences since spoken interactions with voice-AI assistants are generally asymmetrical in terms of conversational role, where users typically issue commands to devices, giving instructions to play music, set a timer, ask for information (weather, news reports), or control smart home integration functions ("internet of things" (IoT) functions, such as turning on lights or TV) (Ammari et al., 2019). Thus, the majority of verbal interactions with voice-AI can be classified as the user adopting a more dominant *giver*-of-information role in the dialogue (e.g., "set a timer", "play my favorite song", "tell me the weather report"). Indeed, there is work suggesting that people expect to interact with technology in different ways than with humans (Edwards, 2018) and some have proposed that we have a routinized way of engaging with technology (Gambino et al., 2020). On the other hand, the Computers are social actors (CASA) framework hypothesizes that people's behavior toward technology mirrors their behaviors from human-human interaction given a cue of humanity is displayed by the system (Nass et al., 1994)); accordingly, the 'giver' and 'receiver' roles might influence behaviors in similar ways for both types of interactions. For example, related work has shown that the local situational context — for instance, whether the participant was winning or losing a game — influences phonetic alignment to the same extent towards both robot and human interlocutors: Ibrahim et al. (2019) examined speech interactions between participants and a robot

or another human while playing a card game. They found that phonetic alignment (specifically of fundamental frequency, f0[1]) was higher when they were winning, but no difference by interlocutor type. Thus, they do observe situational factors affect phonetic imitation toward robots, similar as toward people, arguing that the greater social coordination and rapport that led to winning also led participants to align to a greater extent.

While, to our knowledge, no prior work has tested the interaction of conversational role and interlocutor type (human vs. computer) on phonetic alignment, this line of research can speak to broader theories of alignment, as functionally or socially driven.

1.1. Theories of Alignment

Some theories of alignment, what we refer to as *cognitive-functional accounts*, propose that linguistic alignment is communicatively-strategic and it is used specifically to improve mutual understanding of interlocutors (Clark & Murphy, 1982; Pickering & Garrod, 2004; Zellou et al., 2016). For example, the Interactive Alignment Account proposes that speakers align their output (across multiple levels of the grammar) to enhance intelligibility, matching representations between interlocutors (Pickering & Garrod, 2004). Pickering and Garrod (2004) further claim that imitation relies on automatic, resource-free priming mechanisms. Similarly, Audience Design theory (Clark & Murphy, 1982; Clark & Schaefer, 1987) proposes that speakers choose to converge toward the linguistic patterns of an interlocutor to improve intelligibility. Indeed, Pardo's (2006) findings that participants align more as giver is consistent with the interpretation that increased alignment is found in contexts where it would increase communicative success. Other work in human-computer alignment also suggests a functional role: Branigan et al. (2011) told subjects they were interacting with a computer system or a person in another room across separate experimental blocks, in both pre-scripted speech- and text-based interactions. They found that when participants believed they were communicating with a computer they showed greater alignment in word choice than when they were told they were interacting with a human. The authors interpret this finding as functionally-motivated: people judge the computer as less communicatively able than the human, and consequently align more toward it in an effort to increase communicative success. This supports the stance of many researchers who have proposed that people's speech and language patterns during human-computer interaction are driven by expectations that machines are less communicatively able interlocutors (for review, see Clark et al., 2019). In the current study, a possible outcome is that people view the voice-AI system as communicatively less able than humans and consequently will display more phonetic alignment toward the device voice. Moreover, when participants are in the giver role there is increased pressure to be intelligible since participants are providing instructions to the interlocutor. Thus, it is possible that participants will phonetically align to a greater extent in the giver role toward the device than toward the human.

On the other hand, *socially-mediated accounts*, propose that alignment is driven by the characteristics of and relationships between interlocutors. For example, Communication Accommodation Theory (CAT; Gallois & Giles, 2015; Shepard, 2001) posits that alignment between interlocutors is one way that people enhance, maintain, or decrease social distance. In this approach, phonetic alignment is not viewed as a way to augment intelligibility, rather it is a way

[1] Though one limitation of the Ibrahim at al. (2019) study was that they only looked at one phonetic feature (f0), a feature which could be modulated by factors such as cognitive load and/or emotional expressiveness. In the current study, we assess phonetic imitation holistically (through AXB perception ratings) rather than one acoustic variable.

to display social solidarity or distance. Social dimensions such as gender, regional affiliation, ethnicity, age, sexuality, and more, have been shown to predict patterns of phonetic imitation between interacting talkers in ways that support this stance (Babel, 2012; Namy et al., 2002; Pardo et al., 2010; Yu et al., 2011). There is some evidence that people phonetically align *less* toward voice-AI (in direct comparisons with humans), suggesting a possible social role in mediating alignment toward computers. For example, Cohn et al. (2019) compared phonetic alignment patterns toward human and Siri voices by participants in a word shadowing task (where participants repeat isolated words produced by an interlocutor). Using a perceptual assessment, they found less alignment toward the Siri voices, overall, than toward the human voices (similar patterns were reported using an acoustic assessment of alignment in Snyder et al. (2019)). The observation of *less* alignment toward device voices than toward human voices suggests that people are sensitive to the social distinction between devices and humans, in line with a Communication Accommodation Theory (Giles et al., 1991), as well as accounts proposing more gradient application of CASA (e.g., that individuals show greater phonetic alignment toward machines as their embodiment increases in Cohn et al. (2020)). In other words, people are more socially close to other humans than AI, therefore, they might align more toward humans than voice-AI. Such an account might also predict participants will display less socially-mediated alignment patterns within the voice-AI condition, than for the human condition, as was reported in Cohn et al. (2019) for gender-mediated phonetic imitation patterns. For the current study, if voice-AI have a distinctly lower social status than humans, this will be reflected in a *smaller degree of* phonetic alignment, particularly when the participant is the giver, which is a more dominant conversational role. This interpretation is in line with Pardo (2006), who frames giver and receiver as social roles, reflecting how relational pressures such as dominance and deference shape speech and language patterns between interlocutors of different social statuses (e.g., Bilous & Krauss, 1988).

