INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Speech Enhancement Patterns in Human-Robot Interaction: A Cross-Linguistic Perspective

Jacek Kudera, Katharina Zahner-Ritter, Jakob Engel, Nathalie Elsässer, Philipp Hutmacher, Carolin Worstbrock

Department of Phonetics, Trier University, Germany
kudera@uni-trier.de

Abstract

This paper presents the results of the human-robot interaction (HRI) study with German native speakers addressing the robot in their L1 and in L2 English. The aim of the experiment is to test the strategies of providing clarifications when talking to the voice assistant in a task involving teaching complex vocabulary. The analyses is based on spectral (F1, F2, and mean F0) and temporal (vowel length) features excerpted from the target words. With reference to a theoretical framework of hyperarticulation and hypoarticulation, these acoustic measures were compared across the iterations of the target words (first vs. second iteration). Results showed that participants, when asked for clarification by an inanimate interlocutor, do not hyperarticulate, but try to preserve the surface representation of target words across the iterations. These findings suggest that acoustic characteristics of clarifications directed to voice assistants differ from the ones directed to human interlocutors.
Index Terms: Human-robot interaction, multilingual communication, clarifications, English, German

1. Introduction

Recent advances in ASR (Automatic Speech Recognition) made robot-directed speech a common mode of everyday interaction. Contemporary dialogues that involve voice assistants trained on large speech data permit the use of complex vocabulary and do not limit the interaction to issuing simple requests. Furthermore, several commercial voice assistants already allow for multilingual interaction and exhibit high word recognition scores for several languages [1, 2]. Increasing communicative skills of voice assistants can make overhearers think that robots possess human-like cognitive capabilities [3, 4, 5]. Regardless of attributing high linguistic competence to inanimate talking agents, several studies showed that robot-directed speech acoustically differs from utterances exchanged between human interlocutors hence serves as a good field for testing the enhancement and convergence strategies in HRI [6, 7, 8, 9].

In this study, we investigate speech enhancement patterns applied by German native speakers confronted with clarification requests given by a talking agent in two languages. Even though the term *speech enhancement* is often associated with techniques of noise reduction and speech signal amplification [10], throughout this paper, we use it in the specific sense of referring to the strategies of increasing the speech intelligibility in robot-directed speech. The underlying assumption, this experiment is based upon, relates to a natural tendency of adjusting speech style driven by presupposed communicative competence of an interlocutor. In line with Lindbloom's model of hypoarticulation and hyperarticulation [11], we assume that speakers accommodate selected speech features and hyperarticulate when the voice assistant poses a clarification request [12, 13, 14]. Such adaptation strategy can shift the acoustic parameters of speech in the following relation: the lower the presupposed or actual linguistic competence (or hearing thresholds) of an addressee, the greater the speech enhancement [15, 16]. Previous studies showed that such enhancement strategies entail increased intensity [17], greater F0 range [6, 18], increased formant frequencies [19], increased vowel durations [20], and a larger vowel space [21]. These alternations are often referred to as *clear speech* and constitute the features of hyperarticulation. By referring to previous studies into speech characteristics in HRI, in this paper, we focus on local intelligibility adjustments given by human speakers when encountered the clarification requests and address the following questions. Are clarification requests causing hyperarticulation in HRI? And if so, are the speech enhancement strategies consistent across spoken languages, i.e., participants' native language and their second language? By answering the questions above we wish to better understand the acoustic features of clarifications.

We hypothesize that the acoustics characteristics of speech (measured in F1 and F2 values extracted from the stressed vowels, duration of the stressed vowels and mean F0 of the target words) differ across iterations of target words (first production compared with the one given after a clarification request), and across spoken languages (participants' L1: German vs. L2: English). To test this hypothesis, we use a within-subject experimental design, conceptualized as a task of teaching a voice assistant complex vocabulary in two languages. With the application of a new lexis teaching paradigm, we expect to control for the effects of routinized interactions [22] in addressing the talking agents. The elicitation of the effect of low speech comprehension threshold [23] was achieved by projecting a clarification request (with a clear F0 rise) after the first instances of selected target words.

