INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Expressiveness influences human vocal alignment toward voice-AI

Michelle Cohn, Georgia Zellou

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

Abstract

This study explores whether people align to expressive speech spoken by a voice-activated artificially intelligent device (voice-AI), specifically Amazon's Alexa. Participants shadowed words produced by the Alexa voice in two acoustically distinct conditions: "regular" and "expressive", containing more exaggerated pitch contours and longer word durations. Another group of participants rated the shadowed items, in an AXB perceptual similarity task, as an assessment of overall degree of vocal alignment. Results show greater vocal alignment toward expressive speech produced by the Alexa voice and, furthermore, systematic variation based on speaker gender. Overall, these findings have applications to the field of affective computing in understanding human responses to synthesized emotional expressiveness.

Index Terms: vocal alignment, human-computer interaction, speech perception & production, affective computing

1. Introduction

Humans are increasingly engaging with voice-activated artificially intelligent (voice-AI) devices, such as Amazon's Alexa, in more naturalistic and meaningful ways, e.g., chatbots in [1]. Yet, many text-to-speech (TTS) systems still generate voices that are evaluated as sounding "robotic" or "monotonous" to human users [2]. Some efforts to improve the perceived dynamism of TTS have focused on synthesizing acoustic expressiveness, based on human expressive vocal patterns, cf. [3]–[5]. Yet, how human users perceive and respond to these displays of expressiveness is an empirical question: do users respond to vocal expressiveness in voice-AI speech differently than non-expressive productions? One way to explore reactions to expressive TTS is to examine patterns of *vocal alignment* toward expressive productions in Amazon Alexa's voice, relative to productions without them.

1.1. Vocal alignment

Humans tend to adopt the acoustic-phonetic patterns of their interlocutor; this is known as vocal alignment, entrainment, or phonetic imitation, e.g., [6]–[9]. The ways in which individuals vocally align to each other is mediated by many factors, including the speakers' genders [10]–[12] and perceived attractiveness of their interlocutor [12], [13]. Degree of vocal alignment is proposed by some to serve as a manifestation of interlocutors' social distance; greater vocal alignment is thought to convey speakers' closeness, while divergence is thought to reflect interpersonal distance, cf. Communication Accommodation Theory (CAT) [14], [15]. For example, roommates at the end of the year show greater alignment than at the beginning [9], suggesting that alignment reflects the development and maintenance of social ties.

1.2. Alignment of emotion and expressiveness

Degree of vocal alignment is also considered to be a reflection of interlocutors' feelings of empathy. This is supported by empirical work showing alignment toward displays of emotion or expressiveness, cf. 'emotional mimicry' [16]–[18]. For example, psychologists have been shown to produce speech with reduced vowel spaces, aligning with their depressed patients' speech [18]. In [16], counselors displaying greater vocal alignment toward their patients were rated as having higher degrees of empathy (by independent raters listening to the speech). These observations of alignment toward interlocutors' apparent emotional-cognitive states support theories of *embodied cognition* which propose that humans experience the emotional states of others to some degree as if it was their own, e.g., [19], [20]. For example, both directly experiencing "disgust" and seeing a face showing "disgust" engage the insula, an area of the cortex, cf. [21], suggesting a low-level, neural source for this behavior. In [22], they found some evidence of this in the speech domain: subjects repeated utterances produced in either a "happy" or "sad" voice and external raters confirmed speakers' imitation of these two emotions. Others have found evidence for fine-grained alignment across modalities, such as micro-activations of cheek muscles when subjects hear a "smiling" voice [23] or when subjects read a word varying in valence (e.g., positive, neutral, negative) [24].

1.3. Human-computer alignment

To what extent people align with vocal displays of emotional expressiveness in TTS speech is an under-studied question. The idea that humans might apply socio-linguistic conventions and patterns from human-human conversation to human-computer interactions is in line with theories of computer personification, cf. "Computers Are Social Actors" account, or CASA, in [25]. CASA proposes that humans automatically treat computers like humans when they detect a cue of "humanity" in the system. Indeed, humans have been shown to align to many linguistic features in productions by computer systems, including syntactic structure [26], lexical choice [27], speaking rate [28], and amplitude [29]. Furthermore, when a computer's voice displays speech alignment toward the human's voice, e.g., rate [30], or mean intensity, rate, and pitch [31], it is rated as more "likeable" or "trustworthy" [31], [32]. Taken together, these findings suggest that some of the pro-social mechanisms in human-human alignment may also be relevant in human-device linguistic alignment.