1.2. Current Study

The current study compares participants' phonetic alignment toward a human and device interlocutor when they are in a giver and receiver role during conversational interactions. Many studies of Human-Computer interaction use a "Wizard-of-Oz" paradigm, where participants are led to believe they are interacting with a digital system, yet the system is being controlled in real-time by a remote confederate (Dahlbäck et al., 1993). In the current study, in contrast, participants completed an automated and scripted task where they read target phrases presented on a screen and heard pre-recorded utterances produced by the human and the voice-AI interlocutors (naturally produced recordings and generated via TTS, respectively). In HCI work, this is similar to an approach that implements an approximation of a voice user interface (see Clark et al., 2019 for a review of studies that use a similar approach). In the present study, participants completed blocks as both the giver- and receiver-of-information with both interlocutor types (2 x 2, within-subjects design). These comparisons were made across two studies, varying in the nature of the task: in Study I, participants completed a Word List Task with the interlocutors; the Word List Task was designed to mirror the types of simple, utilitarian, command-based interactions between users and at-home digital voice assistant devices (i.e., dictating a shopping list). In Study II, participants completed a fill-in worksheet Map Task; this task was designed to be a more complex task similar to those used in studies of speech variation in human-human interaction (Scarborough & Zellou, 2013; Zellou & Scarborough, 2019) and also phonetic alignment (Pardo, 2006). The interactions are designed to simulate dialogues between the participant and another human and a digital voice

assistant (across separate blocks), with both top-down guise information (an image of the respective interlocutors) and distinct voices.

A common approach to quantify phonetic alignment is to use a perceptual assessment of imitation where third-party raters complete a perception experiment (typically, an "AXB" rating task) in which they listen to baseline (e.g., "A") and post-exposure (e.g., "B") productions and decide which one sounds more like the model talker's production (e.g., "X") (e.g., Dias & Rosenblum, 2016; Namy et al., 2002; Pardo et al., 2017). If raters select the post-exposure item at an above-chance proportion, this is taken as evidence that phonetic imitation has occurred (Miller et al., 2013). This method has also previously been used as a holistic way to classify phonetic alignment in human and voice-AI comparisons (e.g., Cohn et al., 2019). Thus, in the current study, this approach is adopted to assess phonetic imitation across both studies. As illustrated in Figure 1, both Study I and Study II consist of two experiments: the first experiment collects imitators' productions of words before and after exposure to the model talkers' productions of those words (Goldinger, 1996). The second experiment is an AXB similarity ratings task where a separate group of listeners rate baseline and post-exposure productions from the imitators in the first experiment, allowing for a holistic assessment of vocal alignment (Pardo et al., 2013).

[INSERT FIGURE 1 ABOUT HERE]

Figure 1: *Overview of the current study experiments aimed to investigate phonetic alignment (Experiment 1: Word List Task, Experiment 3: Map Task) by imitators and the Similarity Rating (AXB) paradigm used to evaluate phonetic imitation by a separate group of listeners.*

2. Study I - Word List Phonetic Imitation Task

Study I consists of two experiments. In Experiment 1 (Section 2.1.) participants interacted with both a human and voice-AI interlocutor (2 separate blocks), as both the giver- and receiver-of-instructions (2 separate sub-blocks within each interlocutor block) during the Word List Task. Then, degree of phonetic imitation in participants' test productions is assessed in a separate AXB perceptual similarity ratings experiment (Experiment 2, Section 2.2.).

2.1. Experiment 1: Word List Production (female imitators only)

The first study was designed to investigate whether speakers display different patterns of phonetic imitation based on conversational role (as 'giver' or 'receiver' of information) and whether they are talking to a device or human interlocutor. In light of work that has shown mixed effects of gender on alignment (e.g., Pardo, 2006; Pardo et al., 2013), we hold gender constant in Experiment 1, examining only female imitators.

2.1.1. Target words and pre-recorded model talker utterances

Target words consisted of 15 low frequency monosyllabic items (selected from Babel, 2012) with consonant-vowel-consonant (CVC) or consonant-vowel (CV) structure, and balanced by across 5 vowels: /i/ *cheek, deed, key*; /æ/ *bat, tap, vat*; /ɑ/ *cot, sock, sod*; /o/ *coat, soap, toad*; /u/ *boot, hoop, zoo*.

Model talker utterances were pre-recorded: the human voice stimuli were recorded by a male native California English speaker in a sound-attenuated booth, with a head-mounted

microphone (Shure WH20 XLR), and sampling rate of 44,100 Hz. The device voice stimuli were generated with an Amazon Polly voice (US-English: 'Matthew').

2.1.2. Participants

Since it has been reported that women align more than men in word shadowing tasks (e.g., Namy et al., 2002), only female participants were selected to complete the Word List Task to increase the likelihood of alignment. Participants consisted of 23 female undergraduates (mean age = 20.4, sd = 1.3), all native English speakers, recruited from the UC Davis Psychology subject pool who received course credit for their participation. All except 1 participant reported that they had used digital voice assistant devices before participating in the experiment. 12 out of these 22 participants reported having used Amazon's Alexa, 9/12 reported rare usage, while only 1 reported daily usage, and another reported using Alexa on a weekly basis. Participants who reported using other devices reported using Siri (12/23) and/or Google (5/23). Of these participants, 8 reported daily usage, 5 reported once-a-week usage, while the rest (3 participants) reported rare usage.

2.1.3. Procedure

Participants completed the experiment in a sound-attenuated booth, facing a computer monitor, and wearing a head-mounted microphone (Shure WH20 XLR) and over-ear headphones (Sennheiser Pro). Before beginning the experiment, participants were shown a list of the target words on a piece of paper and were asked to familiarize themselves with the words. Then, participants completed the production experiment in E-Prime 2.0 as one continuous session, consisting of two parts: A pre-exposure production block and a Word List Task paradigm (where participants interact with pre-recorded voices to complete a task).

First, participants completed a **pre-exposure production block** (schematized in Figure 2) in order to collect their baseline productions of words before exposure to the model talkers. On each trial, they saw the target word presented in a frame sentence ("Put __ on the list") and read it aloud. Their utterances were recorded. After 4.5 seconds, the next target sentence would be shown on the screen (ISI between slides = 300 ms). In total in the pre-exposure block, they produced the 15 target words (randomly presented).