2. Method

The experiment included scripted interaction with the talking agent. Participants were instructed to teach the voice assistant some complex vocabulary in two languages. Their pseudo-task was to paraphrase the target word definitions and explain the new lexemes in their own words.

2.1. Experimental procedure

Participants were given the task to teach the robot cognate words in their L1 (German) and L2 (English). They were first presented with a target word accompanied by its concise definition. At this stage, the participants were given an option to listen to a model production of a target word to avoid potential difficulties with recalling its pronunciation in both languages.

10.21437/Interspeech.2023-828

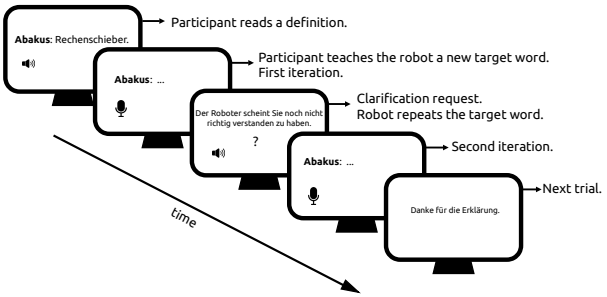

Figure 1: *Schematic interaction between the robot and the participants in a German trial.*

The model pronunciation was followed by a short (0.5 second) pure tone (440 Hz) distraction sound to prevent the participants from mimicking the model pronunciation. Then, the participants were instructed to pronounce the target word clearly to let the robot memorize a new lexical item. On the following screen, participants were given the pseudo-task to explain the meaning of the target word in their own words. Then, they were asked to clarify the target and repeat it to ensure that the robot recorded a clear phonetic representation of the newly learned target word.

The elicitation of the clarifications was achieved by projecting the modified target words in a robotic speech mode with a rising intonation contour. One-third of the words were understood correctly by the robot, i.e., the clarification request was omitted for these items. Such recordings were treated as fillers and were excluded from the analyses (11 of 34 target words). The projection of the stimuli was randomized by the language of instruction, that is, subjects were switching between languages instead of teaching the robot the target words in language blocks. The participants were asked to use a headset to ensure the best possible audio quality. Before the experimental session, participants were able to test the microphone input by recording a short sample and to adjust the recording settings prior to uploading, if necessary. The experiment must have been completed in one uninterrupted session, but participants could revoke their participation any time due to recording-related stress or fatigue. The debriefing stage followed the session to explain the real aims of the experiment. Recordings were saved and uploaded in an uncompressed format with 48 kHz sampling rate and 16-bit depth. The experiment was deployed in LabVanced online platform [24]. The entire session lasted around 30 minutes, depending on how elaborate the explanations of the target words were. The procedure has been approved by the Ethics Committee of the University of Trier (reference no. EK 77/2022).

2.2. Participants

We recorded 50 German native speakers (25 males and 25 females) aged 20 to 72 (mean = 32 years). The subjects were recruited via the outsourcing platform Prolific and paid for their participation. The L2 proficiency of participants (B2-C1) was self-estimated on the basis of a questionnaire which included questions about level of education, foreign language skills (in CEFR: A1-C2 scale), years spent abroad in an English-speaking country, multilingualism, language of everyday communication, as well as diagnosed speech and hearing disorders potentially disqualifying from the participation in the experiment.

2.3. Target words and clarification requests

The target words were German-English cognates to control for the phonological surrounding of the stressed vowels. To account for the influence of the neighbouring sounds, vowels were flanked by equal consonants in the target words in both languages. The target words were recorded in acoustically controlled surroundings (sound attenuated booth) and modified to mimic the robotic sound. The effect of robotic speech was achieved by spectral and temporal manipulation of the recorded samples. The procedure involved multiplication of the recorded signal into four mono tracks. Then the tracks were phased out the by 0.015 s. In addition, two tracks were shifted by +3 and -3 semitones to achieve a robotic sound. The tempo was lowered by 1/10 in comparison to the source samples.