Recent developments in TTS synthesis have improved the acoustic parameters that influence degree of perceived expressiveness of the voices, cf. [3]–[5]. Thus, one question is

http://dx.doi.org/10.21437/Interspeech.2019-1368

whether humans will display alignment to TTS productions that are realized with increased acoustic-phonetic expressiveness, since it conveys robust human-like dynamism in the voice. Some work has demonstrated that *matching* the user's degree of expressiveness in the TTS voice yields more positive social sentiments of the interaction. For example, [17] generated a TTS voice that aligned with the user's degree of expressive speech in a turn-by-turn manner, matching pitch and speaking rate; expressive speech alignment resulted in higher ratings of rapport and mutual understanding by independent raters of the conversations.

Others have provided some evidence that *humans* align to the emotionally expressive behaviors of AI systems: e.g., people imitate robot mood, via body language imitation, in an imitation game [33]. Yet, the extent to which humans align with a device's emotional expression in less *explicit* tasks, where they are not directly told to imitate, remains an open question. We ask whether people subconsciously align to the emotional expression of a TTS voice. While the voice is a conduit for emotional expression, e.g., [22], [23], no work, to our knowledge, has directly tested *human users'* degree of vocal alignment toward a computer voice on the basis of emotional expressiveness.

1.4. Current study

In the present study, we tested whether individuals shadowing an Amazon Alexa voice would show differences in their degree of vocal alignment when the words are produced as "expressive," e.g., wider pitch range, longer segmental durations, compared to "regular" prosodic realizations. Following [34], we collected separate perceptual similarity ratings between the shadowed and model talkers' productions to assess degree of vocal alignment. This is a novel contribution to the literature, extending prior work on human-human emotional mimicry in speech to human-computer interaction.

2. Methods

2.1. Experiment 1: Shadowing task

2.1.1. Stimuli for shadowing task

Stimuli consisted of 24 words (see Table 1) generated by the American English (US) Amazon Alexa TTS system. For each word, we generated recordings of the Alexa voice in two conditions: *Regular* and *Expressive.* The regular productions consisted of Alexa's unmodified, or regular, TTS. The *Expressive* stimuli consisted of hyper-prosodic interjections produced by the Alexa voice as "Speechcons" (see [35] for audio illustrations). All voice recordings were generated in the Alexa Skills Kit, with Speechcons generated with Speech Synthesis Markup Language (SSML). Mean intensity of all tokens were normalized to 70 dB in Praat [36].

Table 1: *Stimuli word list*

awesome	cheers	eureka	jinx	super	yuck
bravo	ditto	great	roger	wow	yum
bummer	dynamite	hurray	shucks	wowzer	zap
cool	darn	howdy	splash	yikes	zing

For each stimulus item, we measured acoustic properties associated with increased expressiveness: fundamental frequency (f0: mean, sd, range) and word duration [37]. We found that *Expressive* tokens, on average, were longer in word duration, and contained greater f0 variation, lower f0 means, and wider f0 ranges (see Table 2).

Table 2: *Acoustic measurements of stimulus items in Regular and Expressive Conditions.*

Condition	*Word duration mean (sd)*	*f0 mean (sd)*	*f0 range means*
Regular	0.496 s (0.08)	215.7 Hz (17.2)	191.4 - 241.9 Hz
Expressive	0.726 s (0.24)	181.5 Hz (25.1)	168.5 - 206.4 Hz