[INSERT FIGURE 2 ABOUT HERE]

Figure 2: *Trial structure for Baseline production block collecting production of words in the frame sentence "Put ___ on the list" recorded prior to the interactions with the human and device model talkers. (Color online.)*

Next, participants completed the novel **Word List Task**, designed to elicit productions of the target words following productions of those items by the model talkers, varying the conversational role (participant as 'giver' role schematized in Figure 3A and 'receiver' role in Figure 3B). The information containing the instructions to the task that participants heard is provided in Appendix A. Each trial consists of 3 scripted turns between participant and model talker. **Turn 1:** the model talker produces the target word in a sentence ("The word is deed."). **Turn 2**: the participant reads a sentence aloud. In the 'giver' role, this is an instruction ("Put deed on the first list.") (A.ii.). In the 'receiver' role, they check that word off (using a dry erase marker on a laminated list, on a clipboard) and gave a verbal response confirming the action ("I checked deed off the list.") (B.ii.). **Turn 3:** the model talker responds to close the mini-dialogue ("Okay," "Got it", "Alright"). If the

participant gave the instructions (in 'giver' role), the target word on the correct list was shown on the screen to provide feedback to the participant (A.iii.). Whenever the model talker (human or TTS voice) produced the target word, it was presented orthographically on the screen so that the target word was unambiguous (and not referring to a different word). Critically, in every mini-interaction trial there is one utterance where the participant produces the target word *following* an utterance where the model talker produced that word.

On each trial (target words were randomly presented within a sub-block), participants read their response from the screen aloud. They had a window of 5 seconds to produce the sentence. Then, the experiment would automatically proceed (ISI=1 second). Participants completed two experimental blocks, blocked by model talker. Within each model talker block, participants completed two separate sub-blocks: one where they were the giver of information and one where they were the receiver. The ordering of model talker block and conversational role sub-block was counterbalanced across participants across four versions. Participants were randomly assigned to one of these versions. In total, participants completed 15 turns (one for each word) in each of the 4 interactive blocks (60 total test trials).

[INSERT FIGURE 3 ABOUT HERE]

Figure 3: *Word List Task 3-turn interaction design. (A) Participant as 'giver' of information, while (B) Participant is 'receiver' of information. (i) The first turn consists of the model talker indicating the target word (e.g., "The word is __"). (ii) The second turn consists of the participant reading a sentence aloud. In the 'giver' role, this is an instruction (A.ii.). In the 'receiver' role, they check that word off (using a dry erase marker on a laminated list, provided on a clipboard) and read the sentence to confirm (B.ii.). (iii) The third turn consists of a confirmation by the model talker (e.g., "Okay"). Turns where the participant reads the sentence from the screen are indicated with a microphone and yellow speech bubble. (Color online.)*

2.2. Experiment 2: AXB Similarity Ratings (Word List Productions)

In order to assess the extent of phonetic alignment in Experiment 1, Experiment 2 utilized a perceptual similarity ratings task (AXB) (Goldinger, 1996; Pardo et al., 2013). Here, a separate group of raters assessed the holistic similarity between the participants' post-exposure productions of the target words and the model's produced target item, relative to participants' baseline productions of the words (recorded prior to exposure to the model talkers).

2.2.1. Word List AXB Methods
2.2.1.1. Stimuli

Stimuli consisted of the productions of the 15 target words from the Word List Task in Experiment 1 (see Section 2.1.1.) excised from participants' pre-exposure utterances (e.g., "Put weave on the list.") and productions from the interactions (e.g., "I checked weave off the list"). We also excised the model talker's productions of the target words in the turns.

To assess participants' imitation of the model talker, we selected recordings for critical turns: when a model talker production was followed by a participant turn. In the 'giver' role, critical turns are Turns 1 and 2 (schematized in Figure 3A). In the 'receiver' role, critical turns are also Turns 1 and 2 (schematized in Figure 3B).

Participants' pre-exposure of the word (e.g., "A"), the model talker's production of that word ("X"), and the participants' production of that word post-exposure (i.e., following the model talker's production of the word) (e.g., "B") were excised from utterances into their own sound files. Individual word files were amplitude-normalized to 60 dB. Then, each AXB set was concatenated into a single sound file (with 400ms between each sound). Order of pre- and post-exposure was counterbalanced for each speaker and model talker.

2.2.1.2. Participants

A total of 56 independent raters (mean age, 19.9 years old; 25 female, 30 male, 1 non-binary) completed Experiment 2, consisting of native English speakers who did not participate in Experiment 1. Raters were recruited through the UC Davis Psychology subjects' pool and received course credit for their participation.

2.2.1.3. Procedure

Raters completed the experiment online, using the Qualtrics survey platform. Before the experiment began, raters completed a sound calibration step where they heard one sentence presented auditorily ("Lubricate the car with grease"), presented in silence at 60 dB, and had to identify the sentence from three multiple choice options. Next, they completed the Perceptual Similarity Ratings ("AXB") task (illustrated in Figure 4). On each trial, raters identify the imitator's token that sounded most similar to the model talker's production. Listeners were told they would hear 3 sounds in a row: the 1st and 3rd were spoken by the same voice, while the 2nd (middle) was spoken by a different voice. Their task was to identify which sound, the 1st or 3rd, is more similar to the middle sound (2nd). The option "1st" and "3rd" was provided on the computer screen as a two-option, forced-choice selection. Raters needed to select one of these options before the experiment would advance to the next trial.

The number of imitators each listener rated was limited in order to keep the experiment a reasonable length, following Pardo et al. (2017). Thus, 14 lists were constructed, containing the full set of stimuli from 2 imitators each. Raters were randomly assigned a list. In total, each list contained 120 AXB similarity ratings (2 imitators x 15 words x 2 model talkers x 2 conversational roles). A listening comprehension question was presented after the experimental trials. In the listening comprehension trial, participants were presented auditorily with a sentence ("Jane heard the pod") and asked to select the final word of the sentence from four multiple choice options ("pod" (correct), "pawn", "pot"; option order randomized across subjects).

[INSERT FIGURE 4 ABOUT HERE]

Figure 4: *AXB trial schematic: raters heard a concatenated recording consisting of "A" (the "1st" sound) (an imitator's pre-exposure production of the word), "X" (model talker's production of the word), and "B" (the "3rd sound") (that imitator's post-exposure production of the word). Raters clicked to indicate whether the 1st sound ("A") or 3rd sound ("B") sounded more similar to "X". (Color online.)*

2.2.2. Statistical Modeling

Mixed effects models are a powerful tool that allows for verification of statistical significance while testing and controlling for effects of multiple variables. In a mixed effects model, predictors that are nested by participant can be included as fixed effects (for example, each AXB task participant heard trials with both types of model talkers and both types of conversational role), which get estimated as traditional regression parameters, and they can also be included as random effects that are allowed to vary around a normal distribution in order to account for participant idiosyncrasy (Baayen et al., 2008). Responses from the AXB ratings task were coded for whether the post-exposure item was selected as more similar to the model talker (=1) or not (=0) and analyzed in a mixed effects logistic regression using the *glmer()* function in the *lme4* package in R (Bates et al., 2014). Estimates for p-values were computed using Satterthwaite approximation in the *lmerTest* package (Kuznetsova et al., 2015). Fixed effects included main effects for Model Talker (Human, Device), Imitator Role (Giver, Receiver), and their interactions. Random effects structure included by-Imitator random intercepts and by-Imitator random slopes for Imitator Role and Model Talker (by-Word and by-Rater random intercepts resulted in singularity errors, indicating model overfit).