Table 1: *English-German cognate target words. Asterisk marks the words with no clarification request.*

English	German	English	German
Abacus	Abakus	*Guilty	*Gültig
*Abductor	*Abduktor	Hand	Hand
*August	*August	Illusion	Illusion
*Autograph	*Autogramm	Jungle	Dschungel
*Bald	*Bald	*Mark	*Marke
Ball	Ball	Mental	Mental
Calamine	Kalamin	Metal	Metall
Caution	Kaution	Normal	Normal
*Conception	*Konzeption	Parallel	Parallel
Dependence	Dependenz	*Personal	*Personal
Depression	Depression	Probe	Probe
Dessert	Dessert	Prosody	Prosodie
Direction	Direktion	Sand	Sand
Elegant	Elegant	*See	*See
*Engaged	*Engagiert	Summer	Sommer
*Familiar	*Familiär	Tolerant	Tolerant
Finger	Finger	Zebra	Zebra

2.4. Acoustic analyses

Recordings were annotated and checked by trained phoneticians. Segment boundaries (stressed vowel and target word) were manually set following standard segmentation criteria [25]. The vowel duration, F1, F2 frequencies and mean F0 values on the target word level were automatically excerpted in Praat [26]. The format frequencies were measured at temporal midpoints of the stressed vowels.

In total, recording session of 9 participants (18% of all data) were discarded from the analyzed data due to technical difficulties, unintelligible speech or high level of background noise captured on the samples. The proportion of excluded data is relatively high also because some participants have changed their strategies of clarification and instead of another production of the target word, they provided the robot with a target synonym or hyperonym. Furthermore, several recordings were excluded from the analyzed set because participants instead of providing the second iteration of the target word simplified its definition.

2.5. Statistical analyses

The statistical analyses was done in R (version 4.1.1) [27] using the *lme4* package (version 1.1.27.1) [28]. To estimate the effects of the explanatory variables (language and iteration) on

the specified acoustic measures, four linear mixed regression models (one for each dependent variable: F1, F2, mean F0, and vowel duration) were fitted in the form of a factorial 2 x 2 design using Restricted Maximum Likelihood (REML) and Boundary Optimizer Based on Quadratic Approximation (BOBYQA). The participant and target word were included as separate random effects (intercept adjustment). The theoretical assumption for the current test can be represented as:

$$Y_{ijk} \sim N(\mu_{ij}, \delta^2) \tag{1}$$

where Y_{ijk} stands for the dependent variable for participant j, target word i and iteration k; μ_{ij} represents the value of the dependent variable for participant j and target word i for all iterations; and δ^2 is the variance component (parameter of the random effects model) that describes the variability within the iteration parameter. On the basis of this assumption, the regression model was computed according to the following equation:

$$\begin{aligned}\mu_{ij} = \beta_0 + \beta_1 \cdot language_{ij} + \beta_2 \cdot iteration_{ij} \\ +\beta_3 \cdot language_{ij} \times iteration_{ij} + u_{ij} + \tilde{u} + \epsilon_{ijk}\end{aligned} \tag{2}$$

where μ_{ij} stands for the value of dependent variable total for participant j and target word i for all iteration categories; β_0 represents the parameter of the fixed effects model, which is the true average value of the dependent variable ijk for all participants, the target word, and iteration, assuming the value of the reference levels for categorical variables; β_{1-3} stand for model coefficients representing effect sizes, provided that the values of the other variables of the model are controlled; u_{ij} indicates how far the actual values of the dependent variable participant j for the target word i differ from the average dependent variable for all target words i; $\tilde{u_i}$ indicates how far the average dependent variable deviates for all observations for all target words from the average value of the dependent variable in all target words, participants and repetitions; and ϵ_{ijk} determines how far the actual values of the dependent variable ijk deviate from the average for participant j and target word i.

