Social and functional pressures in vocal alignment: Differences for human and voice-AI interlocutors

Georgia Zellou and Michelle Cohn

UC Davis, Phonetics Lab, Department of Linguistics, Davis, CA, USA
gzellou | mdcohn @ucdavis.edu

Abstract

Increasingly, people are having conversational interactions with voice-AI systems, such as Amazon's Alexa. Do the same social and functional pressures that mediate alignment toward human interlocutors also predict align patterns toward voice-AI? We designed an interactive dialogue task to investigate this question. Each trial consisted of scripted, interactive turns between a participant and a model talker (pre-recorded from either a natural production or voice-AI): First, participants produced target words in a carrier phrase. Then, a model talker responded with an utterance containing the target word. The interlocutor responses varied by 1) communicative affect (social) and 2) correctness (functional). Finally, participants repeated the carrier phrase. Degree of phonetic alignment was assessed acoustically between the target word in the model's response and participants' response. Results indicate that social and functional factors distinctly mediate alignment toward AI and humans. Findings are discussed with reference to theories of alignment and human-computer interaction.

Index Terms: vocal alignment, human-computer interaction, social vs. functional pressures

1. Introduction

Interacting talkers systematically align toward the acoustic-phonetic patterns of each other's speech to sound more alike. This phenomenon is known as vocal alignment (also phonetic imitation) and has been well reported in the literature [1]–[3]. There is also evidence of vocal alignment toward speech generated by non-human entities: people align toward the speaking rate [4] and amplitude [5] of synthetic computer voices, and even toward speech patterns of modern voice-activated artificially intelligent (voice-AI) systems (e.g., Apple's Siri, Amazon's Alexa [6]–[10]). That we see an application of this human behavior toward technology is in line with the 'Computers as social actors' theory (CASA, [11], [12]), which posits that people's behavior during interactions with technology mirrors their behavior toward humans.

Above and beyond documenting the presence of alignment toward computers and voice-AI, a growing body of work has suggested that the *magnitude* of speech alignment may differ by interlocutor: individuals tend to show weaker vocal alignment toward voice-AI, relative to human interlocutors [6], [8], [9]. Conversely, others have found that people adopt the lexical choices [13] and syntactic structures [14] for the computers to a greater extent than for human interlocutors. Why do we see these conflicting results? Aside from this alignment occurring at different linguistic levels (e.g., phonetic vs. syntactic), one possibility is that modern voice-AI has different *functional* and *social* pressures in communication than computer systems or avatars. For one, the primary way we communicate with voice-AI systems is using speech, a uniquely human form of communication. Also, unlike computer systems in the past, voice-AI systems have improved text-to-speech (TTS) synthesis [15] and automatic speech recognition (ASR) abilities; therefore, the functional pressures may be lessened. This might explain mixed findings of less alignment toward voice-AI, but greater alignment toward computer systems. On the other hand, modern voice-AI systems exhibit greater *social* cues, such as having names (e.g., "Alexa") and apparent genders. There is even evidence that humans engage with voice-AI in for purely social purposes, for example having short, non-utilitarian conversations with "chatbots" [16]. Therefore, another possibility is that humans will respond to apparent sociality in voice-AI more similarly as they would toward another human. An alternative possibility, following the 'Uncanny Valley of the Mind' framework [17], [18] is that as cues of human-likeness near 'real' human levels, it can trigger feelings of uneasiness or 'uncanniness' toward robots.

The current study investigates the nature of human-AI and human-human interaction by examining whether degree of vocal alignment is different based on social and functional pressures within a verbal exchange. Moment-by-moment, the dynamics of an interaction vary: speakers may become more animated and emotionally expressive (social) or make errors that require correction (functional). Parametrically manipulating these factors across interlocutors (human vs. AI) can pinpoint differences, and similarities, in the way humans engage with voice-AI, relative to humans, and can shed light more broadly on the nature of human-AI interaction. Furthermore, this study can contribute to understanding the mechanisms of vocal alignment, as driven by social dynamics (§1.1.) and/or pressures to improve intelligibility (§1.2.).