2.1.2. Participants and procedure

Subjects consisted of 10 native English speakers, balanced by gender (5 female). All subjects reported normal hearing. All speakers reported having used digital devices in the past (e.g., Apple's Siri, Amazon's Alexa, Google Assistant, etc.). Subjects were recruited through the UC Davis Psychology undergraduate subject pool (mean age: 18.8 years, sd: 0.78). Subjects were fitted with headphones (Seinheiser Pro) and a head-mounted microphone (Shure WH20 XLR) and seated in front of a computer monitor in a sound attenuated booth. Subjects began with a pre-exposure phase, where they were asked to read the words presented on the screen aloud; these words consisted of the 24 target items, randomly presented, each in two blocks, for a total of 48 trials. Next, subjects were introduced to "Alexa", the digital device, along with a picture of a silver Amazon Echo. During the shadowing phase, Alexa first produced each word (randomly selected and balanced by condition within each block). Subjects were asked to simply repeat the word after Alexa (i.e., their "Post-Exposure" productions) and were given no explicit instructions to mimic or imitate. In total, subjects shadowed the 48 items (24 words x 2 Conditions) twice, in two separate blocks. The shadowing experiment took roughly 20 minutes.

2.2. Experiment 2: Perceptual similarity ratings

2.2.1. Stimuli for AXB similarity ratings

Stimuli consisted of the second Pre-exposure and second Post-exposure recordings (for both the *Expressive* and *Regular* conditions*)* for the 5 female and 5 male participants from the shadowing study.

2.2.2. Participants and procedure

A separate group of subjects (*n*=43) participated in the perceptual similarity ratings experiment. Subjects were native English speakers and recruited from the UC Davis Psychology subject pool (33 female; mean age: 20.1 years, sd: 2.0). Subjects completed the experiment in a sound attenuated booth, wearing headphones, and seated in front of a computer monitor and button box (E-Prime). On each trial, subjects heard three tokens presented in a row, following the format of an AXB paradigm [34]: "A" and "B" were the Pre-exposure (e.g., $\text{awesome}_{\text{PREEXP}}$) and Post-exposure recordings (e.g., $\text{awesome}_{\text{POST-EXPRESSIVE}}$) for a given word, while "X" was the Alexa recording of that same word in the same condition (i.e., *Expressive* or *Regular*).

Subjects were asked to determine whether "A" or "B" (the same speaker) sounded most like "X" (Alexa, the modeled token). Pre-exposure and Post-exposure ordering was counterbalanced across trials. In total, subjects completed 480

trials (10 shadowers x 2 conditions x 24 words). The AXB task took roughly 35-40 minutes in total.

3. Results

Raters' responses were coded as binomial data based on whether they selected the "post-exposure" token as being more similar to the model talker's production of the word (=1), relative to selection of the "pre-exposure" token (=0), and were modeled with a mixed effects logistic regression using the *lme4* R package [38]. The model included two fixed effects: Condition (Expressive or Regular) and Speaker Gender (Female, Male). The interaction of Condition and Gender was also included. Random effects included by-Rater and by-Speaker random intercepts and by-Speaker random slopes by Condition. Figure 1 presents mean similarity ratings by Condition and Speaker Gender.

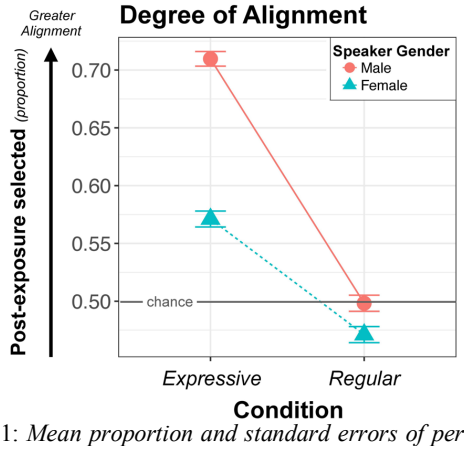

Figure 1: *Mean proportion and standard errors of perceptual similarity ratings, by Condition (Expressive, Regular) and Speaker Gender (M, F).*

Table 3 presents the output of the mixed effects logistic regression. The model revealed a main effect of Condition, where *Expressive* post-exposure productions were rated as more similar to the model talker (p<0.001), seen in Figure 1.

Additionally, as seen in Figure 1, there was a main effect of Speaker Gender, such that male speakers' post-exposure productions were rated as significantly more similar to the model talker than for females (p<0.001). Furthermore, we observed a two-way interaction between Condition and Speaker Gender, with males even more likely to align to the *Expressive* condition (p<0.01).