Post-hoc analyses were conducted to test whether there is a greater than chance probability that the model production is perceived to be more similar to the post-exposure then pre-exposure imitator production (0.50), using one-sample binomial tests (with the *biom.test()* R base function) on the four subsets of data (Human-Giver, Human-Receiver, Device-Giver, Device-Receiver; p = 0.50), following Miller et al. (2013).

2.2.3. Word List AXB Results

The output of the logistic regression modeling raters' responses is provided in Table 1 and mean AXB responses are plotted in Figure 5. The model computed only a significant main effect of Model Talker on perceived alignment. As seen, participants' post-exposure productions were rated as similar to the Human Model Talker's items, but there was no phonetic alignment toward the Device Model talker. There was no effect of Imitator Role, nor any interaction between Role and Model Talker, on perceived similarity.

The post-hoc binomial tests revealed greater than chance perception of imitation for the Human-Giver trials, with an average of 0.53 (CI: 0.51-0.56, p<0.01). All other Model and Role subsets did not significantly differ from chance.

[INSERT FIGURE 5 ABOUT HERE]

Figure 5: *Mean proportion and standard errors of proportion of "post-exposure" tokens from the Word List task (females only) selected in the AXB ratings task by Model Talker (Human vs. Device), Imitator Conversational Role (Giver, Receiver). (Color online.)*

Fixed Effects	*Est*	*Std.Err*	*z*	*p*
(Intercept)	0.01	0.03	0.47	0.64
Model Talker (Human)	**0.07**	**0.03**	**2.45**	**0.01**
Imitator Role (Giver)	0.02	0.03	0.76	0.45
Model * Imitator Role	0.03	0.03	1.28	0.20

Random Effects	*Variance*	
Imitator		
(Intercept)	7.7e-04	
Model Talker	3.0e-03	
Imitator Role	1.8e-04	
Num. observations = 5,916, *n*=23 imitators, *n*=56 raters,		

Table 1. Summary statistics of the fixed and random effects for the mixed effects logistic regression from the AXB task for Experiment 2.

3. Study II - Map Task Phonetic Imitation Task

Study II extends the questions from Study I, asking whether participant conversational role might influence how people align towards devices and humans in a more complex task. Extending Experiment 1 which held gender constant (examining only female participants), Experiment 3 additionally examines male and female participants.

As with Study I, Study II consists of two experiments. Experiment 3 (Section 3.1.) tests the effect of imitator conversational role (giver vs. receiver), Model Talker (human vs. device), and Imitator Gender (male vs. female) in an interactive Map Task. In Experiment 3, participants completed an interactive fill-in task to place target words on corresponding locations in a 4 x 6 grid of landmarks varying in color and shape. As in Experiment 1, each participant completed the task with two model talkers: one human and one digital voice assistant device interlocutor in both a giver and a receiver role. The task was pre-scripted: participants gave instructions (giver) or confirmations (receiver) that included a target word after the interlocutor produced an utterance with that item. Phonetic alignment was assessed with an AXB similarity ratings task by a separate group of raters (Experiment 4, Section 3.2.).

3.1. Experiment 3: Map Task Production (female and male imitators)

3.1.1. Target words and pre-recorded model talker utterances

Target words consisted of 24 low frequency items, a larger subset taken from Babel (2012): /i/ *cheek, deed, key, peel, teethe, weave*; /æ/ *bat, tap, vat, nag, wag, wax*; /ɑ/ *cot, sock, sod, tot, wad, pod*; /u/ *boot, hoop, zoo, doom, toot, dune*. As in Experiment 1, all model talker productions were pre-recorded and all responses were predetermined (rather than contingent). The same two talkers from Experiment 1 were used to generate model talker recordings (human male, Amazon Polly 'Matthew' voice).

3.1.2. Participants

Participants (n=50; mean age = 20.9 years old, sd = 3.3; 25 female, 25 male) were all native speakers of American English. Experiment 3 extended participation to males as well, to examine whether the same patterns of alignment are seen across genders. They were undergraduates recruited from the UC Davis Psychology subject pool who received course credit for their participation. As in Experiment 1, all except 1 participant reported that they had used a digital voice assistant before participating in the experiment. Participants reported using Siri (40), Alexa

(10), and/or Google Assistant (1). 29 of these participants reported once-a-week usage, 4 reported daily usage, and the rest (7 participants) reported rare usage.

3.1.3. Procedure

The study took place in a sound-attenuated booth, and participants wore a head-mounted microphone (Shure WH20 XLR) and over-ear headphones (Sennheiser Pro). They were seated in a soundbooth facing a computer monitor. They received two blank color-number grids (laminated and attached to a clipboard) labeled for each interlocutor (device ('Matthew') vs. human ('Carl')) for use when instructed in the receiver blocks. As in Experiment 1, participants were given the list of target words on a piece of paper to familiarize themselves prior to the start of the study.

Participants completed the experiment in one continuous session in E-Prime 2.0, consisting of two parts: a pre-exposure block (to get baseline productions prior to the interaction) and a Map Task paradigm (where participants interact with pre-recorded voices to complete a task). Participants began with a **pre-exposure block**, producing target words in a frame ("Repeat the word ___ to me"), using the same design as the pre-exposure in in Experiment 1 (this procedure is schematized in Figure 2 for Experiment 1; after 4.5 seconds, the next target sentence would be shown on the screen (ISI = 300)). Each of the 24 target words were randomly presented in the pre-exposure block.

Next, participants began an **interlocutor block** (either human or device, order counterbalanced). At the beginning of each interlocutor's block, a 'connecting' screen was presented to simulate that a live interaction was beginning: for the human, this consisted of the Skype loading sound, while for the Alexa this consisted of a related 'loading' sound. Then, the interlocutor provided a short introduction (e.g., "Hi! I'm [Carl/Matthew]. I'm a digital device on Amazon products. Today we're going to do a simple task together…."), with an image of a human or an Amazon Echo corresponding to the speaker. The introductions that participants heard are provided in Appendix B.