1.1. Social factors

Synchrony between interlocutors is thought to serve as 'social glue'. 'Communication Accommodation Theory' (CAT [19]) proposes that alignment is used to foster social closeness: the gender [1], regional affiliation [20], attractiveness [1], and emotional state [7] of interlocutors predicts alignment. In the present study, one social dimension is whether the interlocutor is human or AI; here, as in prior work, we might predict less alignment toward voice-AI, relative to humans [6], [8].

Further, the current study manipulates *within* speaker social dynamics by including an expressive interjection in some model talker utterances. Interjections, or emotively used words such as "Yipee!" or "Darn!", are conventionalized phrases. Interjections often provide no additional linguistic meaning to an utterance; their function is purely social in that they convey the speaker's emotional or cognitive state [21], [22]. There is also some evidence that individuals are sensitive to these interjections, even when they are produced by voice-

AI: [7] found that participants aligned more to Amazon's Alexa voice when they shadowed interjections realized in hyper-expressive prosody, relative to Alexa's neutral prosody.

Therefore, in the current study, we utilize these expressive interjections, not as target words, but as additions to responses made by the model talker as a way to increase the socio-communicative force of the utterance. If increased expressiveness modulates vocal alignment, we predict that participants will align toward the interlocutor to a greater extent when the interlocutor's response contains an expressive interjection (e.g., "Super! I think I heard boot"), relative to when they do not (e.g., "I think I heard boot"). Yet, interjections add no semantic content to the utterance. If the addition of these expressive-encoding items does not add to the sociality of the interaction, we predict no difference in alignment patterns when they are present or not.

With respect to social factors, we might also predict different alignment patterns toward the human and voice-AI model talker as a function of the socio-communicative expressiveness of the interaction. For one, modern-day voice-AI can play a social role for humans. As previously mentioned, voice-AI systems are increasingly assuming more human-like qualities (e.g., more realistic voices, better speech recognition, etc.). Thus, one prediction is that alignment toward AI will increase with interactions containing interjections, paralleling what we expect for alignment toward human interlocutors. This would support computer personification frameworks, e.g., CASA [11]. Alternatively, displays of socio-expressiveness might be negatively perceived by participants. This would support observations of an 'uncanny valley' in people's behavioral responses toward technology which take on near-human qualities [17]. For example, [18] observed that conflicting cues to human-likeness in a non-human entity violates people's expectations of technology and leads to feelings of uneasiness or discomfort toward voice-AI. Therefore, we might predict that when the AI model talker displays highly expressive socio-communicative responses, participants will diverge, doing the opposite of how we expect them to behave toward the human interlocutor in more expressive interactions (converge).

1.2. Functional factors

Other accounts propose that alignment serves a functional, intelligibility-driven role: to help a speaker be better understood by their listener by matching their linguistic representations for better mutual intelligibility [23], [24]. For example, participants who actively imitate novel speech patterns later display improved recognition of that speech (unfamiliar accent [25]; disordered speech [26]). Furthermore, [27] assessed alignment between dyads completing a map task. They found that when participants were giving information during the task, the 'giver' displayed greater vocal alignment than the 'receiver'. These results lead us to predict that speakers will actively align toward the speech of another talker when there is strong pressure to be more intelligible.

Functional pressures are also very much present in interactions with technology: people adopt the lexical choices [13] and syntactic structures [14] of a computer system during interactions to a greater extent than toward a human under similar conditions. These findings suggest that linguistic behavior toward AI is distinct from that toward human interlocutors, in some cases triggering less alignment toward voice-AI (speech alignment) and in some cases more alignment toward computers (lexico-syntactic alignment). One possibility is that the social and functional pressures in the different paradigms from prior studies led to greater alignment toward the human or the technological interlocutor. A more formal investigation into which of these factors leads to more or less alignment toward voice-AI, relative to humans, can inform our understanding of the mechanisms of vocal alignment and the dynamics of human-computer interactions.

