Interspeech 2024
1-5 September 2024, Kos, Greece

The Use of Modifiers and f0 in Remote Referential Communication with Human and Computer Partners

Iona Gessinger[1], *Bistra Andreeva*[2], *Benjamin R. Cowan*[1]

[1]ADAPT Centre, University College Dublin, Ireland
[2]Language Science and Technology, Saarland University, Saarbrücken, Germany

iona.gessinger@ucd.ie, andreeva@lst.uni-saarland.de, benjamin.cowan@ucd.ie

Abstract

The present study investigates referring expressions in a remote interaction context with a human or computer partner (both simulated). Across these conditions, we compare the effect of competitor information being available to both partners (common ground) or only the speaker (privileged ground) on target item descriptions. We analyse the number of adjectival modifiers uttered and show that participants responded to the manipulation of information status in both partner conditions. In addition, we examine whether the information status also affects the prosodic realisation of the descriptions. No sufficient evidence was found for this. As expected, adjectives showed a slightly higher peak f0 when a competitor was present in the common ground than when there was no competitor. However, when analysing the overall f0 contour, there was no systematic difference between conditions.

Index Terms: prosody, referring expressions, privileged ground, common ground, remote, human-computer interaction

1. Introduction

When producing utterances in dialogue, speakers may consider only their own perspective (egocentric behaviour), or they may take the perspective of their interlocutor into account (allocentric behaviour). While it has been shown that speakers often apply audience design [1], it was proposed that egocentric language production may be the default behaviour [2, 3].

Referential communication is commonly studied in controlled visually-situated tasks such as the director-matcher task: Participants are presented with a set of objects. The speaker (i.e., director) describes an object which the listener (i.e., matcher) subsequently has to select. In this scenario, speakers often produce redundant item descriptions, i.e., they over-specify items [4]. Figure 1 shows an example of a visual scene. Referring to the target item highlighted in red as “green square” would be an over-specification, since there is no other square (i.e., competitor) in the scene. Such over-specification may be due to egocentric behaviour – use of a visually salient feature facilitates attribute selection and production for the speaker [5], or allocentric behaviour – early reduction of uncertainty regarding the referent facilitates comprehension for the listener [6].

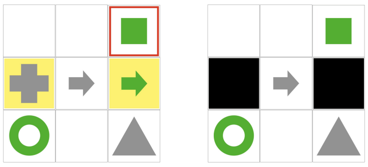

Figure 1: *Speaker perspective (left) and supposed listener perspective (right). Items with yellow background are in the privileged ground of the speaker, i.e., the listener does not see them. The target item is highlighted in red and has no competitor.*

From the listener’s perspective, the use of the adjective “green” could be interpreted as narrowing down the set of possible target objects. However, it could also indicate that the speaker sees another square which is not in the common ground, i.e., the use of the adjective could leak information about the speaker’s privileged ground [7]. The prosodic realisation of referring expressions provides another layer of information that could disambiguate these interpretations, e.g., through a stronger pitch accent on “green” highlighting a contrast to another square. Work by [8] explored the assumption that such information-leaking effects on the prosodic level are enhanced by asking speakers to conceal privileged information (due to ironic processes [9]). Comparing conditions with a competitor in common vs. privileged ground, adjectives from the latter were indeed produced with higher relative maximum f0 compared to the noun – and this was perceived by listeners as well.

Recent work investigated whether egocentric and allocentric language production occur to similar degrees when the interlocutor is human or a computer [10]. Results from [11] investigating lexical data suggest that speakers take the perspective of a computer partner less into account than that of a human partner when producing referring expressions. However, this effect was reduced when the computer was presented as a “separate” entity and not as an integrated part of the system in which the interaction took place – the latter may be seen as omniscient.

On the phonetic level, it has been shown that computer-directed speech is often carefully articulated for improved intelligibility, e.g., speech rate is slowed down [12] or vowels are hyperarticulated [13]. This is in line with computers being seen as inflexible “at-risk” listeners who may require speakers to adapt to their level of communicative ability [14].

