DO HUMANS CONVERGE PHONETICALLY WHEN TALKING TO A ROBOT?

Tom Offrede[1], Chinmaya Mishra[2], Gabriel Skantze[2,3], Susanne Fuchs[4], Christine Mooshammer[1]

[1]Humboldt-Universität zu Berlin, [2]Furhat Robotics,
[3]KTH Royal Institute of Technology, [4]Leibniz-Centre General Linguistics (ZAS)
offredet@hu-berlin.de, chinmaya@furhatrobotics.com, skantze@kth.se, fuchs@leibniz-zas.de, mooshamc@hu-berlin.de

ABSTRACT

Phonetic convergence—i.e., adapting one's speech towards that of an interlocutor—has been shown to occur in human-human conversations as well as human-machine interactions. Here, we investigate the hypothesis that human-to-robot convergence is influenced by the human's perception of the robot and by the conversation's topic. We conducted a within-subjects experiment in which 33 participants interacted with two robots differing in their eye gaze behavior—one looked constantly at the participant; the other produced gaze aversions, similarly to a human's behavior. Additionally, the robot asked questions with increasing intimacy levels.

We observed that the speakers tended to converge on F0 to the robots. However, this convergence to the robots was not modulated by how the speakers perceived them or by the topic's intimacy. Interestingly, speakers produced lower F0 means when talking about more intimate topics. We discuss these findings in terms of current theories of conversational convergence.

Keywords: phonetic convergence; fundamental frequency; human-robot interaction; intimacy

1. INTRODUCTION

Humans having a conversation are known to change their speech depending on how their partner is speaking. Among other terms, this phenomenon is called convergence [1]. On a phonetic level, for example, it has been shown that speakers alter their speech rate, intensity, and shimmer [2] in the direction of their interlocutor's speech acoustics. The current study is concerned with convergence on the fundamental frequency (F0) level, as an initial feature to be investigated. Phonetic convergence has been observed not only in interactions between two humans, but also between a human and a voice AI system [3, 4], and, preliminarily, between a human and a robot [5]. Importantly, phonetic convergence tends to be a variable and often subtle effect [1].

Various theories propose different explanations for this phenomenon. For instance, the Interactive Alignment Model argues that a priming mechanism causes speakers to entrain to each other [6]. On the other hand, [7]'s Communication Accommodation Theory (CAT) argues that convergence (or accommodation) depends on interindividual and intergroup factors, and that it can be used to manage the social distance between two interlocutors. For example, the speakers' evaluations of each other's personal history and group-belonging dynamics may affect their (non-)accommodating behavior.

In addition to social factors, situational elements may also influence speech overall and phonetic convergence. [8, 9] have suggested that the topic being discussed may influence convergence. One example could be the emotional content of the conversation: more intimate themes, especially in interaction with the interlocutors' perceptions of each other, might influence their speech behavior. Further, [10] have demonstrated that speakers tended to converge to their interlocutor on a few acoustic features when they felt engaged in the conversational task, but not when the task was not engaging. Interestingly, these speakers still adapted their speech according to the task, independently of convergence to their partner.

One important communicative cue during conversation is eye gaze—specifically, gaze aversion (GA), i.e., the breaking of eye contact between interlocutors. Different functions have been attributed to GA, such as management of cognitive load [11], regulation of turn-taking [12], and modulation of intimacy expressions [13]. Since this behavior is ubiquitous in human-human interaction, we hypothesize that gaze behavior may also have an impact on human-to-robot convergence. Specifically, we expect that a robot that produces GA more or less similarly to a human, in comparison to a robot that looks continuously at its human interlocutor, would be perceived more positively. This, in turn, might affect the human's phonetic (F0) convergence behavior.

Hence, in this study, we investigated (1) whether humans converged on F0 to a robotic interlocutor, and whether (2) social factors—the participants' perception of the robot—and (3) the conversation topic—intimacy of the questions asked by the robots—would modulate such convergence.