Table 3: *Mixed effects logistic regression output*

	Coef	*SE*	*z*	*p*
(Intercept)	0.27	0.07	3.8	<0.001 ***
Condition - *Expressive*	0.34	0.05	7.5	<0.001 ***
SpeakerGender- *Male*	0.19	0.06	3.4	<0.001 ***
Condition - *Expressive* * SpeakerGender - *Male*	0.13	0.05	2.9	<0.01 **
Subjects, n= 43		*Num. Observations* = 20,640		

3.1. Post-hoc analysis: Emotional properties of words

We were additionally interested in whether the emotional properties of the words used in this study mediate vocal alignment of expressiveness (e.g., lexical valence). Prior work has explored how emotion is expressed via multiple dimensions that affect word recognition, including valence (positive-negative) and arousal (calm-excited) [24]. For example, listeners respond faster in lexical decision tasks to positive valence words, relative to negative or neutral words [39], while another study found differences in a semantic categorization task for words with high versus low arousal [40]. One prediction for the present study is that speakers might show greater vocal alignment for words with positive emotional valence; this stems from prior work showing that *positive* social sentiments toward the interlocutor predicts greater vocal alignment [12]. Additionally, we predict that speakers may show greater vocal alignment for words that express greater "excitement", or higher arousal, based on increased attention to their production.

We tested these predictions on a subset of our words (16 items) that had ratings of emotional valence and arousal in a large-scale norming experiment [41]. In the norming study, 1,827 English speakers provided their emotional responses on several dimensions when reading the word: e.g., valence (1=unhappy, to 9=happy) and arousal (1=calm to 9=excited). We selected the ratings for each lemma, the base form of words, where available. (Subset word list used for post-hoc analysis: *awesome, great, super, cheer(s), darn, bummer, yum(my), bummer, yuck, zap, ditto, zing, splash, dynamite, jinx, cool*).

We modeled the effect of each emotional property (arousal, valence) on AXB similarity ratings in separate post-hoc logistic regression models. Main effects included either Word Valence Rating or Word Arousal Rating and Condition (Regular, Expressive), and the interaction of these two factors, with by-Rater and by-Speaker random intercepts and by-Speaker random intercepts by Condition.

The Word Valence model showed no effect of Word Valence Rating (p=0.44), but a main effect of Condition (β=0.28, z=3.0, p<0.01): expressive speech productions were rated as more similar to the model talker, a finding in line with our main analysis. No interaction between Word Valence Rating and Condition was observed (p=0.86).

Results for the word arousal model revealed a significant main effect of Word Arousal Rating (β=0.15, z=6.8, p<0.001), with higher ratings of post-exposure similarity for words with higher arousal ratings. While we observed no main effect of Condition (p=0.27), we did find an interaction between Condition and Arousal (β=0.09, z=3.9, p<0.001): words with higher arousal ratings produced by the expressive voice showed higher alignment ratings (Figure 2).

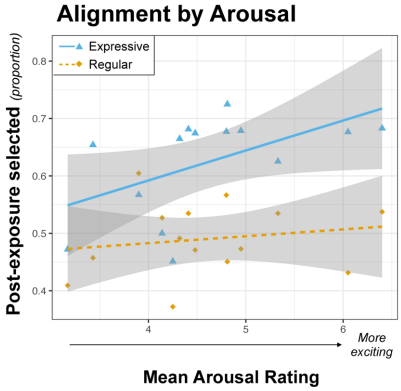

Figure 2: *Mean proportion perceptual similarity ratings as a function of word Arousal ratings (taken from [38]) by Condition (Expressive, Regular).*

4. Discussion

In this study, we tested whether human vocal alignment to Amazon's Alexa, a voice-AI system, varied based on the apparent expressiveness of the TTS voice, as realized by distinct acoustic realizations. Global similarity ratings revealed that speakers showed greater vocal alignment to expressive realizations of lexical items (Alexa "Speechcons" [35]), which contained wider f0 ranges and longer word durations, relative to the standard Alexa productions. Our results are broadly in line with theories of embodied cognition that propose that humans experience emotional expressions of others to some degree as their own (cf. [19], [20]), as reflected in their productions while shadowing the expressive Alexa voice. A novel finding of this study is that humans show subconscious alignment to the emotional expressiveness of a *non-human* voice, responding to the human-like indices of emotion (e.g., duration, f0 differences, etc.) produced by an artificially-intelligent entity.