In the context of spoken human-computer interaction (HCI), similar to human-human interaction over the phone, exploiting prosodic information for a better interpretation of the speaker’s perspective may contribute to smoother interactions. The present study therefore investigates whether the findings from [8], which was an in-person study, replicate in a remote context where the human partner is simulated. It further compares this to a condition in which the partner is a (simulated) voice assistant. Figure 1 shows an example turn of the present study. The visual scenes contain either *no competitor*, a *competitor in common ground* (visible to both interlocutors), or a *competitor in privileged ground* (only visible to the speaker). Competing items vary in colour or size. Participants are told that their partner must not find out what items are in their privileged ground (i.e., they are asked to conceal this information). The target language was Dutch in [8] and is German in the present study. For such West Germanic languages, it has been

10.21437/Interspeech.2024-1169

shown that differences in prosodic prominence correspond to the marking of information status. This involves differences in the categorical choice of accent type, as well as in the modulation of continuous phonetic parameters that characterise them. In German, high and rising accents with late peak are often used for new information, whereas low and falling accents with early peak are preferred for given information [15, 16, 17]. In production, speakers employ larger tonal onglide, higher f0 peak, and later peak alignment for contrastive focus compared to narrow focus [18, 19]. In perception, accents with rising onglide as well as higher f0 scaling of pitch accents are most frequently rated as encoding contrastive information [20, 21].

In this study, we examine proportions of adjectival modifiers produced and the shape of f0 contours of target item descriptions. We expect the competitor conditions to influence the amount of adjectives, with the highest occurrence when the competitor is in common ground and the lowest when there is no competitor. Regarding the privileged ground condition, we expect speakers to produce either as many adjectives as if there were a competitor in the common ground – since they see a competitor in their privileged ground – or as few adjectives as in the no competitor condition – since they believe their partners do not see a competitor. Partner type may affect adjective production, if participants take the perspective of the computer partner less into account than that of the human partner [11]. This would mean more adjectives occuring in the privileged ground condition, since participants behave egocentrically.

In terms of prosodic realisation, in line with previous research [18, 19], we expect the f0 contours to be more prominent for adjectives with competitors in common ground (*contrastive focus*) than for those without competitor (*narrow focus*), revealing, e.g., higher f0 peak and later peak position. Regarding the privileged ground condition, same as above, we expect speakers to produce adjectives either as prominently as if there were a competitor in the common ground, or at a similar prominence level as in the no competitor condition. Given the results in [8], the fact that we instruct participants to conceal information in privileged ground may (ironically) even lead to higher prominence than in the common ground condition. Partner type may affect the prosodic realisation in that the computer, as an "at-risk" listener [14], is addressed more carefully, manifesting in increased prominence compared to the human partner overall. This may possibly eliminate the effect of the competitor condition for the computer partner, as the prosodic prominence may already be at an upper limit.

2. Material and Methods

2.1. Participants

We recruited 47 native speakers of German via *Prolific Academic*, seven of whom had to be excluded due to technical difficulties. The final data set hence consisted of 40 native speakers of German, which were randomly assigned to a *human* or *computer* partner in the experiment. The human group (10 female, 10 male) had a mean age of 33.4 years (SD: 5.2, range: 25-45). The computer group (9 female, 9 male, 2 non-binary) had a mean age of 32.5 years (SD: 5.9, range: 25-44).

2.2. Manipulation Check: Partner Modelling

It is challenging to credibly simulate live interaction partners, be that a computer or even more so a human. Asking participants whether they believed in their partner's authenticity during the experiment may not lead to accurate answers, as the question itself causes suspicion. We hence used the German version of the Partner Modelling Questionnaire (PMQ), a validated self-report measure of perceived communicative ability of machines as dialogue partners [22, 23, 24], to investigate whether the two partner types where perceived differently. Perceptions of linguistic ability influence language production in speakers [25] and can therefore offer valuable context for the production of referring expressions in the present study. The PMQ uses 18 semantic differentials such as *consistent/ inconsistent*, *social/transactional*, and *spontaneous/predetermined* (rated on 7-point scales) to determine competence and dependability, human-likeness, and flexibility of the interlocutor. Given that our priority was to credibly simulate the interaction partners, we removed the items *human-like/machine-like* and *life-like/tool-like* from the PMQ presented to the participants of the human condition, as these items would explicitly raise the suspicion that they might not be talking to a real person. We considered this reduced 16-item PMQ for both participant groups. Analysis of the PMQ scores using Mann Whitney U Tests showed no statistically significant difference in the perception of competence and dependability ($p > .05$), yet found that participants rated the human partner as significantly more human-like ($W = 66.5, p < .001$; human: $median = 4.00, SD = .83$; computer: $median = 2.00, SD = 1.21$) and more flexible ($W = 93, p = .003$; human: $median = 3.83, SD = .93$; computer: $median = 2.33, SD = .90$) after interaction than the computer partner. This suggests that the human condition was indeed perceived differently than the computer condition.