For one, we expect people to align in different ways toward human and voice-AI model talkers as a function of differing intelligibility pressures in the interaction. Voice-AI systems are often used for practical and utilitarian tasks, with the user usually in the 'giver' role when producing an utterance (e.g., "set a timer", "play a song", "tell me the weather"). Thus, we might expect that vocal alignment patterns toward AI will be more intelligibility-motivated, relative to those toward a human, reflecting the utilitarian purpose of voice-AI as task-oriented interlocutors. This would explain studies comparing alignment toward computers and humans which find *greater* alignment toward the computer [14] where the task was goal-oriented: the bias toward greater alignment toward technology could be explained by stronger functional pressures.

Thus, the current study examines whether a correct or incorrect response from the interlocutor influences patterns of phonetic imitation. If feedback about intelligibility influences alignment, we predict that a trial containing an uncertain response from the interlocutor, that they have not with certainty understood the target word, will elicit more robust phonetic imitation by participants than trials where the interlocutor correctly responds. For our predictors, we expect to see an effect of *Correctness* on degree of alignment.

1.3. Present study

The present study examines identical communicative interactions in a laboratory setting, varying social and functional factors, across human-human and human-AI interactions. To our knowledge, no prior work investigates the relative weight of these social and functional factors *within* an interaction with a single interlocutor; most vary the interlocutor as a way of assessing social variables (e.g., gender, attractiveness, [1]). Doing so allows us to probe both social and functional pressures that may differentially, or similarly, affect human-human and human-AI interaction.

2. Methods

2.1. Participants, Stimuli, and Procedure

Participants (n=54; 27 F) were native English speakers (mean age=20.2 years old, sd=2.4) recruited from the UC Davis subject pool. None reported having any hearing impairments.

Target words were 16 low frequency CVC items selected from [1]: *bat, boot, cheek, coat, dune, hoop, moat, pod, tap, toot, tot, weave, cot, soap, deed, sock*. Both model talkers (human and AI) produced the target words in a 'correct' frame ("I think I heard sock") and one in an 'incorrect' frame ("I think I heard sock or sack"), where the distractor was a minimal pair differing in vowel backness (order of the target word and minimal pair were counterbalanced across trials and differed across interlocutors). For the human model talker, a female native English speaker recorded all utterances in a sound attenuated booth with a head-mounted microphone. For the voice-AI model talker, we generated recordings by default

female Alexa voice (US-En) using the Alexa Skills Kit. Both talkers produced all sentences in their correct and incorrect forms in a neutral speaking style, as well as all introductions, voice-overs, and final responses.

Both model talkers also produced 16 interjections (e.g., "Yipee!") in a hyper-expressive manner, balanced by valence (8 positive, 8 negative). These were produced naturally by the human talker; for the Alexa voice, emotionally expressive interjections recorded by the Alexa voice actor, or 'Speechcons', were added to the TTS output using speech synthesis markup language (SSML) tags (a limitation of TTS is that emotion is otherwise difficult to synthesize).

Participants sat in a sound-attenuated booth, wearing a head-mounted microphone and headphones, facing a computer screen. Subjects first were presented a screen introducing them to the two model talkers: an Amazon Echo device named "Alexa" and a female human named "Melissa", with images of them. Then, participants produced baseline productions of target words, reading a carrier sentence containing the target word, "The word is __". Subjects completed two baseline blocks, where order of sentences were randomly presented.