Additionally, our findings parallel reported patterns from human-human alignment, where speakers have been shown to align to the acoustic-phonetic properties indexing their interlocutor's emotional state [16], [18], [22]. That humans appear to engage with the device voice on the basis of cognitive-emotional expression from human-human interaction is in line with theories of computer personification, e.g., CASA [25]. In other words, human-device interaction appears to be mediated by similar pro-social behaviors as in human-human interactions. To further support this, we find differences in the magnitude of alignment on the basis of *speaker gender*: male speakers showed greater alignment for expressive speech relative to the female speakers. That speakers vary in degree of alignment on the basis of their gender is consistent with prior work in human-human interaction that reported greater alignment for male speakers [9]. This finding also supports the CASA theory of human-computer interaction [25]: human behavior toward devices appears to be mediated by the same social patterns from human-human interaction, including gender asymmetries.

Still, there are alternative explanations as to why the expressive realizations of words may have been phonetically imitated to a greater extent. One possibility is that that expressive condition may have triggered increased perceptual attention. Specifically, the acoustic properties of the "expressive" productions are *hyper*-expressive, i.e., phonetically exaggerated [35]. The presence of increased, exaggerated expression may drive listeners' attention to the acoustic-phonetic features of the word [42], possibly leading to more robust vocal alignment.

In support of this interpretation, our post-hoc analysis revealed that subjects showed greater vocal alignment toward expressive items that also had higher word arousal ratings. We see this as evidence that congruence between expressiveness of the voice and expressiveness of the word's meaning can predict stronger alignment in human-computer alignment, perhaps via greater attention to the word's pronunciation. While attention has also previously been linked to valence [43], we did not see an effect of valence in our post-hoc analysis, a finding in line with prior work demonstrating the dissociability of these subdimensions of emotional expression (e.g., separate neural underpinnings for valence and arousal in [44]).

Another explanation for the increased alignment for expressive tokens is based on durational differences. As summarized in Table 1, the expressive condition items were longer in duration than the regular condition items. Longer productions contain, by definition, greater acoustic phonetic information which listeners have more time to attend to. This is another attentional/acoustic aspect of expression that could lead to greater degree of alignment.

One question raised by the findings in this study is whether humans behave similarly toward apparent expressiveness of device voices as for human voices. For example, does Alexa's expressive "darn" communicate the same degree of disappointment as a human "darn"? Future work comparing vocal alignment toward human and device productions of words displaying variations in expressiveness can additionally test theories of computer personification, e.g., CASA, i.e., whether vocal alignment to expressiveness in device voices is comparable to vocal alignment to human emotional expressiveness.

Furthermore, our findings are relevant to the field of affective computing; this study contributes to our understanding of human responses to synthesized emotional expressiveness. For one, our results demonstrate that humans automatically align with more expressive vocal productions; this may have applications in fostering a degree of empathy in human-computer interaction, in inducing 'emotional mimicry'. For example, in voice user interfaces (VUI) that lack a screen or visual avatar to provide other paralinguistic information, adding more expressive–or perhaps *hyper*-expressive– features into TTS utterances may further improve rapport and user satisfaction. Still, whether humans respond to more nuanced emotional expressiveness in a VUI remains an open question. Overall, we see shadowing paradigms as one method to more implicitly test humans' responses to synthesized emotional expressiveness, rather than relying on more explicit ratings by the user or external rater.

5. Conclusions

We observe that humans show greater vocal alignment to more expressive productions by the Amazon Alexa TTS voice. This is the first study, to our knowledge, to demonstrate that the acoustic variation signaling cognitive-emotional state in the TTS voice has a direct consequence on the speech patterns of the users. This raises important questions as to human perception of emotion by *non*-human entities and, overall the degree to which there is personification of voice-AI systems. As humans interact with voice-AI systems more and more, the role that they play in our speech communities will grow. Understanding how people behave when they interact with voice-AI, and how talking to a TTS system may impact *human language* more broadly, is relevant for models of language perception and production, as well as for theories of human-computer interaction.