2.3. Stimuli

The speech material for the human condition was recorded by a male native speaker of German aged 35 years. For the computer condition, it was synthesized using the male Standard German voice *Alex* by *CereProc* text-to-speech. The stimuli consisted of one longer utterance during which the partners introduced themselves, and 32 short utterances with the following structure: "Der/Das *[The]* <item> ist *[is]* <colour/size>." The introduction was delivered straightforwardly (11 s) by the computer voice and interspersed with delays and hesitations (49 s) by the human speaker to give the impression that the latter was taking part in the experiment just as spontaneously as the participant. The short utterances were produced once for every combination of item (arrow, circle, cross, square, star, triangle) with colour (green, grey) or size (big, small) by the computer voice, but several times by the human speaker to avoid reusing the same audio file more than once in the experiment for this condition. Only stimuli that were to be used during the practice trials of the experiment included intended delays and hesitations on the part of the human speaker, while stimuli for the experimental trials were produced in a similar fashion by computer and human. Since the trials in the online experiment had to be advanced by pressing a key on the keyboard, we made sure that the sound of a key press was audible in the human stimuli.

2.4. Procedure

The study was approved by the *Human Research Ethics Committee* at University College Dublin. It was implemented to run in a web browser using the JavaScript framework *jsPsych* [26]. Participants took part online on their personal computers after testing their microphone and providing informed consent. Crucially, it was explained in detail when audio recordings would be made and it was clearly signposted throughout the study when the participants' microphone was active. The participants were told that they would be connected live with a human partner or a voice assistant. For the human condition, “finding a partner” took longer and only worked during specific hours of the day to increase credibility. The study took approx. 30 min and was compensated with £4.50.

The experiment started with short introductions by the simulated partner (see 2.3) and the participant. Next, participants evaluated the communicative ability of their respective partner using the PMQ (see 2.2).[1] This was followed by a referential communication game including 6 practice trials and 36 experimental trials (14 per competitor condition; see Section 1). Each trial started with a participant turn where the participant asked their partner to click on the respective target item (see left grid in Figure 1). Since items varied regarding shape, colour, and size (see 2.3), participants could use any combination of these for their target item descriptions. They were reminded each time in writing that their partner must not find out which objects were hidden (yellow background). Two consecutive participant turns never contained the same target item shape, as this may lead to a contrastive pitch accent with respect to the previous trial and not the competitor item. Each participant turn was followed by a partner turn, which differed structurally to avoid priming a specific use of modifiers and f0. The participant view of a partner turn was similar to the right grid in Figure 1. However, a maximum of one item was visible in these turns and the partner stated: “The <item> is <colour/size>.” The participant had to indicate via button press whether – from their perspective – this statement was *correct*, *unclear*, or *false*. While the interaction in [8] was unilateral with participants always taking the role of the director, it was necessary for the simulated partners in the present study to take turns with the participants to signal their presence and establish their human or computer identity in the remote interaction. After the game, the PMQ was administered a second time. The study ended with the collection of demographic data. Participants were then fully debriefed.

3. Analysis and Results

The statistical analysis was carried out in R [27]. When fitting mixed models, we started with a base model only containing random intercepts for PARTICIPANT and added relevant fixed factors only if they improved model fit as determined by a likelihood ratio test and a minimum two-point decrease in the Akaike information criterion (AIC). Random slopes are explicitly mentioned or were otherwise omitted due to convergence issues.