In the first step, an unconditional null model (without predictors) was computed. Then, the *language* and *iteration* predictors were added. A significant ANOVA result examining the difference in variances between the models without and with a variable permitted the inclusion of the variables in the final model. Similarly, the next step tested the inclusion of interactions between the *language* and *iteration* predictors in the model. Collinearity parameters were estimated via variance inflation factors (VIF). The interpretation of the VIF values along with linearity of fitted models, homogeneity of variance, and normality of random effects was based on the recommendations by [29]. The analyzed sample consisted of 3741 observations (see supplementary material), measured on the pool of 41 participants (unique categories) across 23 target words in each tested language.

3. Results

The descriptive statistics for the numerical variables in tested models showed a small to moderate skewness (< 2.0) and kurtosis (< 8.0). The intercept-only model was significant: $Int = .13,\ 95\%CI(.12, .15), t(3737) = 15.40, p < .001$ (for vowel duration). Adding the predictors to the model resulted in significant decrease of AIC (Akaike Information Criterion) [30] and BIC (Bayesian Information Criterion) [31] and the log-likelihoods which provided the justification of the selected model ($\chi^2(2) = 116.91, p < .001$). However, accounting for interactions between predictors did not result in significant changes in deviance, ($\chi^2(2) = 2.16, p = .152$), therefore, only main effects were considered in the final model (henceforth: *mod1*). The model's total explanatory power was substantial (R^2 conditional $= .49$, fixed effects R^2 marginal $= .02$). The temporal differences across the iterations and languages are plotted on Figure 2 and the results of the fitted regression model are shown in Table 2. Across all target words the differences between stressed vowels were in the range of .13 s. The overall effect regarded the marginal increase of vowel duration in second iteration consistently across languages (see Figure 2).

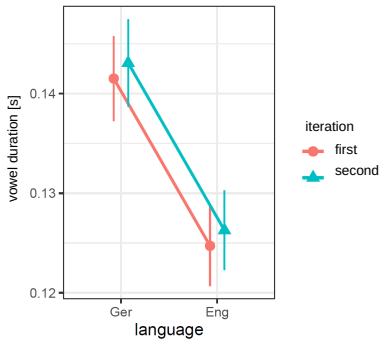

Figure 2: *Marginal means for vowel durations [s] across iterations and languages in the final model: mod1*

The null models accounting for spectral measures of formant frequencies moderated by target word and iteration factors were significant. The model testing the F1 scored: $Int = 603.41, 95\%CI(550.92, 655.90), t(3737) = 22.54, p < .001$; whereas the model testing the F2 scored: $Int = 1559.01, 95\%CI(1417.58, 1700.43), t(3737) = 21.61, p < .001$). Adding the predictors to the models resulted in an non-significant reduction in AIC, BIC, and log-likelihoods for first format values ($\chi^2(2) = 1.08, p = .583$). Adding the predictors to model accounting for second formant values resulted in significant decrease of the AIC, BIC, and log-likelihoods ($\chi^2(2) = 8.77, p = .012$), which proved that the model fitted the data well. The model's total explanatory power was substantial (R^2 conditional = .45, fixed effects R^2 marginal = .01). However, accounting for interactions between predictors did not result in significant changes in deviance ($\chi^2(1) = .01, p = .907$), therefore, only main effects were considered in the final model. No interaction effect was observed on the basis of the analyzed data.

The null model testing the mean f0 values on the target word level was significant and fitted the data well: $Int = 168.10, 95\%CI(154.26, 181.94), t(3737) = 23.81, p < .001$). Again, the addition of predictors did not significantly reduce the AIC, BIC, and log-likelihoods ($\chi^2(2) = .33, p = .845$). Similarly, the interactions did not reach the level of statistical significance ($\chi^2(3) = 2.26, p = .521$) and no significant differences were found between the tested categories.