In each interlocutor block, participants completed two **conversational role sub-blocks**, one as 'giver' of instructions and one as 'receiver' of instructions (order counterbalanced). In each, their task was to complete a novel **Map Task paradigm:** consisting of 3-turn dialogues with the participant and model talkers (human and device) varying conversational role as to where to place target words on a color/shape grid. Similar to Experiment 1 (List Task) the three turns had a consistently structured format: Turn 1 consists of asking where to put a target word on the color/shape grid ("Where should I put…"), Turn 2 consists of instructions to place the target word on a landmark ("Write sod on the green square"), and Turn 3 consists of a confirmation ("Okay, I wrote sod on the green square"). We created two versions of this dialogue, varying conversational role.

When the participant was the 'giver' of information (schematized in Figure 6A), they heard the model talker (human or device) ask where to place the word. Then they read the instructions directly from the screen. Finally, they saw feedback that the interlocutor 'heard' them correctly, showing the word in the correct location on the color/shape grid.

When the participant was the 'receiver' of information (schematized in Figure 6B), they were asked to pick up the clipboard labeled for that interlocutor (either human or device). Each clipboard had a laminated color/shape grid and a dry erase marker. Participants were told that they would ask where to place the word, and then *they* would write that word in the correct location. On each trial, participants began the 3-turn dialogue by reading the question written on the screen

("Where should I put the word, cot."). Next, they heard the instruction from the model talker (human or device). Then, participants wrote the word in the appropriate location and read the confirmation sentence written on the screen (e.g., "Okay, I wrote cot on the green square."). At the end of the 'receiver' block for each interlocutor, participants completely filled out the color/shape grid.

Within each of the interlocutor/conversational role blocks, participants produced all target words (order randomized). The word to shape/color location was pseudorandomized for each of the two interlocutors (device, human) and conversational role blocks. Overall, subjects completed 96 dialogue trials (24 words x 2 conversational roles x 2 model talkers). The experiment took roughly 45 minutes to complete.

[INSERT FIGURE 6 ABOUT HERE]

Figure 6: *Map Task Paradigm 3-turn interaction design. (A) Participant as 'Giver' of information, while (B) Participant is 'Receiver' of information. (i) The first turn consists of asking where to place the word on the color/shape grid. (ii) The second turn consists of instructions for where to write the word. (iii) The third turn consists of a confirmation. When the Model Talker (device vs. human) asks for or gives instructions, their image and the word are displayed on the screen while the pre-recorded utterance is played (either human voice or TTS voice). When the Model Talker 'confirms', they show the word in the correct position on the grid on the screen, seen in (A.iii.). When the participant 'confirms', they write the word on the color/shape location (with a laminated color/shape grid, provided on a clipboard, and dry erase marker) and then read the sentence on the screen to confirm they completed the action (B.iii.). Turns where the participant reads the sentence from the screen are indicated with a microphone and yellow speech bubble. (Color online.)*

3.2. Experiment 4: Map Task AXB Perceptual Similarity Assessment

In order to assess holistic similarity between the post-exposure productions and the model talkers' productions, we conducted an AXB perceptual similarity task, with a separate group of raters.

3.2.1. Map Task AXB Methods
3.2.1.1. Stimuli

The target lexical items from the 50 participants (25 F, 25 M) who completed the Map Task production experiment were extracted from the participants' pre-exposure and post-exposure productions, as well as from the Model Talkers' utterances. The stimuli for the ratings task were prepared following the same procedure from the ratings task for Experiment 2 (Section 2.2.1.1.).

3.2.1.2. Participants and Procedure

A total of 227 raters (mean age = 20.3 years old; 166 female, 61 male) completed the holistic AXB perceptual assessment of the Map Task imitation productions. Raters, none of whom participated in the production studies, were recruited through the UC Davis Psychology subjects' pool, completed the experiment online on Qualtrics, and received course credit for their participation (the ratings task procedure was identical to that used in Experiment 2, section 2.2.1.3.). A total of

25 lists were constructed, containing the full set of stimuli from 2 imitators each. Raters were randomly assigned a list. Each list contained a total of 192 AXB similarity ratings (2 imitators x 24 words x 2 model talkers x 2 conversational roles).

3.2.2. Statistical Analysis

Responses were coded for whether the target item was selected as more similar to the model talker (=1) or not (=0). These data were modeled following the same procedure from Experiment 2 (see Section 2.2.2. for details about the model structure and justification), using a mixed effects logistic regression with fixed effects of Model Talker, Imitator Role, Imitator Gender, and all two-way interaction and the three-way interaction. The random effects structure included by-Rater, by-Word and by-Imitator random intercepts, as well as by-Imitator random slopes for Model Talker and Imitator Conversational Role.

3.2.3. Map Task Results

The summary statistics from the logistic regression are provided in Table 2. The model revealed a significant main effect of model talker. Overall, participants' responses to the Human Model Talker were more similar to that interlocutor's productions than when they responded to the Device Model Talker. There was also a significant two-way interaction between Model talker and Imitator conversational role. This interaction is illustrated in Figure 7. As seen, for the human model talker, both female and male imitators converged when they were in the giver role, and did not align as receiver. There was no alignment toward the digital device model talker in both the giver and the receiver roles. No other main effects or interactions were significant.

Post-hoc analyses tested if listeners perceived similarity of the post-exposure tokens was greater-than-chance, using one-sample binomial tests on the four subsets of data relative to chance (0.50) (Human-Giver, Human-Receiver, Device-Giver, Device-Receiver) (Miller et al., 2013). The post-hoc binomial tests revealed greater than chance perception of similarity for the Human-Giver trials, with an average of 0.51 [CI: 0.50-0.52, $p<0.05$]. The Human-Receiver trials did not differ significantly from 0.50. However, both Device roles showed significantly *less* than chance perception of imitation for the Device-Giver (mean=0.48) [CI: 0.47-0.49, $p<0.001$] and Device-Receiver (mean = 0.49) [CI: 0.48-0.50, $p<0.05$]. Put another way, listeners chose the *baseline* productions as sounding 'more similar' to the model (relative to the Device post-exposure productions).