Next, participants completed either the AI or Human shadowing blocks (order counterbalanced between subjects). Instructions were given at the beginning of each shadowing block. The interlocutors (AI, Human) were introduced ("Hi! I'm Melissa. I'm a research assistant in the Phonetics Lab." / "Hi! I'm Alexa. I'm a digital device through Amazon.") and went through a voice-over example. During each trial, the target words occurred in pre-scripted dialogues between the participant and the model talker. Each trial contained multiple turns: ***1. Initial turn***: participants saw a phrase containing the target word printed on the screen and read it aloud (e.g., "The word is weave."). ***2. Interlocutor response turn***: The interlocutor provided a response, telling them what she 'heard'. ***3. Participant shadowing response turn***: Participants repeated the sentence a second time. ***4. Interlocutor concluding turn:*** The interlocutor provided final feedback (e.g., "Great!", "Thanks, got it!", "Perfect!", etc.; randomly presented).

Subjects completed two interlocutor blocks, where trials were manipulated to vary in the social and functional properties of Interlocutor response turns. To manipulate **intelligibility pressures**, there were two correctness conditions: In 50% of trials, the interlocutor responded with the correct target word (e.g., "I think I heard weave.") while in 50% of trials the interlocutor did not correctly identify the target word with certainty (e.g., "I'm not sure I understood. I think I heard weave or wove."). (Note that while incorrect responses differed from correct responses in information structure, the former eliciting 'corrective focus', our critical prediction is that there are differences in alignment *across interlocutor types* for a given response type.)

To manipulate **socio-expressiveness**, there were two expressiveness conditions: In 50% of trials, the interlocutor responses took the form described above; in the other 50%, the interlocutor response contained an emotionally expressive interjection that corresponded to the intended sentiment of the response: either a positive interjection (e.g., "Yipee! I think I heard sock.") for correct turns, or a negative interjection (e.g., "Darn! I'm not sure I understood. I think I heard sock or sack.") for incorrect turns. Within each block, correctness and expressiveness trials were intermixed and randomly presented.

There were 128 trials (16 items, 2 correctness conditions, 2 expressiveness conditions, 2 interlocutors).

Figure 1 presents a sample trial and the four response conditions for turn 2, varying by expressiveness and correctness.

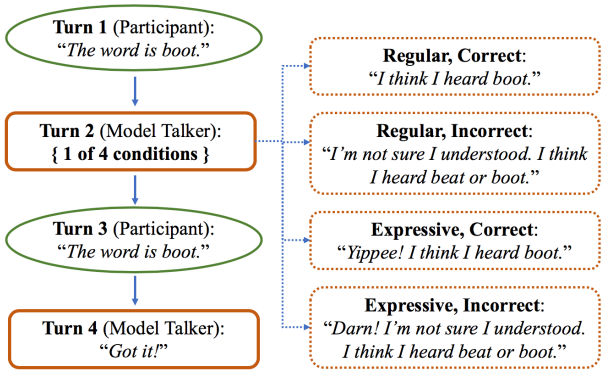

Figure 1: *Sample dialogue of a trial.*

2.2. Acoustic Analysis

Interlocutor and participant recordings were force-aligned with FAVE [28] and segment boundaries were hand-corrected by phonetically-trained research assistants. Vowel duration was measured for each target word vowel from the model talker response turn (turn 2) and the participant's response (turn 3). We assessed degree of alignment with *Difference in Distance* (DID) = |baseline-model|-|shadowed-model| [1]. This relative difference score reflects overall alignment, taking into account baseline similarity between participants and model talkers. Positive DID values indicate alignment toward, while negative values indicate divergence from, the model talker, relative to participants' baseline productions.

2.3. Statistical Analysis

We modeled DID duration values in a linear mixed effects model. The model was run in R using the *lmer()* function in the *lme4* package [29]. Fixed effects included Model Talker (2 levels: AI, human), Correctness (2 levels: correct, incorrect), Expressiveness (2 levels: regular, expressive). The model included all two-way interactions, as well as the three-way interaction, between predictors. Predictors were sum-coded. By-participant and by-item random intercepts and by-participant random slopes for each main effect, each two- and three-way interactions were also included.

3. Results

There was a significant main effect of Correctness [β=-1.52, t=-2.64, p<0.05]: overall, participants displayed greater alignment when the model talker produced an uncertainty response than for a correct response. Correct responses made by the Human, and all responses made by the voice-AI, triggered less alignment to the model talkers' productions. There was no effect of Expressiveness (p=.14).