2. METHOD

Each participant interacted with two Furhat robots separately under two conditions in a within-subjects design. One robot looked constantly at the participant (*Fixed Gaze* condition). The other one averted its gaze away from its interlocutor every few seconds [14], similarly to what a human would do (*Gaze Aversion* condition). This GA behavior respected known gaze cues related to conversational floor management: for example, at the end of the robot's turn, it always made eye contact with the participant, as humans usually do. We recorded the participants' speech, eye movement behavior (data discussed elsewhere) and ratings of their perception of the robots and conversations.

2.1. Participants

Thirty-three participants assigned male at birth participated in the study. Their ages ranged from 21 to 56 (M = 30.55; SD = 8.07). Five participants spoke English as a first language (L1)—the language in which the study took place—while the others' L1s were one or two of 16 different languages. According to their LexTALE scores and Lemhöfer & Broersma's classification [15], 16 of these participants would be categorized at the C1 to C2 level, 15 at B2, and two at B1 [16]. We obtained similar results on F0 behavior regardless of whether the dataset included the participants with lower English proficiency.

All participants provided written consent and received a 100 SEK-voucher as compensation.

2.2. Procedure and Materials

Before each conversation with the robot, the eye-tracker was calibrated. Then, the first interaction began. The robot started out introducing itself and explaining that they would have a conversation together. It then proceeded to ask the participant six questions, which the participant answered freely. After the participant's answer, the robot would also answer its own question before moving on to the next one. Each participant interacted with two robots (each differing in eye gaze behavior; counterbalanced in order), and after each interaction, the participant responded to a questionnaire that measured their perception of the robot and the conversation. Between the two conversations, they also filled in [17]'s version of the Big Five personality inventory and the LexTALE, a test that measures lexical knowledge in English and that correlates with general proficiency [15].

The questionnaire the participants filled in after each interaction contained items about their perception of the robot and of the conversation itself, which they rated on a 1–9 Likert scale. Since we observed moderate to strong correlation in the answers to multiple questions, we conducted a Principal Components Analysis using the R package *parameters* [18] to reduce dimensionality and identify patterns in the data. Two principal components were determined. *Conversational Quality* included items such as "My conversation with the robot flowed well" and "I was able to understand when the robot wanted me to speak." The component *Evaluation of Robot* included items such as "The robot responded to me at the appropriate time" and "The robot's behavior was very human-like."

The questions asked by the robots during the interaction always went from less (e.g., *What did you have for breakfast this morning?*) to more intimate (e.g., *For what in your life do you feel most grateful?*). To determine their intimacy level, 28 questions were taken from [19, 20] and rated on a 1–9 Likert scale by residents of the city where the data were collected (N = 130). The researchers then selected 12 questions with increasing intimacy ratings and divided them into two sets with similar intimacy distributions. Each robot asked one set of questions in a counterbalanced fashion across conditions.

2.3. Feature Extraction and Analysis

Using Praat [21], we automatically extracted voiced and voiceless portions during speech using autocorrelation. In a next step, we defined pauses as unvoiced portions with at least 350 ms duration. This threshold was determined through visual inspection. Unvoiced portions of 100, 150, 200, and 300 ms had also been considered; however, these periods often corresponded to unvoiced segments or consonant clusters within the speech stream. The speech streams between pauses are called Interpausal Units (IPU). From these, we extracted F0 information of the participant's and robot's audio streams, obtaining the mean F0 in each IPU.

For all statistical analyses, we used the R package *lmerTest* [22] to fit linear mixed-effects models.

All the models contained random intercepts for participants. All the continuous predictors were centered around their mean value in each given conversation. We consider p values below 0.05 as indicating a statistically significant effect. F0 data are given in Hertz, intimacy data are the mean ratings of the questions (1 to 9), and intensity data (see below) are given in decibels.