Taken together, on the basis of the analyzed spectral and temporal features, no hyperarticulation effects were present in the second iteration of the target words.

Table 2: *Estimated marginal means (EMM) for the temporal predictors in a linear model mod1 and contrasts between the predictors*

Language	Iteration	EMM	SE	95% CI	Language / Iteration Contrast	SE	z.ratio	*p*
German	1	.142	.009	$.124 - .159$	German 1 - English 1	.002	10.85	$< .001$
English	1	.125	.009	$.108 - .142$	Ger/Eng 1 - Ger/Eng 2	.002	-1.01	.744
German	2	.143	.009	$.126 - .160$	German 1 - English 2	.002	6.95	$< .001$
English	2	.126	.009	$.109 - .143$	German 2 - English 1	$-.002$	-8.39	$< .001$

4. Discussion

Overall, we rejected our assumptions that robot-initiated clarification requests cause hyperarticulation and elicit clear speech. Our outcomes resonate with the findings by [32] and partly refer to [22]. On the basis of the gathered data, we rejected the hypothesis regarding the local intelligibility adjustments in HRI. The spectral measures (first and second format frequencies of stressed vowels and mean fundamental frequency) were not significantly different in the tested conditions. In line with findings reported by [32], we did not observe the exaggeration in the second iteration of the target word. Even though the spectral measures did not support our hypothesis, due to lack of the expansion of F1, F2 on the temporal mid-point of the vowels, and mean F0 on a word level - we managed to partly replicate the results by [33] that suggest segmental lengthening observed in computer-directed speech. We find a tendency for increased vowel duration upon clarification request issued by the robot.

However, this finding should be interpreted with caution, not only due to the reported statistical power, but also because lengthening fluctuation does not exclusively correspond to speech rate changes. The temporal measure of the stable segments of spectrum is a function of speech rate, lexical stress placement, idiosyncratic characteristics of speaker, and phonological surrounding of syllable nuclei. The effect of less prominent phonetic encoding present in routinized commands directed to talking agents also impacts the length of stressed continuants. Against our expectations, the spectral characteristics of the signal, measured in temporal mid point of the stressed vowels did not show significant differences across repetitions (in none of the languages). We only observed the marginal differences in temporal characteristics of the stressed vowels. Relatively small alternations of vowel length across iterations suggest that clarification requests initiated by the voice assistant cause only small-scale changes in the temporal domain.

5. Conclusions

This study examined the local adjustments of speech intelligibility in human-robot interaction. Our methodological approach involved clarifications elicited by a request issued by the talking agent in English and German. By referring to human-human interaction, and in line with hyperarticulation and hypoarticulation model [11], we assumed that clarification requests in HRI evoke hyperarticulation (defined on spectral and temporal plane) similarly to interaction between animate speakers. However, the gathered data did not confirm our initial hypotheses.

Possible explanations of speech lacking the hyperarticulation in the second iteration of the clarified tokens can touch upon the nature of interaction in which participants are aware that the robot has recorded the first instance of a target word, hence, when asked for a repetition, most similar pronunciation of the targets is desired. It seems that participants in order to clarify the misheard target words try to preserve the quality of the first iteration while showing a tendency to lower the speech rate only. In contrast to human-human interaction, participants may assume that talking agents will benefit from identical sound quality between the iterations and the enhancement strategy relies on changes of speech tempo. Therefore, selective application of clear speech features in HRI seems to differ from the enhancement patterns typical to human-human interaction. Participants may assume that similar surface representations of the words may facilitate the process of teaching the voice agent complex vocabulary. If the voice assistant tries to map the properties of both provided iterations of the target words, the degree of phonetic resemblance across the repetitions should be high. The only dimension which may help the robot comprehend the target word is the speech rate. Changing the phonetic encoding between the target words may result in further misunderstandings and evoke more clarification requests. Alternatively, natural and frequent interactions with voice assistants could have made robot-directed speech a distinctive mode of interaction that should no longer be studied in comparison with human-human interaction models. The hyper- and hypo-articulation theory should be fine-tuned with respect to animacy of an interlocutor. Such a supplement of the well-established theory finds its justification in a growing field of studies into HRI. In the future, studies need to show if HRI can be treated as a distinctive interaction scheme, possibly diverging from typical human-human interaction, even though many contemporary voice assistants are designed to imitate natural human-like interplay.