Furthermore, the model computed a significant three-way interaction between Model talker, Expressiveness, and Correctness [β=-2.1, t=-4.3, p<0.001]. Figure 2 displays this interaction. For one, speakers align more toward voice-AI in Incorrect trials that include the expressive interjection (e.g., "Darn! I'm not sure I understood. I heard beat or boot."). Meanwhile, they also display *less* alignment toward the human

in Correct trials without interjections (e.g., "I'm not sure I understood..."). No other interactions were observed.

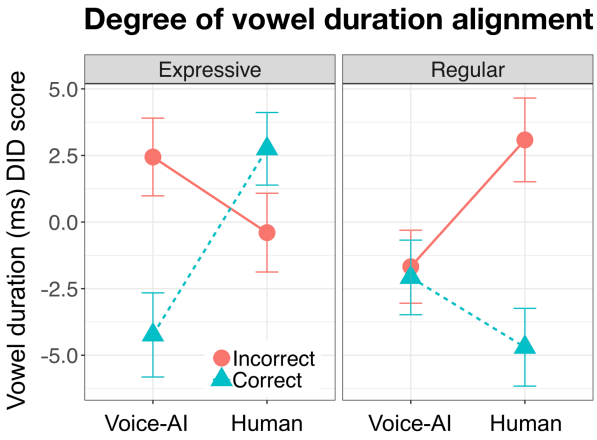

Figure 2: *Vowel duration DID scores (means and standard errors of the mean) to AI and Human model talkers' target words, by Correctness and Expressiveness.*

4. Discussion

The current study investigates how social and functional factors mediate vocal alignment toward a human and voice-AI interlocutor in an interactive shadowing task. Overall, we observe that functional pressures do play a role in predicting degree of alignment during this task: participants displayed greater alignment when the model talker responded with uncertainty about the correct target word than when the model talker responded unequivocally with the correct target word. This observation supports theories that functional pressures in an interaction influence speech alignment [23], [24].

Table 1 summarizes the interaction between social and functional factors and interlocutor type on alignment patterns.

Table 1: *Summary of alignment patterns seen in this study.*

	Incorrect (+ functional pressure)	**Correct (- functional pressure)**
Regular (less social)	Converge to human Diverge from AI	No alignment (divergence)
Expressive (more social)	Converge to AI No alignment to human	Converge to human Diverge from AI

First, we observe converge toward the human interlocutor under two conditions: when there is pressure to be intelligible but with a non-expressive response and with a socially expressive response with no pressure to be intelligible. These are also the conditions where we observe the greatest *divergence* from the voice-AI interlocutor. That we see these *simple* factors leading to convergence toward the human, but divergence from voice-AI, is contrary to prior reports of greater alignment toward computers than toward human interactors [13], [14]. Yet, it does align with recent work reporting less alignment toward voice-AI systems, relative to naturally produced human voices [6], [8], [9]. The observation of the same factors leading to different patterns of alignment toward humans and voice-AI does not support predictions made by theories of computer personification, e.g. CASA [11], that people automatically behave toward technology as a person. Rather, we observe distinct patterns of speech alignment that people apply toward voice-AI systems.

Additionally, and further in support of distinct vocal behavior toward human and voice-AI interlocutors, we observe that participants do align toward voice-AI only with additive social *and* functional pressures. Thus, one interpretation for this is that in order for functional pressures to trigger alignment toward voice-AI, the system needs to display even *more* social signals than for a human interlocutor. This is not what we would expect from an 'Uncanny Valley' hypothesis for human-computer interaction [18]: that people feel discomfort when technology displays near-humanlike behavior. We expected highly emotive responses from the AI system, containing expressive interjections, to be perceived as off-putting, or uncanny, by participants. Yet, these interactions received the most alignment for the voice-AI. This also aligns with recent work showing greater alignment toward these interjections during word shadowing [7] and improved user ratings during conversational interactions with a chatbot that produced these interjections compared to when it did not [30]. Thus, contra an 'Uncanny Valley' hypothesis, people respond positively to socially-expressive utterances from voice-AI.

