IMITATING SIRI: SOCIALLY-MEDIATED VOCAL ALIGNMENT TO DEVICE AND HUMAN VOICES

Michelle Cohn, Bruno Ferenc Segedin, and Georgia Zellou

University of California, Davis
mdcohn@ucdavis.edu, bferencsegedin@ucdavis.edu, gzellou@ucdavis.edu

ABSTRACT

The current study explores the extent to which humans vocally align to digital device voices (i.e., Apple's Siri) and human voices. First, participants shadowed word productions by 4 model talkers: a female and a male digital device voice, and a female and a male real human voice. Second, an independent group of raters completed an AXB task assessing perceptual similarity between imitators' pre- and post-exposure items to model talkers' productions. Results show that people *do* imitate device voices, but to a lesser degree than they imitate real human voices. Furthermore, similar social factors mediated vocal imitation toward both device and human voices: people imitated male device and human voices to a greater extent than female device and human voices.

Keywords: phonetic imitation, speech production, human-device interaction

1. INTRODUCTION

Speakers adjust the acoustic-phonetic properties of their productions, often aligning to many of the vocal characteristics of their interlocutor [1, 7, 10, 15]. The extent to which phonetic imitation patterns are driven by automatic updating of experiences to speech representations or socially-mediated mechanisms, however, is a source of theoretical debate. One way to test these differing stances is to investigate imitation toward a new type of interlocutor in our speech communities: digital devices.

Talking to hyper-naturalistic voice-activated and artificially intelligent digital devices, such as Apple's Siri or Amazon's Alexa, is becoming a ubiquitous daily behavior for many individuals [11, 14, 18]. However, the impact of human-device interactions on *humans' linguistic patterns* is a vastly understudied area. Prior work has shown that individuals align with the linguistic properties of computer voices, such as speech rate [3] and syntactic structure [4]. Human-device alignment is one phenomenon that can be revealing to the underlying mechanisms of phonetic imitation. Studying human imitation of device speech patterns could also be relevant for theories of sound change. Imitation has been hypothesized to play a role in the spread of sound change, e.g., [6, 19]. If humans imitate the acoustic properties of devices, then characterizing the nature of these imitations could be relevant for understanding the role devices may have in affecting human speech patterns.

1.1. Representationally-mediated alignment

People pay attention to and encode many acoustic properties of the voices they hear, even in situations where these acoustic properties are seemingly irrelevant to the task they are engaged in, as when hearing isolated words through headphones in a research lab [9, 16]. Hence, one proposal is that alignment may be driven by the nature of mental representations for words. For example, exemplar-based phonological theories [10, 12, 17] predict that vocal imitation is a natural consequence of the episodic nature of speech representations. Goldinger [10] demonstrated that degree of imitation toward various talkers' voices varied with the usage frequency of the words being shadowed: Low-frequency words were imitated more strongly than high-frequency words. This result is expected if, following an exemplar account, speech production patterns are influenced by the constellation of past memory traces associated with a given word. Because people have relatively fewer encounters with low-frequency words, their exposure to such words in the study would have disproportionate influence on their subsequent pronunciation of those words, compared to exposure to high-frequency words. Along these lines, a strong lexical-representation perspective might propose that only degree of exposure to a given word by a given talker, *not the type of interlocutor producing that word*, would influence degree of imitation [10]. Hence, for the present study, this type of representational account might predict that speakers will imitate words (here, we select only low-frequency items to increase the likelihood of imitation) spoken by digital device voices to the same degree as they would for those words spoken by human voices if the degree of exposure to those words by the speakers is equal.

1.2. Socially-mediated alignment

A second type of account of the mechanisms involved in vocal alignment comes from Communication Accommodation Theory (CAT) [8]. CAT proposes that vocal alignment is a socially mediated process by which speakers' goal is to emphasize or minimize social differences between themselves and their interlocutors via speech and language patterns. This view predicts that vocal imitation would be mediated by social properties of the speaker and their interlocutor. Supporting this is evidence that social attributes of interlocutors mediate degree of vocal imitation. For instance, in a direction-giving map-task, male speakers showed greater convergence toward their interlocutor [15]. These gender effects are also mediated by role: female speakers converged toward the interlocutor who was *receiving* instructions, while male speakers converged toward the instruction-*giver* [15]. Additionally, it has also been shown that rated attractiveness of the model talker plays a role in degree of alignment [1]. Thus, while phonetic imitation appears to be a robust phenomenon, it is not simply fully automatic, or mediated via experiential factors alone, as evidenced by these various social factors that have been shown to influence patterns of imitation.