This study would certainly benefit from extending the set of target words and including other language pairs. The limitations of this approach also relate to quite monotonous experimental design, in which participants are exposed to numerous clarification requests. The randomization of the number of the clarification requests per token would also prevent the participants from habituating the pattern of the second iteration being understood by the talking agent. The experience in interacting with various ASR systems can also influence the speech enhancement strategies. Users more accustomed to talking agents may exhibit different clarification patterns than speakers rarely addressing the voice assistants. To fully understand the strategies of providing clarifications in HRI, some other characteristics of hyper- and hypo-articulation should be included, such as VOT, formant dynamics, or voice quality measures. A conversational study design would further help us to focus on suprasegmental features of phrases exceeding the level of the target words. Other possible extension of the study objectives could shift the focus to lexical analyses to test the semantic relations between the definitions of target words provided in human-human interaction compared with HRI.

6. Data availability

The supplementary material, experimental data, and code are publicly available in the following Open Science Framework (`https://osf.io/qwyzv/`) repository.

7. References

[1] T. Ammari, J. Kaye, J. Y. Tsai, and F. Bentley, "Music, Search, and IoT: How People (Really) Use Voice Assistants," *ACM Transactions on Computer-Human Interaction*, vol. 26, no. 3, pp. 1–28, 2019.

[2] M. Cohn and G. Zellou, "Prosodic Differences in Human- and Alexa-Directed Speech, but Similar Local Intelligibility Adjustments," *Frontiers in Communication*, vol. 6, 2021, 675704.

[3] S. Kriz, G. Anderson, and J. G. Trafton, "Robot-directed speech: Using language to assess first-time users' conceptualizations of a robot," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010, pp. 267–274.

[4] I. Lopatovska, "Personality dimensions of intelligent personal assistants," in *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, 2020, pp. 333–337.

[5] A. Gampe, K. Zahner-Ritter, J. J. Müller, and S. Schmid, "How children speak with their voice assistant Sila depends on what they think about her," *Computers in Human Behavior*, 2023, 107693.

[6] E. Raveh, I. Steiner, I. Siegert, I. Gessinger, and B. Möbius, "Comparing phonetic changes in computer-directed and human-directed speech," *Studientexte Zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung*, pp. 42–49, 2019.

[7] O. Ibrahim and G. Skantze, "Revisiting robot directed speech effects in spontaneous Human-Human-Robot interactions," in *Human Perspectives on Spoken Human-Machine Interaction*, 2021, pp. 6–10.

[8] A. Bonarini, "Communication in human-robot interaction," *Current Robotics Reports*, vol. 1, no. 4, pp. 279–285, 2020.

[9] I. Gessinger, E. Raveh, S. Le Maguer, B. Möbius, and I. Steiner, "Shadowing Synthesized Speech-Segmental Analysis of Phonetic Convergence," in *Proc. Interspeech*, 2017, pp. 3797–3801.

[10] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2006.

[11] B. Lindblom, "Explaining phonetic variation: A sketch of the H&H theory," *Speech Production and Speech Modelling*, pp. 403–439, 1990.

[12] V. Freeman, "Hyperarticulation as a signal of stance," *Journal of Phonetics*, vol. 45, pp. 1–11, 2014.