[23] argue that alignment during interactive dialogue is automatic and facilitates comprehension by converging interlocutors' linguistic representations. Our findings do not support such a strong stance for how functional pressures modulate alignment. Both the type of interlocutor and the socio-communicative force of the interaction mediate how functional pressures influence alignment.

Future work could also examine how other linguistic factors mediate alignment toward voice-AI. For example, the current study used low frequency words, which are shown to be more susceptible to exposure. Comparing imitation of high versus low frequency items across interlocutor types could reveal differences in representational factors at play during these interactions, cf. [2]. Also, in the current study we presented comprehension errors with minimal pairs differing in vowel backness. Future work examining how varying types of phonological confusions trigger different types of alignment patterns across interlocutor types could also tease apart what linguistic-communicative pressure influence speech alignment. Other phonetic variables (e.g., formant frequencies, pitch, and intensity) could also be explored in future work.

Investigations of vocal alignment toward voice-AI can present novel tests and theoretical extensions to human-AI interaction frameworks. As speech becomes a more dominant mode of interfacing with technology, understanding how voice-AI systems influence human language patterns will be more important. Humans often interact with voice-AI in the more functional, 'giver' role. Our findings suggest this could lead to more alignment toward voice-AI, if it is expressive. The findings from the current study also have implications for voice user interface design. For one, based on a greater degree of alignment, including expressive interjections in interactions with AI appears to improve user responsiveness and interactive behavior toward the voice-AI interlocutor. Furthermore, this study suggests that the communicative success of the interaction (e.g., presence of ASR errors) might dynamically interact with apparent *sociality* of the system, and shape the extent users apply human-human speech rules to their interactions with voice-AI.

5. Acknowledgements

This material is based upon work supported by the National Science Foundation SBE Postdoctoral Research Fellowship under Grant No. 1911855 to MC.

6. References

[1] M. Babel, "Evidence for phonetic and social selectivity in spontaneous phonetic imitation," *J. Phon.*, vol. 40, no. 1, pp. 177–189, Jan. 2012, doi: 10.1016/j.wocn.2011.09.001.

[2] S. D. Goldinger, "Words and voices: episodic traces in spoken word identification and recognition memory," *J. Exp. Psychol. Learn. Mem. Cogn.*, vol. 22, no. 5, pp. 1166–1183, Sep. 1996, doi: 10.1037//0278-7393.22.5.1166.

[3] J. S. Pardo, "On phonetic convergence during conversational interaction," *J. Acoust. Soc. Am.*, vol. 119, no. 4, pp. 2382–2393, Apr. 2006, doi: 10.1121/1.2178720.

[4] L. Bell, "Linguistic Adaptations in Spoken Human-Computer Dialogues - Empirical Studies of User Behavior," 2003, Accessed: Apr. 15, 2020. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-3607.

[5] N. Suzuki and Y. Katagiri, "Prosodic alignment in human–computer interaction," *Connect. Sci.*, vol. 19, no. 2, pp. 131–141, Jun. 2007, doi: 10.1080/09540090701369125.

[6] E. Raveh, I. Siegert, I. Steiner, I. Gessinger, and B. Möbius, "Three'sa Crowd? Effects of a Second Human on Vocal Accommodation with a Voice Assistant," *Proc Interspeech 2019*, pp. 4005–4009, 2019.

[7] M. Cohn and G. Zellou, "Expressiveness Influences Human Vocal Alignment Toward voice-AI," in *Interspeech 2019*, Sep. 2019, pp. 41–45, doi: 10.21437/Interspeech.2019-1368.

[8] M. Cohn, B. Ferenc Segedin, and G. Zellou, "Imitating Siri: Socially-mediated vocal alignment to device and human voices," *Proc. 19th Int. Congr. Phon. Sci.*, pp. 1813–1817, 2019.