The present study investigates this social account by examining whether "humanity" of the interlocutor mediates imitation: will shadowers imitate a digital device voice to a different extent than they imitate a real human voice? We might predict, based on CAT, that people imitate human voices more than device voices, if the goal of imitation is to minimize social distance between themselves and other humans, but not necessarily artificially intelligent entities.

1.2.1 Are digital devices social actors?

Above and beyond the simple empirical question of whether people phonetically imitate the vocal properties of device and human voices in a similar manner, there is the question of whether, if device-imitation is seen, humans apply the same socially-mediated imitation patterns to the apparent-social attributes of the device voices. Specifically, we ask whether people imitate female and male device voices following the same patterns of gender-mediated imitation of human voices, e.g. [15]. Theoretical accounts of computer personification predict that when a person detects any sense of humanity in a digital system, they will automatically begin to treat the computer as a person by applying human social rules and norms (e.g., Computers are Social Actors framework: [13]). For voice-AI digital devices, these "cues" of humanity are more robust than for voice-based computer avatars in the past; modern voice-AI systems often have names, apparent genders, and personas. Critically, these devices also differ from earlier voice technology because the way we interact with them is primarily through speech, a uniquely human mode of communication.

On the one hand, we might predict that speakers will show similar patterns of gender-mediated imitation for *both* humans and devices. This would support a stance that people indeed apply human social rules to devices, a prediction in line with computer personification frameworks [13]. On the other hand, we might predict that speakers have distinct ways interacting with digital devices, relative to how they interact with humans, and this will be reflected in their phonetic convergence patterns. Therefore, we might see gender-mediated imitation patterns only toward human voices, and not toward device voices.

1.3. Current Study

To examine the mechanisms underlying speech imitation and to test our specific hypotheses about the nature of vocal alignment toward devices, relative to humans, we conducted a lexical shadowing study (2.1). Participants shadowed single word productions from female and male real human voices, as well as female and male digital device voices, while viewing a corresponding image of either a human face or a device. Degree of imitation in these productions was then assessed using an AXB similarity task from an independent group of raters (2.2).

2. METHODS

2.1. Shadowing paradigm

2.1.1. Stimuli

Target words consisted of 12 monosyllabic real English words containing a vowel-nasal (VN) sequence. The stimulus list, presented in Table 1, consisted of low usage frequency items (mean log frequency: 1.6, range: 1.1-2.1, taken from SUBTLEX [5]) since prior work has observed imitation is most robust for low-frequency words.

Table 1: Target words used for stimulus items.

bomb	sewn	vine	pun	yawn	shun
chime	shone	wane	tame	wren	hem

Stimulus items consisted of recordings of the target words from 4 distinct voices. For the human voices, the target words were recorded by a female and a male speaker, both native English speakers, using a Shure

WH20 XLR head-mounted microphone in a sound-attenuated booth. The digital assistant voices were created using the command line on an Apple computer (OSX 10.13.6), and changing the Siri voice preference (American female, American male). All sound files were amplitude normalized (60 dB).

2.1.2. Participants and Procedure

10 participants, balanced for gender (5F, 5M), were recruited from the UC Davis Psychology subject pool to complete the shadowing task. All participants were native speakers of American English, 18-39 years old (mean = 22.4y, sd = 6.3y), and received course credit for their participation. First, subjects read an introduction, where they were told they would be repeating words produced by four interlocutors: Siri and Alex (digital devices) and Melissa and Carl (humans). This introductory slide included four pictures of the model talkers: the images for Siri and Alex were two separate iPhones displaying an "active" Siri mode (e.g., "What can I help you with today?") while the images for the human voices, consisted of two stock images of smiling adult humans of corresponding genders.

In the baseline phase, following the introduction, each of the 12 target words were shown on a computer screen one at a time (randomly selected) and participants produced each word in isolation (two times). In the shadowing phase, participants heard one of four interlocutor voices saying the word and were asked to repeat, while seeing the interlocutor and target word on the screen (e.g., "Carl says 'shone'"). The words and interlocutors were randomly selected on each trial. In total, subjects repeated the 12 words twice for each speaker (12*2*4 = 96 shadowed tokens per participant). Each word production was recorded, digitized at a 44kHz sampling rate, using Shure WH20 XLR head-mounted microphone in a sound-attenuated booth.