6. Acknowledgements

This work was partially funded by an Amazon Faculty Research Award to GZ.

7. References

[1] C.-Y. Chen, D. Yu, W. Wen, Y. Yang, M. Zhou, K. Jesse, A. Chau, A. Bhowmick and S. Iyer, "Gunrock: Building A Human-Like Social Bot By Leveraging Large Scale Real User Data," *2nd Proc. Alexa Prize*, 2018.

[2] F. S. Baker, "Emerging realities of text-to-speech software for nonnative-English-speaking community college Students in

the freshman year," *Community Coll. J. Res. Pract.*, vol. 39, no. 5, pp. 423–441, 2015.

[3] R. J. Skerry-Ryan, E. Battenberg, Ying Xiao, Daisy Stanton Joel Shor, Ron Weiss, Rob Clark, and Rif A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *ArXiv Prepr. ArXiv180309047*, 2018.

[4] E. Székely, "Expressive speech synthesis in human interaction," University College Dublin, 2015.

[5] M. Tahon, G. Lecorvé, and D. Lolive, "Can we Generate Emotional Pronunciations for Expressive Speech Synthesis?," *IEEE Trans. Affect. Comput.*, 2018.

[6] S. D. Goldinger, "Signal detection comparisons of phonemic and phonetic priming: The flexible-bias problem," *Percept. Psychophys.*, vol. 60, no. 6, pp. 952–965, 1998.

[7] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[8] N. Lubold and H. Pon-Barry, "Acoustic-prosodic entrainment and rapport in collaborative learning dialogues," in *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, 2014, pp. 5–12.

[9] J. S. Pardo, "On phonetic convergence during conversational interaction," *J. Acoust. Soc. Am.*, vol. 119, no. 4, pp. 2382–2393, 2006.

[10] R. Levitan, A. Gravano, L. Willson, S. Benus, J. Hirschberg, and A. Nenkova, "Acoustic-prosodic entrainment and social behavior," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, 2012, pp. 11–19.

[11] U. D. Reichel, Š. Beňuš, and K. Mády, "Entrainment profiles: Comparison by gender, role, and feature set," *Speech Commun.*, vol. 100, pp. 46–57, 2018.

[12] M. Babel, "Evidence for phonetic and social selectivity in spontaneous phonetic imitation," *J. Phon.*, vol. 40, no. 1, pp. 177–189, 2012.

[13] J. Michalsky and H. Schoormann, "Pitch Convergence as an Effect of Perceived Attractiveness and Likability.," in *INTERSPEECH*, 2017, pp. 2253–2256.

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

[15] T. L. Chartrand and J. A. Bargh, "The chameleon effect: the perception–behavior link and social interaction.," *J. Pers. Soc. Psychol.*, vol. 76, no. 6, p. 893, 1999.

[16] B. Xiao, P. G. Georgiou, Z. E. Imel, D. C. Atkins, and S. Narayanan, "Modeling therapist empathy and vocal entrainment in drug addiction counseling.," in *INTERSPEECH*, 2013, pp. 2861–2865.

[17] J. C. Acosta and N. G. Ward, "Achieving rapport with turn-by-turn, user-responsive emotional coloring," *Speech Commun.*, vol. 53, no. 9–10, pp. 1137–1148, 2011.

[18] B. Vaughan, C. De Pasquale, L. Wilson, C. Cullen, and B. Lawlor, "Investigating Prosodic Accommodation in Clinical Interviews with Depressed Patients," in *Int. Symp. Pervasive Computing Paradigms for Mental Health*, 2018, pp. 150–159.

[19] P. Winkielman, P. M. Niedenthal, and L. M. Oberman, "Embodied perspective on emotion-cognition interactions," in *Mirror Neuron Systems*, Springer, 2008, pp. 235–257.

[20] P. M. Niedenthal, P. Winkielman, L. Mondillon, and N. Vermeulen, "Embodiment of emotion concepts.," *J. Pers. Soc. Psychol.*, vol. 96, no. 6, p. 1120, 2009.

[21] M. Iacoboni, "Neural mechanisms of imitation," *Curr. Opin. Neurobiol.*, vol. 15, no. 6, pp. 632–637, 2005.