[9] C. Snyder, M. Cohn, and G. Zellou, "Individual Variation in Cognitive Processing Style Predicts Differences in Phonetic Imitation of Device and Human Voices," in *Interspeech 2019*, Sep. 2019, pp. 116–120, doi: 10.21437/Interspeech.2019-2669.

[10] K. Metcalf *et al.*, "Mirroring to build trust in digital assistants," *ArXiv Prepr. ArXiv190401664*, 2019.

[11] C. Nass, J. Steuer, and E. R. Tauber, "Computers are social actors," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1994, pp. 72–78.

[12] C. I. Nass, Y. Moon, and J. Morkes, "Computers Are Social Actors: A Review of Current," *Hum. Values Des. Comput. Technol.*, no. 72, p. 137, 1997.

[13] S. E. Brennan, "Lexical entrainment in spontaneous dialog," *Proc. ISSD*, vol. 96, pp. 41–44, 1996.

[14] H. P. Branigan, M. J. Pickering, J. Pearson, and J. F. McLean, "Linguistic alignment between people and computers," *J. Pragmat.*, vol. 42, no. 9, pp. 2355–2368, 2010.

[15] A. van den Oord *et al.*, "Wavenet: A generative model for raw audio," *ArXiv Prepr. ArXiv160903499*, 2016.

[16] A. Ram *et al.*, "Conversational ai: The science behind the alexa prize," *ArXiv Prepr. ArXiv180103604*, 2018.

[17] M. Mori, "The Uncanny Valley: The Original Essay by Masahiro Mori - IEEE Spectrum," *IEEE Spectr.*, p. 6, 2017.

[18] R. K. Moore, "A Bayesian explanation of the 'Uncanny Valley'effect and related psychological phenomena," *Sci. Rep.*, vol. 2, p. 864, 2012.

[19] C. A. Shepard, "Communication accommodation theory," *New Hand-Book Lang. Soc. Psychol.*, pp. 33–56, 2001.

[20] R. Y. Bourhis and H. Giles, "The language of intergroup distinctiveness," *Lang. Ethn. Intergroup Relat.*, vol. 13, p. 119, 1977.

[21] F. Ameka, "Interjections: The universal yet neglected part of speech," *J. Pragmat.*, vol. 18, no. 2–3, pp. 101–118, 1992.

[22] E. Goffman, *Forms of talk*. University of Pennsylvania Press, 1981.

[23] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behav. Brain Sci.*, vol. 27, no. 2, pp. 169–190, 2004.

[24] S. Garrod and G. Doherty, "Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions," *Cognition*, vol. 53, no. 3, pp. 181–215, 1994.

[25] P. Adank, P. Hagoort, and H. Bekkering, "Imitation improves language comprehension," *Psychol. Sci.*, vol. 21, no. 12, pp. 1903–1909, 2010.

[26] S. A. Borrie and M. C. Schäfer, "The role of somatosensory information in speech perception: Imitation improves recognition of disordered speech," *J. Speech Lang. Hear. Res.*, vol. 58, no. 6, pp. 1708–1716, 2015.

[27] J. S. Pardo, I. C. Jay, and R. M. Krauss, "Conversational role influences speech imitation," *Atten. Percept. Psychophys.*, vol. 72, no. 8, pp. 2254–2264, Nov. 2010, doi: 10.3758/bf03196699.

[28] I. Rosenfelder, J. Fruehwald, K. Evanini, and J. Yuan, "FAVE (forced alignment and vowel extraction) program suite," *URL Httpfave Ling Upenn Edu*, 2011.

[29] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *ArXiv Prepr. ArXiv14065823*, 2014.

[30] M. Cohn, C.-Y. Chen, and Z. Yu, "A Large-Scale User Study of an Alexa Prize Chatbot: Effect of TTS Dynamism on Perceived Quality of Social Dialog," in *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, 2019, pp. 293–306.