Fixed Effects	*Est*	*Std.Err*	*z*	*p*
(Intercept)	-0.04	0.04	-1.14	0.25
Model Talker (Human)	**0.03**	**0.01**	**2.01**	**0.04**
Imitator Role (Giver)	0.01	0.01	1.37	0.17
Imitator Gender (Female)	-0.04	0.03	-1.43	0.15
Model * Imitator Role	**0.03**	**0.01**	**2.59**	**<0.01**
Model * Imitator Gender	0.02	0.01	1.36	0.178

Imitator Role * Imitator Gender	0.01	0.01	0.95	0.34
Model * Imitator Role * Imitator Gender	-6.4e-04	0.01	-0.07	0.95
Random Effects	*Variance*			
Imitator				
(Intercept)	0.02			
Model Talker	3.2e-03			
Imitator Role	1.8e-04			
Word (Intercept)	0.02			
Rater (Intercept)	0.01			
Num. observations = 43,212, raters = 227, imitators =50, words = 24				

Table 2: *Summary statistics of the fixed and random effects for the mixed effects logistic regression from the AXB ratings study run on productions from the Map Task.*

[INSERT FIGURE 7 ABOUT HERE]

Figure 7. *Mean proportion and standard errors of proportion of "post-exposure" tokens from the Map Task selected in the AXB ratings task by Model Talker (Human vs. Device) and Imitator Role (Giver, Receiver). (Color online.)*

4. Discussion

The current study tested the effect of participants' conversational role on their phonetic alignment toward voice-AI and human interlocutors. This study addresses a gap in the literature as prior work has not investigated how conversational role influences linguistic alignment in human-computer interaction. We designed two pre-scripted, yet interactive, dialogue tasks to explore this question. The Word List Task (Study I) was a simplistic, utilitarian dialogue where interlocutors coordinate where to place, or confirm, target words on appropriate lists, simulating the types of simple interactions between a user and a digital voice assistant (e.g., "put milk on the grocery list"). The Map Task (Study II) had a more complex and game-like task.

First, we observed differences in phonetic imitation by conversational role, but critically only for the human interlocutor. In the Map Task, we observed greater alignment toward the human (relative to the device) when the participant was in the 'giver' role. While we did not observe role-based differences in the simpler, Word List task (Study I) in the full model, the binomial tests testing difference from chance showed similar patterns: for both studies, there was significantly greater-than-chance perception of imitation for the Human-Giver role. This finding aligns with prior work in human-human interaction showing that conversational role mediates phonetic imitation and, specifically, that participants align toward their interlocutor when giving, than when receiving, information in an interaction (Pardo, 2006; Pardo et al., 2010). Furthermore, since the giver's role is to provide information to their interlocutor, greater alignment in this role can be viewed as facilitating communication, supporting *cognitive-functional* accounts of alignment (e.g.,

Interactive Alignment Account: Pickering & Garrod, 2004). Yet, in the current study, role-based phonetic alignment was only observed in the more interactive task (here, Map Task) in our main modeling, which takes into account speaker variability. Thus, it is possible that the effect of conversational role on phonetic imitation is sensitive to the level of interaction and engagement in the task, as claimed by Pardo et al. (2018). Future work making direct comparisons of task can explore this question further.

Second, across both tasks, we see similar patterns on phonetic imitation based on interlocutor: more phonetic alignment toward the human, relative to device. This interlocutor-based asymmetry can be viewed as supporting predictions made by *socially-mediated accounts* of alignment, such as Communication Accommodation Theory (Shepard, 2001), which propose that linguistic alignment is a way to signal social closeness to an interlocutor. Here, one interpretation is that a voice-AI interlocutor has a lower social status than humans and *lack* of alignment could be seen as reflecting a lack of socially-downward accommodation toward the voice-AI who is the less socially-dominant actor (e.g., Giles, 1973; Giles et al., 1991). This interpretation is supported by prior studies comparing phonetic imitation of digital device and human voices in word shadowing tasks which report less phonetic alignment toward the device voices (Cohn et al., 2019; Snyder et al., 2019) and less phonetic alignment toward less anthropomorphized device systems than devices that have a more embodied human form (Cohn et al., 2020). More specifically, we observe apparent *divergence* toward the device interlocutor observed in the Map Task study: productions toward the device showed significantly *less* than chance perception of alignment. That is, the baseline production sounded more 'similar' to the model talker, indicating that the speaker might have diverged from the device in the interaction. This, too, is in line with *socially-mediated accounts* wherein people diverge to create social distance from an interlocutor they do not feel socially close to.

Our observation of less phonetic alignment toward voice-AI than the human model talker contrasts with *cognitive-functional accounts* (e.g., Pickering & Garrod, 2004; Branigan et al., 2011) that predict greater alignment toward a computer interlocutor. For example, increased lexical alignment toward computer interlocutors is thought to be driven by the perception they are seen as less communicatively capable than humans (Branigan et al., 2011). Here, one possibility is that communicative pressures mediate alignment in different ways for different linguistic features (e.g., syntactic, lexical, phonetic). This is supported by Kim, Horton, and Bradlow (2011) which examined map task dialogues between native and non-native English-speaking interlocutors and found that native English speakers displayed no phonetic alignment toward English learners. Kim et al. (2011) did observe phonetic alignment between pairs where both interlocutors were non-native English speakers and argued that interlocutor language distance was negatively correlated with degree of phonetic alignment. In the current study, a similar might be realized as *greater* phonetic alignment toward the more similar interlocutor (the human voice) relative to the synthesized device voice. Future work examining and comparing both speech and lexical alignment can tease apart whether differential alignment toward computers across studies are indeed due to differential constraints on alignment across linguistic features. If so, this would contrast with theoretical claims that alignment in dialogue applies automatically and across all levels of the grammar (Pickering & Garrod, 2004).

Contra predictions from CASA (Nass et al., 1994) — yet, in line with more recent work proposing that people have routinized ways of engaging with technology that are distinct from human-human interaction (Gambino et al., 2020) — we did not observe 'transfer' of speech behaviors from human-human interaction to voice-AI in our tasks. For one, the *lack* of phonetic

alignment toward the device interlocutors contrasts with prior work showing phonetic alignment to computers (Bell, 2003; Gessinger et al., 2021, 2017; Thomason et al., 2013) and modern voice-AI systems (e.g., Raveh et al., 2019; Cohn et al., 2019; Zellou & Cohn, 2020). Additionally, we did *not* observe transfer of conversational-role based alignment from human-human to human-computer interaction. This contrasts with prior studies that have observed display similar socially-mediated differences in phonetic alignment for human and voice-AI talkers (e.g., more alignment toward male than female voices in Cohn et al., 2019; Snyder et al., 2019). In those studies, participants 'shadowed' single words (repeating after the interlocutor); one possibility is that the more interactive dialogue in the present study highlighted different social constraints on alignment between humans and voice-AI.