[13] A. J. Stent, M. K. Huffman, and S. E. Brennan, "Adapting speaking after evidence of misrecognition: Local and global hyperarticulation," *Speech Communication*, vol. 50, no. 3, pp. 163–178, 2008.

[14] I. Gessinger, A. Schweitzer, B. Andreeva, E. Raveh, B. Möbius, and I. Steiner, "Convergence of pitch accents in a shadowing task," in *International Conference on Speech Prosody*, 2018, pp. 225–229.

[15] R. Smiljanić and A. R. Bradlow, "Speaking and Hearing Clearly: Talker and Listener Factors in Speaking Style Changes," *Language and Linguistics Compass*, vol. 3, no. 1, pp. 236–264, 2009.

[16] J. Kudera, "Attuning to linguistically less-fluent interlocutors: evidence from convergence in Danish and Finnish foreigner talk," *Kwartalnik Neofilologiczny*, no. No 1, pp. 102–123, 2020.

[17] R. Lunsford, S. Oviatt, and A. M. Arthur, "Toward open-microphone engagement for multiparty interactions," in *Proceedings of the 8th international conference on Multimodal interfaces*, 2006, pp. 273–280.

[18] 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," in *Proc. Interspeech*, 2019, pp. 3980–3984.

[19] K. Vertanen, "Speech and Speech Recognition During Dictation Corrections," in *Proc. Interspeech*, 2006, pp. 1890–1893.

[20] J. J. Ohala, "Acoustic study of clear speech: A test of the contrastive hypothesis," in *Proceedings of the international symposium on prosody*, 1994, pp. 75–89.

[21] I. Siegert and J. Krüger, ""Speech Melody and Speech Content Didn't Fit Together"—Differences in Speech Behavior for Device Directed and Human Directed Interactions," *Advances in Data Science: Methodologies and Applications*, pp. 65–95, 2021.

[22] A. Gambino, J. Fox, and R. A. Ratan, "Building a stronger CASA: Extending the computers are social actors paradigm," *Human-Machine Communication*, vol. 1, pp. 71–85, 2020.

[23] B. R. Cowan, H. P. Branigan, M. Obregón, E. Bugis, and R. Beale, "Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human- computer dialogue," *International Journal of Human-Computer Studies*, vol. 83, pp. 27–42, 2015.

[24] H. Finger, C. Goeke, D. Diekamp, K. Standvoß, and P. König, "Labvanced: a unified javascript framework for online studies," in *International conference on computational social science (cologne)*. University of Osnabrück Cologne, 2017, pp. 1–3.

[25] A. Turk, S. Nakai, and M. Sugahara, "Acoustic segment durations in prosodic research: A practical guide," *Methods in empirical prosody research*, vol. 3, pp. 1–28, 2006.

[26] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," version 6.2.32, https://www.fon.hum.uva.nl/praat/.

[27] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, version 4.1.1, https://www.R-project.org/.

[28] D. Bates, R. Kliegl, S. Vasishth, and H. Baayen, "Parsimonious Mixed Models," *arXiv preprint:1506.04967*, 2015.

[29] J. Gareth, W. Daniela, H. Trevor, and T. Robert, *An introduction to statistical learning: with applications in R*. Spinger, 2013.

[30] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *2nd International Symposium on Information Theory*, B. Petrov and F. Csaki, Eds. Budapest: Akademiai Kiado, 1973, pp. 267–281.

[31] G. Schwarz, "Estimating the Dimension of a Model," *The Annals of Statistics*, pp. 461–464, 1978.

[32] J. Schertz, "Exaggeration of featural contrasts in clarifications of misheard speech in English," *Journal of Phonetics*, vol. 41, no. 3-4, pp. 249–263, 2013.

[33] D. Burnham, S. Joeffry, and L. Rice, "Computer- and Human-Directed Speech Before and After Correction," *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*, pp. 13–17, 2010.