2.2. AXB Perceptual Similarity Paradigm

We used an AXB paradigm to assess global imitation of the shadowed productions, following [15].

2.2.1. Stimuli

The second pre- and post-exposure productions from each shadower were selected for each of the twelve words. The silence was removed before and after the word production from the recordings. All pre- and post-exposure tokens were amplitude-normalized (60 dB). In total, there were 480 tokens (10 shadowers*12 words*4 model talkers).

2.2.2. Participants and Procedure

30 native English speakers, none of whom participated in the shadowing task, were recruited from the UC Davis Psychology subject pool to complete the AXB similarity ratings task. On a given trial, subjects heard three productions of a word in a row. The 1st and 3rd items were either a pre- or post-exposure production of the word by one shadower. The 2nd item was the model talker's production of that item. Subjects indicated whether the 1st ("A") or 3rd ("B") recording was most similar to the 2nd ("X") item, using a labeled button box. Orders of pre- and post-exposure tokens occurred equally within each subject and were randomized by item. Trial order was additionally randomized. The experiment lasted roughly 40 minutes.

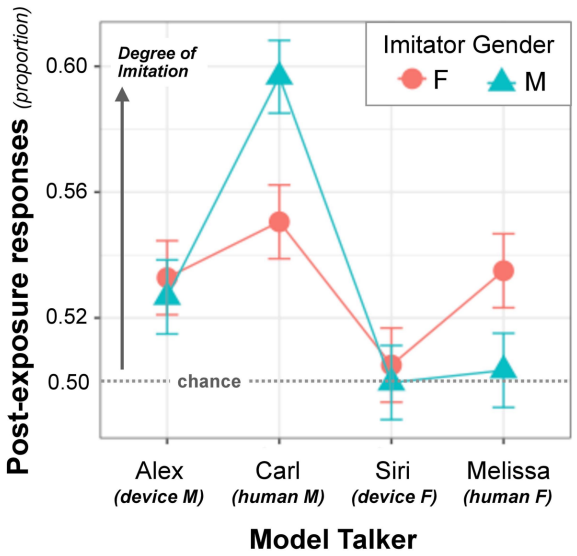
3. ANALYSIS AND RESULTS

3.1. Perceptual evaluation of vocal alignment

Responses were coded for whether the Post-exposure item was rated as more similar-sounding to the Model item with "1", or a "0" if not. We analyzed responses using a mixed effects logistic regression with the *lme4* R package [2]. Fixed effects included Model Humanity (device, human), Model Gender (F, M), and Imitator Gender (F, M), the three-way interaction between these variables, all possible two-way interactions, and by-Rater random intercepts.

The logistic regression revealed a significant main effect of Model Gender, with less imitation for female model talkers (β=-0.02, F=24.37, p<0.001). Model Humanity was also a significant main effect, with less imitation for devices than humans (β=-0.02, F=13.43, p<0.001) (see Figure 1).

Figure 1: Mean accuracy and standard errors of perceptual similarity ratings.



Additionally, we observed a significant interaction between Model Gender and Imitator Gender, where female speakers imitated female models more than male speakers (β=0.01, F=5.41, p=0.02). The three-way interaction between Model Gender, Model Humanity, and Imitator Gender was also significant (β=-0.001, F=5.67, p=0.018), where females imitated other females less for devices. The perceptual rating of males' and females' post-exposure productions to *Siri* are at-chance (50%); that is, neither group displays imitation of the device female voice. For male talkers, this was also true for the human female voice (see Figure 1, where *Melissa* is at chance).

4. GENERAL DISCUSSION

In this study, we examined human-device vocal imitation. We find that people do imitate the vocal patterns of device voices, however to a lesser extent than they do for human voices. This observation does not align with proposals that phonetic imitation is strictly representationally mediated, where lexical representations are updated proportional to exposure to a particular voice. Rather, interlocutor humanity this a factor that mediates patterns of vocal alignment. Despite the fact that modern-day digital devices are more naturalistic than ever, people still engage with devices in a distinct manner from the way they engage with other humans, as evidenced by phonetic imitation.