There are several limitations of the present study, as well as open questions that provide avenues for future work. For one, it is important to note that in the current studies there were no comprehension errors by either the human or voice-AI in the controlled experimental context. There is some recent work showing differences in alignment following an error made by the voice-AI and human interlocutors at a very high error rate (50% accurate) (Zellou & Cohn, 2020). Indeed, a comprehension error by the interlocutor might increase functional pressures acting on alignment. Future research can test the effect that interlocutor errors have in mediating role-based patterns of alignment toward voice-AI and human interlocutors. Relatedly, one limitation of the present study is that it is a pre-scripted, artificial task. Future work comparing alignment toward humans and voice-AI during unscripted conversations is a next step to understand how these patterns play out in more naturalistic interactions. We also only assessed imitation through AXB perceptual ratings; examining acoustic patterns of alignment can shed light on the patterns of convergence (and potential divergence) observed across interlocutor types and conversational roles.

Another future direction is to investigate alignment across users with different language backgrounds and different age and cognitive profiles (since the current study examined only English-speaking college-age adults). Moreover, in the current study imitators in Experiment 1 (and raters in Experiment 4) were not fully balanced for participant gender. Future work fully balancing imitator, rater, and model talker gender can explore the role that this social factor has on linguistic alignment. An additional limitation is that the different model talkers were cued by differences across multiple auditory and visual properties (different voices and different guises with images of a human vs. device). Thus, whether the differences in phonetic alignment toward the human vs. voice-AI was triggered by voice or guise cues cannot be fully teased apart in this study. Future work varying both voice and label can address this gap. Finally, the present study did not manipulate degree of alignment *by* the interlocutors; mutual alignment and accommodation is a natural behavior between two interlocutors in human-human interaction (e.g., Szabo, 2019). Such behavior might be expected with a text-to-speech (TTS) system, as well. For example, Levitan (2014) found that participants were more likely to entrain to the speech of a TTS system when the interlocutor entrained back toward them. In the present study, the model talker's productions were pre-recorded and not modified to adapt to the participant. These factors raise many avenues for future work to fully understand the conditions under which people do and do not align toward the speech patterns of voice-AI assistants.

4.1. Conclusion

Overall, the current studies contribute to our understanding of conversational-internal factors in alignment by exploring the impact of participant role (as giver or receiver) as well as the type of

interlocutor (human or voice-AI). Together, these studies provide important first steps in comparing alignment across multiple studies, and varying degrees of interactivity, across interlocutor types. At the same time, this work addresses larger questions about the nature of voice-AI interaction. For example, phonetic imitation has been proposed as mechanisms for the spread of sound change (Garrett & Johnson, 2013); will voice-AI influence human speech patterns? As people increasingly use speech to interface with technology, understanding how voice-AI influences human language patterns will be more important.

Acknowledgements

Thank you to editor Stefan Frank and three anonymous reviewers for their helpful comments and feedback. This material is based upon work supported by an Amazon Faculty Research Award to GZ and by the National Science Foundation SBE Postdoctoral Research Fellowship under Grant No. 1911855 to MC. Thanks to Bruno Ferenc Segedin for his assistance in data collection.

References

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

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*(1), 177–189. https://doi.org/10.1016/j.wocn.2011.09.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *ArXiv Preprint ArXiv:1406.5823*.

Bell, L. (2003). *Linguistic Adaptations in Spoken Human-Computer Dialogues—Empirical Studies of User Behavior*. [Doctoral Dissertation, KTH University]. http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-3607

Bentley, F., Luvogt, C., Silverman, M., Wirasinghe, R., White, B., & Lottridge, D. (2018). Understanding the long-term use of smart speaker assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(3), 1–24. https://doi.org/10.1145/3264901

Bilous, F. R., & Krauss, R. M. (1988). Dominance and accommodation in the conversational behaviours of same-and mixed-gender dyads. *Language & Communication*., *8*(3-4), 183-194. https://doi.org/10.1016/0271-5309(88)90016-X

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition*, *121*(1), 41–57. https://doi.org/10.1016/j.cognition.2011.05.011

Clark, H. H., & Murphy, G. L. (1982). Audience design in meaning and reference. *Advances in psychology*, *9*(C), 287–299. https://doi.org/10.1016/S0166-4115(09)60059-5

Clark, H. H., & Schaefer, E. F. (1987). Concealing one's meaning from overhearers. *Journal of Memory and Language*, *26*(2), 209–225. https://doi.org/10.1016/0749-596X(87)90124-0

Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., Aylett, M., Cabral, J.,

Munteanu, C., Edwards, J., & others. (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

Cohn, M., Ferenc Segedin, B., & Zellou, G. (2019). Imitating Siri: Socially-mediated vocal alignment to device and human voices. *Proceedings of the 19th International Congress of Phonetic Sciences*, 1813–1817.

Cohn, M., Jonell, P., Kim, T., Beskow, J., & Zellou, G. (2020). Embodiment and gender interact in alignment to TTS voices. *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, 220–226.

Cohn, M., & Zellou, G. (2019). Expressiveness Influences Human Vocal Alignment Toward voice-AI. *Interspeech 2019*, 41–45. https://doi.org/10.21437/Interspeech.2019-1368

Cowan, B. R., Branigan, H. P., Obregón, M., Bugis, E., & Beale, R. (2015). Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human- computer dialogue. *International Journal of Human-Computer Studies*, *83*, 27–42. https://doi.org/10.1016/j.ijhcs.2015.05.008

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies—Why and how. *Knowledge-Based Systems*, *6*(4), 258–266. https://doi.org/10.1016/0950-7051(93)90017-N

Dias, J. W., & Rosenblum, L. D. (2016). Visibility of speech articulation enhances auditory phonetic convergence. *Attention, Perception, & Psychophysics*, *78*(1), 317–333. https://doi.org/10.3758/s13414-015-0982-6

Edwards, A. (2018). Animals, humans, and machines: Interactive implications of ontological classification. *Human-Machine Communication: Rethinking Communication, Technology, and Ourselves*, 29–50.

Fandrianto, A., & Eskenazi, M. (2012). Prosodic entrainment in an information-driven dialog system. *Thirteenth Annual Conference of the International Speech Communication Association*. 342-345.

Gallois, C., & Giles, H. (2015). Communication accommodation theory. *The International Encyclopedia of Language and Social Interaction*, 1–18. https://doi.org/10.1002/9781118611463.wbielsi066

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: extending the computers are social actors paradigm. *Human-Machine Communication*, *1*(1), 5. https://doi.org/10.30658/hmc.1.5

Garrett, A., & Johnson, K. (2013). Phonetic bias in sound change. *Origins of Sound Change: Approaches to Phonologization*, 51–97.