3. RESULTS

3.1. Do humans converge on F0 to a robotic interlocutor?

To investigate whether the participants displayed overall convergence to the robot, we fit a mixed-effects model with F0 mean as the dependent variable and the robot's previous F0 mean (i.e., the robot's F0 values in the IPUs of the preceding turn) as a predictor (*human's F0 ~ robot's previous F0*). We considered that convergence happened if the robot's F0 was a significant positive predictor of the human's F0. This method of convergence measure was based on [23]. The model revealed that the robot's F0 mean in the preceding turn was a significant predictor of the human's current F0 mean (see Figure 1 and the model's coefficients on Table 1). This suggests that the participants' F0 tended to converge to their robotic interlocutors. Importantly, there was great interindividual variability, meaning that not all speakers converged to the robot. This drove the overall convergence effect to be small: marginal $R^2 = 0.002$ and conditional $R^2 = 0.58$ [24], as calculated with the R package *MuMIn* [25].

Variables	Estimate	St. Err.	*t* value	*p* value
F0 mean				
Intercept	115.86	3.05	38.02	< 0.001
Robot's F0	0.06	0.02	3.09	0.002

Table 1: Regression coefficients indicating the effect of robot's F0 mean on human's F0 mean.

3.2. Does the participants' perception of the robot modulate convergence?

Next, we investigated whether the participants converged to the robots differently depending on the gaze condition (Fixed Gaze and Gaze Aversion). First, we evaluated whether the participants had different perceptions of the robots and conversations according to the condition. In linear regression models, we included the participants' ratings of the *Conversational Quality* and *Evaluation of Robot* components as the dependent variables and gaze condition as a predictor. Contrary to our predictions, neither of the models revealed a significant difference between conditions in the participants' ratings (Conversational Quality: $p = 0.91$; Evaluation of Robot: $p = 0.19$).

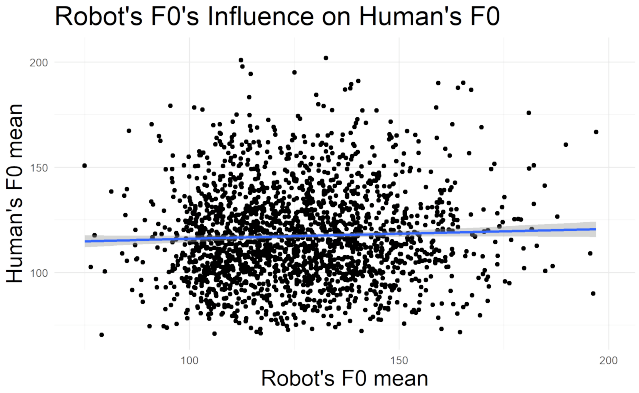

Figure 1: The influence of the robot's F0 mean on the human's F0 mean.

Despite this null result, we fit a regression model to test whether gaze condition influenced F0 convergence, since differences in perception or attitude might not be captured by our previous analysis. We added gaze condition as an interacting predictor to the aforementioned model measuring convergence (*human's F0 mean ~ robot's F0 mean : gaze condition*). This model suggested a difference between the conditions. However, a comparison between the AICs of the models with and without condition as a predictor indicated that the one without it was preferred. We thus conclude that there was no reliable difference in convergence between the conditions. Neither the participants' perception of the robot nor its gaze behavior affected their convergence on F0.

3.3. Does the conversation topic (intimacy of the questions asked by the robots) modulate convergence?

To evaluate whether intimacy of the topic influenced the convergence effect, we also included intimacy value in the model described above (*human's F0 ~ robot's F0 : intimacy*). This model did not output any significant values that might indicate convergence ($p = 0.77$). Thus, the intimacy of the topics being talked about did not influence how much people converged to the robots.

Interestingly, however, increases in the topic's intimacy level were related with a decrease in the F0 mean of the participants' speech. The robot's

gaze condition did not interact significantly with the effect of intimacy on F0.

It could be hypothesized that these decreases in F0 were related to lower intensity: participants might speak less loudly when discussing more personal topics, which would then lead to lower F0 values. Hence, for an additional post hoc analysis, using Praat's default settings we obtained the mean intensity of the IPUs for which we had calculated F0. After a mixed-effects model confirmed that intimacy also negatively predicted intensity, we ran a mediation analysis using the R package *mediation* [26]. F0 mean was the dependent variable, intimacy the independent variable, and speech intensity the mediator. Intimacy's direct effect on F0 when accounting for the mediator was -0.77 ($p < 0.001$, with F0's intercept being 115.14 and the 95% confidence interval of the effect between -1.07 and -0.49). This analysis corroborated intimacy's direct effect on F0.