[22] R. Neumann and F. Strack, "' Mood contagion': the automatic transfer of mood between persons.," *J. Pers. Soc. Psychol.*, vol. 79, no. 2, p. 211, 2000.

[23] P. Arias, P. Belin, and J.-J. Aucouturier, "Auditory smiles trigger unconscious facial imitation," *Curr. Biol.*, vol. 28, no. 14, pp. R782–R783, 2018.

[24] J. Künecke, W. Sommer, A. Schacht, and M. Palazova, "Embodied simulation of emotional valence: Facial muscle responses to abstract and concrete words," *Psychophysiology*, vol. 52, no. 12, pp. 1590–1598, 2015.

[25] C. Nass and Y. Moon, "Machines and mindlessness: Social responses to computers," *J. Soc. Issues*, vol. 56, no. 1, pp. 81–103, 2000.

[26] 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.

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

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

[29] N. Suzuki and Y. Katagiri, "Prosodic alignment in human–computer interaction," *Connect. Sci.*, vol. 19, no. 2, pp. 131–141, 2007.

[30] N. Ward and S. Nakagawa, "Automatic user-adaptive speaking rate selection," *Int. J. Speech Technol.*, vol. 7, no. 4, pp. 259–268, 2004.

[31] R. Levitan, "Acoustic-prosodic entrainment in human-human and human-computer dialogue," Doctoral Dissertation, Columbia University, 2014.

[32] Š. Beňuš, M. Trnka, E. Kuric, L. Marták, A. Gravano, J. Hirschberg, R. Levitan, "Prosodic entrainment and trust in human–computer interaction," in *Proc. 9th Intl. Conf. on Speech Prosody*, 2018, pp. 220–224.

[33] J. Xu, J. Broekens, K. Hindriks, and M. A. Neerincx, "Robot mood is contagious: effects of robot body language in the imitation game," in *Proc. 2014 Intl. Conf. on Autonomous Agents & Multi-Agent Systems*, 2014, pp. 973–980.

[34] J. S. Pardo, I. C. Jay, R. Hoshino, S. M. Hasbun, C. Sowemimo-Coker, and R. M. Krauss, "Influence of role-switching on phonetic convergence in conversation," *Discourse Process.*, vol. 50, no. 4, pp. 276–300, 2013.

[35] Amazon, "Speechcon Reference (Interjections): English (US) | Custom Skills," 2018. [Online]. Available: https://developer.amazon.com/docs/custom-skills/speechcon-reference-interjections-english-us.html. [Accessed: 09-Dec-2018].

[36] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*. 2018.

[37] I. R. Murray and J. L. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," *J. Acoust. Soc. Am.*, vol. 93, no. 2, pp. 1097–1108, 1993.

[38] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, Oct. 2015.

[39] Z. Estes and J. S. Adelman, "Automatic vigilance for negative words is categorical and general.," *Emotion,* vol. 8, no. 4, pp. 453-457, 2008.

[40] N. Delaney-Busch, G. Wilkie, and G. Kuperberg, "Vivid: How valence and arousal influence word processing under different task demands," *Cogn. Affect. Behav. Neurosci.*, vol. 16, no. 3, pp. 415–432, Jun. 2016.

[41] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," *Behav. Res. Methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[42] Y. Zhang, T. Koerner, S. Miller, Z. Grice-Patil, A. Svec, D. Akbari, L. Tusler, and E. Carney, "Neural coding of formant-exaggerated speech in the infant brain," *Dev. Sci.*, vol. 14, no. 3, pp. 566–581, 2011.

[43] E. Fox, R. Russo, R. Bowles, and K. Dutton, "Do threatening stimuli draw or hold visual attention in subclinical anxiety?," *J. Exp. Psychol. Gen.*, vol. 130, no. 4, p. 681, 2001.

[44] J. E. Warren, D.A. Sauter, F. Eisner, J. Wiland, A.M. Dresner, R.JS. Wise, S. Rosen, and S. Scott, "Positive emotions preferentially engage an auditory–motor 'mirror' system," *J. Neurosci.*, vol. 26, no. 50, pp. 13067–13075, 2006.