However, we did observe socially-mediated alignment patterns towards both types of interlocutors: speakers imitated male voices more than female voices for both human and device voices. This observation raises a theoretically interesting possibility that people are applying the same social rules from human-human interactions to their interactions with device interlocutors based on their apparent social characteristics. For example, we see more robust imitation towards male relative to female voices, which is in line with prior findings [15]. This supports a CAT [8] perspective that alignment is mediated by social properties of our interlocutors. Additionally, we see gender-mediated vocal imitation for both human *and* device voices. This is in line with the Computers are Social Actors (CASA) account [13]; our findings suggest that speakers are applying principles from human-human interaction. Still, we observe less imitation for device voices than for human voices. This means that people do not treat these device voices identically to human voices; their alignment is tempered by their artificiality.

Some scholars have suggested that phonetic imitation is one mechanism for the spread of sound change, e.g., [6, 19]. That we see imitation of device voices in the present study is a starting point in considering whether devices might influence humans' speech patterns more broadly. Future work studying imitation of device speech *over time* could serve as a new testing ground for investigating language change via imitation, ultimately to inform theories of sound change.

Nevertheless, the current study is constrained by certain limitations. For one, we only report global perceptual similarity measures. To further explore the nature of phonetic imitation of digital devices, future work can explore the contributions of individual acoustic properties. For example, we might predict that speakers align more with the prosodic features of device voices, relative to acoustic-phonetic characteristics (e.g., formant frequencies). Another limitation is in our study design: there was a confound in the apparent "humanity" of the voices and the recordings being either naturally produced by a real talker or text-to-speech (TTS) synthesized. Thus, from our observed differences in imitation as a function of humanity, it is difficult to tease apart whether people are responding to the apparent humanity or to the different digital properties of the voices. While the design of our current study reflects the confound that actually exists when we interact with real humans vs. digital devices in our everyday lives, future work can explore which aspect of the digital voices constrains imitation.

Ultimately, examining how people interact with digital devices can shed light on the mechanisms underlying speech production and inform models of linguistic communication. This work also has applications for improving speech and voice technology.

5. REFERENCES

[1] Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*(1), 177-189.

[2] Bates, D., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Retrieved from doi:10.18637/jss.v067.i01

[3] Bell, L. (2003). Linguistic Adaptations in Spoken Human-Computer Dialogues-Empirical Studies of User Behavior. *Institutionen för talöverföring och musikakustik*.

[4] Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Nass, C. (2003, July). Syntactic alignment between computers and people: The role of belief about mental states. In *Proceedings of the Twenty-fifth Annual Conference of the Cognitive Science Society* (pp. 186-191).

[5] Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new

and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990.

[6] Garrett, A., & Johnson, K. (2013). Phonetic bias in sound change. *Origins of sound change: Approaches to phonologization*, 51-97.

[7] Garrod, S., & Pickering, M. J. (2007). Alignment in dialogue. *The Oxford handbook of psycholinguistics*, 443-451.

[8] Giles, H., Coupland, J., Coupland, N., & Oatley, K. (1991). Contexts of Accommodation: Developments in *Applied Sociolinguistics.* Cambridge University Press.

[9] 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.

[10] Goldinger, S. D. (1998). Signal detection comparisons of phonemic and phonetic priming: The flexible-bias problem. *Perception & Psychophysics*, 60(6), 952–965.

[11] Hoy, M. B. (2018). Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. *Medical Reference Services Quarterly,* 37(1), 81–88.

[12] Johnson, K. (2007). Decisions and mechanisms in exemplar-based phonology. *Experimental Approaches to Phonology. In Honor of John Ohala*, 25–40.

[13] Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103.

[14] Olmstead, K. (2017). Nearly half of Americans use digital voice assistants, mostly on their smartphones. Pew Research Center.

[15] Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382-2393.

[16] Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(2), 309.

[17] Pierrehumbert, J. (2002). Word-specific phonetics. *Laboratory Phonology*, 7.

[18] Plummer, D. C., Reynolds, M., Golvin, C. S., Young, A., Sullivan, P. J., Velosa, A., … Bhat, M. (2016). Top strategic predictions for 2017 and beyond: Surviving the storm winds of digital disruption. *Gartner report G00315910.* Gartner. Inc.

[19] Yu, A. C. L., Abrego-Collier, C., & Sonderegger, M. (2013). Phonetic Imitation from an Individual-Difference Perspective: Subjective Attitude, Personality and "Autistic" Traits. *PLOS ONE*, 8(9), e74746. https://doi.org/10.1371/journal.pone.0074746