4. DISCUSSION

Our data suggest that human speakers tend to converge on F0 mean to a robotic interlocutor—albeit to a small degree, as has been shown elsewhere [1]. To the extent of our knowledge, this had been shown to happen to spoken dialogue systems [3, 4], but not systematically to a physical robot and in a conversational setting—i.e., where the participant could produce unscripted speech. Although we have only analyzed F0, we will investigate convergence on other acoustic features in future work.

We attempted to modulate the humans' perception of their robotic interlocutor by having them interact with robots that differed in their eye gaze behavior. However, their ratings of the robots and conversations were not significantly different across conditions. In addition, gaze condition did not reliably influence the participants' convergence to the robot. That is, in our data, factors related to interpersonal dynamics did not seem to modulate convergence, as theories such as CAT [7] would predict. Rather, it could be that the mechanisms behind our participants' convergence behavior are of (nonsocial) cognitive nature. For instance, [6] would argue that this happens through priming: speakers produce speech that matches the acoustics they have been exposed to in the conversation. This entrainment process would then facilitate alignment on mental representations and, in turn, the conversational flow and information sharing in general.

The level of intimacy of the topics discussed also did not interact with the robot's F0 to influence the humans' speech. Interestingly, however, intimacy had a direct impact on the participants' F0 mean. This finding is in line with [10]'s observation that the circumstances of the interaction (in their case, the speakers' engagement with the conversational task) modified individuals' speech.

The impact of intimacy on F0 cannot be fully accounted for in terms of the valence of the emotions evoked. Some of the more intimate questions evoked positive emotions (*For what in your life do you feel most grateful?*) while others evoked negative memories (*What is one of the more embarrassing moments in your life?*).

One possible explanation for the effect of intimacy on F0 is [9]'s Audience Design model. Speakers produce different speech depending on their addressees, accounting for relevant social factors such as level of formality. According to [9], speakers also associate certain topics to specific addressees, causing them to vary their speech when talking about said topics even in the absence of those addressees. This might explain why our participants produced lower F0 when talking about more intimate topics.

It is important to acknowledge that the questions' intimacy is confounded with their order (every conversation went from less to more intimate), which could be the reason for the F0 effect. We argue that this is not related to such a habituation effect to the robots because the F0 reduction was similar for both conversations the participants had. When they started the second interaction, they did not have to get used to the robot as they did to the first. Still, this effect was herein analyzed in an exploratory manner and should be further investigated experimentally.

5. CONCLUSION

This paper has demonstrated that human speakers may converge on F0 to robotic interlocutors, as they do to other humans. Their perception of these interlocutors, however, does not modulate the convergence behavior, as has been shown to happen in human-human interactions. We have also observed that the intimacy level of the topics being discussed had a direct influence on the speakers' F0.

6. ACKNOWLEDGMENTS

We are thankful to Melina Pfundstein for her work on data annotation. COBRA is a European project funded by the European Union's Horizon

2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement n° 859588. This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–SFB 1412, 416591334.

7. REFERENCES

[1] J. S. Pardo, A. Urmanche, S. Wilman, and J. Wiener, "Phonetic convergence across multiple measures and model talkers," *Attention, Perception, & Psychophysics*, vol. 79, pp. 637–659, 2017.

[2] R. Levitan, A. Gravano, L. Willson, Š. Beňuš, 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.

[3] M. Cohn, K. Predeck, M. Sarian, and G. Zellou, "Prosodic alignment toward emotionally expressive speech: Comparing human and alexa model talkers," *Speech Communication*, vol. 135, pp. 66–75, 2021.

[4] I. Gessinger, B. Möbius, S. Le Maguer, E. Raveh, and I. Steiner, "Phonetic accommodation in interaction with a virtual language learning tutor: A wizard-of-oz study," *Journal of Phonetics*, vol. 86, p. 101029, 2021.