Gessinger, I., Möbius, B., Fakhar, N., Raveh, E., & Steiner, I. (2019). A Wizard-of-Oz experiment to study phonetic accommodation in human-computer interaction. *International Congress of Phonetic Sciences (ICPhS), Melbourne*, 1475–1479.

Gessinger, I., Raveh, E., Steiner, I., & Möbius, B. (2021). Phonetic accommodation to natural and synthetic voices: Behavior of groups and individuals in speech shadowing. *Speech Communication*, *127*, 43–63. https://doi.org/10.1016/j.specom.2020.12.004

Giles, H. (1973). Accent mobility: A model and some data. *Anthropological Linguistics*, 87–105. https://www.jstor.org/stable/30029508

Giles, H., Coupland, N., & Coupland, I. (1991). Accommodation theory: Communication, context, and. *Contexts of Accommodation: Developments in Applied Sociolinguistics*, *1*.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and

recognition memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *22*(5), 1166–1183. https://doi.org/10.1037//0278-7393.22.5.1166
Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279. https://doi.org/10.1037/0033-295x.105.2.251
Ibrahim, O., Skantze, G., Stoll, S., & Dellwo, V. (2019). Fundamental frequency accommodation in multi-party human-robot game interactions: The effect of winning or losing. *Interspeech*, 3980–3984.
Kim, M., Horton, W. S., & Bradlow, A. R. (2011). Phonetic convergence in spontaneous conversations as a function of interlocutor language distance. *Laboratory Phonology*, *2*(1), 125–156. https://doi.org/10.1515/labphon.2011.004
Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B., & others. (2015). Package 'lmertest.' *R Package Version*, *2*(0).
Levitan, R. (2014). *Acoustic-prosodic entrainment in human-human and human-computer dialogue* [Doctoral Dissertation, Columbia University].
Levitan, R., Gravano, A., Willson, L., Beňuš, Š., Hirschberg, J., & Nenkova, A. (2012). Acoustic-prosodic entrainment and social behavior. *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 11–19.
Miller, R. M., Sanchez, K., & Rosenblum, L. D. (2013). Is speech alignment to talkers or tasks? *Attention, Perception, & Psychophysics*, *75*(8), 1817–1826. https://doi.org/10.3758/s13414-013-0517-y
Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, *21*(4), 422–432. https://doi.org/10.1177/026192702237958
Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78.
Oben, B., & Brône, G. (2015). What you see is what you do: On the relationship between gaze and gesture in multimodal alignment. *Language and Cognition*, *7*(4), 546–562. https://doi.org/10.1017/langcog.2015.22
Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382–2393. https://doi.org/10.1121/1.2178720
Pardo, J. S., Jay, I. C., Hoshino, R., Hasbun, S. M., Sowemimo-Coker, C., & Krauss, R. M. (2013). Influence of Role-Switching on Phonetic Convergence in Conversation. *Discourse Processes*, *50*(4), 276–300. https://doi.org/10.1080/0163853X.2013.778168
Pardo, J. S., Jay, I. C., & Krauss, R. M. (2010). Conversational role influences speech imitation. *Attention, Perception & Psychophysics*, *72*(8), 2254–2264. https://doi.org/10.3758/bf03196699
Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, *79*(2), 637–659. https://doi.org/10.3758/s13414-016-1226-0
Pardo, J. S., Urmanche, A., Wilman, S., Wiener, J., Mason, N., Francis, K., & Ward, M. (2018). A comparison of phonetic convergence in conversational interaction and speech shadowing. *Journal of Phonetics*, *69*, 1–11. https://doi.org/10.1016/j.wocn.2018.04.001
Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*(2), 169–190. https://doi.org/10.1017/S0140525X04450055
Raveh, E., Siegert, I., Steiner, I., Gessinger, I., & Möbius, B. (2019). Three'sa Crowd? Effects of

a Second Human on Vocal Accommodation with a Voice Assistant. *Proc. Interspeech 2019*, 4005–4009.

Scarborough, R., & Zellou, G. (2013). Clarity in communication: "Clear" speech authenticity and lexical neighborhood density effects in speech production and perception. *The Journal of the Acoustical Society of America*, *134*(5), 3793–3807. https://doi.org/10.1121/1.4824120

Shepard, C. A. (2001). Communication accommodation theory. *The New Hand-Book of Language and Social Psychology*, 33–56.

Shockley, K., Baker, A. A., Richardson, M. J., & Fowler, C. A. (2007). Articulatory constraints on interpersonal postural coordination. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(1), 201-208. https://doi.org/10.1037/0096-1523.33.1.201

Snyder, C., Cohn, M., & Zellou, G. (2019). Individual Variation in Cognitive Processing Style Predicts Differences in Phonetic Imitation of Device and Human Voices. *Interspeech 2019*, 116–120. https://doi.org/10.21437/Interspeech.2019-2669

Szabo, I. (2019). Phonetic Selectivity in accommodation: The effect of chronological age. *Proceedings of the 19th International Congress of Phonetic Sciences*, 3195–3199.

Thomason, J., Nguyen, H. V., & Litman, D. (2013). Prosodic entrainment and tutoring dialogue success. *International Conference on Artificial Intelligence in Education*, 750–753. https://doi.org/10.1007/978-3-642-39112-5_104

Yu, A., Abrego-Collier, C., Baglini, R., Grano, T., Martinovic, M., Otte III, C., Thomas, J., & Urban, J. (2011). Speaker attitude and sexual orientation affect phonetic imitation. *University of Pennsylvania Working Papers in Linguistics*, *17*(1), 235-242. https://repository.upenn.edu/pwpl/vol17/iss1/26

Zellou, G., & Cohn, M. (2020). Social and functional pressures in vocal alignment: Differences for human and voice-AI interlocutors. *Proc. Interspeech 2020*, 1634–1638.

Zellou, G., Cohn, M., & Ferenc Segedin, B. (2020). Age-and gender-related differences in speech alignment toward humans and voice-AI. *Frontiers in Communication*, *5*:600361, 1-11. https://doi.org/10.3389/fcomm.2020.600361

Zellou, G., & Scarborough, R. (2019). Neighborhood-conditioned phonetic enhancement of an allophonic vowel split. *The Journal of the Acoustical Society of America*, *145*(6), 3675–3685. https://doi.org/10.1121/1.5113582

Zellou, G., Scarborough, R., & Nielsen, K. (2016). Phonetic imitation of coarticulatory vowel nasalization. *The Journal of the Acoustical Society of America*, *140*(5), 3560–3575. https://doi.org/10.1121/1.4966232