[5] O. Ibrahim, G. Skantze, S. Stoll, and V. Dellwo, "Fundamental frequency accommodation in multi-party human-robot game interactions: The effect of winning or losing." Interspeech, 2019.

[6] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," *Behavioral and brain sciences*, vol. 27, no. 2, pp. 169–190, 2004.

[7] C. Gallois, T. Ogay, and H. Giles, "Communication accommodation theory: A look back and a look ahead," in *Theorizing about intercultural communication*. Thousand Oaks: Sage, 2005, pp. 121–148.

[8] U. C. Priva and C. Sanker, "Distinct behaviors in convergence across measures." in *CogSci*, 2018.

[9] A. Bell, "Language style as audience design," *Language in society*, vol. 13, no. 2, pp. 145–204, 1984.

[10] T. Biro, J. C. Toscano, and N. Viswanathan, "The influence of task engagement on phonetic convergence," *Speech Communication*, 2022.

[11] G. Doherty-Sneddon and F. G. Phelps, "Gaze aversion: A response to cognitive or social difficulty?" *Memory & cognition*, vol. 33, no. 4, pp. 727–733, 2005.

[12] G. Brône, B. Oben, A. Jehoul, J. Vranjes, and K. Feyaerts, "Eye gaze and viewpoint in multimodal interaction management," *Cognitive Linguistics*, vol. 28, no. 3, pp. 449–483, 2017.

[13] A. Abele, "Functions of gaze in social interaction: Communication and monitoring," *Journal of Nonverbal Behavior*, vol. 10, no. 2, pp. 83–101, 1986.

[14] C. Mishra and G. Skantze, "Knowing where to look: A planning-based architecture to automate the gaze behavior of social robots," in *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2022, pp. 1201–1208.

[15] K. Lemhöfer and M. Broersma, "Introducing lextale: A quick and valid lexical test for advanced learners of english," *Behavior research methods*, vol. 44, no. 2, pp. 325–343, 2012.

[16] C. of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division, *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press, 2001.

[17] E. R. Thompson, "Development and validation of an international English big-five mini-markers," *Personality and individual differences*, vol. 45, no. 6, pp. 542–548, 2008.

[18] D. Lüdecke, M. S. Ben-Shachar, I. Patil, and D. Makowski, "Extracting, computing and exploring the parameters of statistical models using R." *Journal of Open Source Software*, vol. 5, no. 53, p. 2445, 2020.

[19] M. Kardas, A. Kumar, and N. Epley, "Overly shallow?: Miscalibrated expectations create a barrier to deeper conversation." *Journal of Personality and Social Psychology*, 2021.

[20] E. Hart, E. M. VanEpps, and M. E. Schweitzer, "The (better than expected) consequences of asking sensitive questions," *Organizational Behavior and Human Decision Processes*, vol. 162, pp. 136–154, 2021.

[21] Boersma, Paul and Weenink, David, "Praat: doing phonetics by computer [computer program]," http://www.praat.org/, 2022, version 6.2, retrieved August 23, 2022.

[22] A. Kuznetsova, P. B. Brockhoff, and R. H. B. Christensen, "lmerTest package: Tests in linear mixed effects models," *Journal of Statistical Software*, vol. 82, no. 13, pp. 1–26, 2017.

[23] A. Schweitzer and N. Lewandowski, "Convergence of articulation rate in spontaneous speech." in *INTERSPEECH*, 2013, pp. 525–529.

[24] S. Nakagawa, P. C. Johnson, and H. Schielzeth, "The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded," *Journal of the Royal Society Interface*, vol. 14, no. 134, p. 20170213, 2017.

[25] K. Bartoń, *MuMIn: Multi-Model Inference*, 2023, r package version 1.47.5. [Online]. Available: https://CRAN.R-project.org/package=MuMIn

[26] D. Tingley, T. Yamamoto, K. Hirose, L. Keele, and K. Imai, "Mediation: R package for causal mediation analysis," 2014.