3.1. Use of Modifiers

We excluded all target item descriptions that deviated from the expected structure of 0 to 2 adjectives followed by a noun (e.g., those using a relative clause or the grid location). Figure 2 indicates the number of target item descriptions per condition remaining in the analysis, and shows the proportions of adjectives used. We fitted a cumulative link mixed model (*ordinal* R package) for the dependent variable of producing 0, 1, or 2 adjectives. Including COMPETITOR CONDITION as a fixed factor (treatment coded, reference level: common; w/ slopes) in the base model improved the model fit ($\chi^2(7) = 604.12, p < .001; \Delta_{AIC} = 590.1$), while including PARTNER TYPE did not (w/o interaction, w/ slopes: $\chi^2(5) = 1.74, p = .88$; w/ interaction, w/o slopes: $\chi^2(3) = 3.11, p = .38$). We therefore omitted the latter concluding that the partner type had no influence on the number of adjectives produced. Pairwise comparisons using estimated marginal means indicate that more modifiers were used in common ground condition than privileged ground ($3.43, SE = .94, z = 3.67, p < .001$) and no competitor condition ($4.40, SE = .76, z = 5.79, p < .001$). In addition, more modifiers were used in privileged ground than no competitor condition ($.97, SE = .38, z = 2.54, p = .03$).

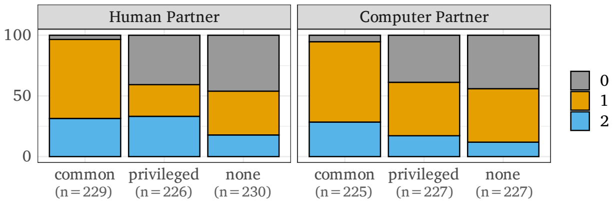

Figure 2: *Proportions of 0, 1, or 2 adjectives used in common ground, privileged ground, and no competitor conditions.*

3.2. Use of f0

For the acoustic analysis, we further reduced the data set to target item descriptions that contained only one adjective followed by a noun. Since participants where free in their choice of words, we restricted this to bisyllabic adjective realisations (*graue/n/r/s, große/n, gruene/n/r/s, kleine/n*) and monosyllabic (*Kreis, Kreuz, Pfeil, Plus, Ring, Stern*) or bisyllabic (*Donut, Dreieck, Kästchen, Quadrat, Rechteck, Viereck*) noun realisations. We further included only cases where the modifier referred to the correct contrast in the visual scene, i.e., if there was a competitor which differed in size, the adjective should refer to the size and not the colour of the item. Finally, we excluded utterances where major issues with recording quality would affect the f0 measurement, utterances that included a pause, as well as cases where a target item description was uttered a second time after a false start, as this is expected to affect f0. Figures 3 and 4 indicate the number of target item descriptions per condition remaining in the respective analysis. First, we applied the acoustic analysis method reported in [8] (see 3.2.1), then we conducted a functional analysis of the f0 contours (see 3.2.2) [28, 29, 30].

3.2.1. Difference in peak f0

Similar to [8], we measured peak f0 (Hz) in the adjective and noun using Praat [31] (autocorrelation method, range: 100-500 Hz for female voices, 75-300 Hz for male voices)[2] and visually checked for measurement errors. Cases where peak f0 could not be measured in either adjective or noun were removed from the data. Measured f0 values were converted to ERB [32] to approximate perceived pitch. Since pitch accents are perceived relative to one another [33], the peak of the noun was subtracted from the peak of the adjective. Figure 3 shows the mean difference of peak f0 (ERB) per condition, for which we fitted a linear mixed-effects model (LMM; *lme4* R package). Including COMPETITOR CONDITION as a fixed factor in the base model improved the model fit ($\chi^2(2) = 6.31, p = .04; \Delta_{AIC} = 2.13$), while including PARTNER TYPE did not (w/o interaction: $\chi^2(1) = .18, p = .67$; w/ interaction: $\chi^2(3) = .88, p = .83$). We therefore omitted the latter concluding that partner type had no influence on the difference of peak f0. Pairwise comparisons indicate that peak f0 was higher in adjectives with a competitor in common ground compared to the no competitor condition ($0.14, SE = .06, t =$

[1] This first administration of the PMQ was included as part of another research study and is not analysed in the present work.

[2] Based on their mean f0, the two non-binary speakers were grouped with the female speakers in the acoustic analysis.

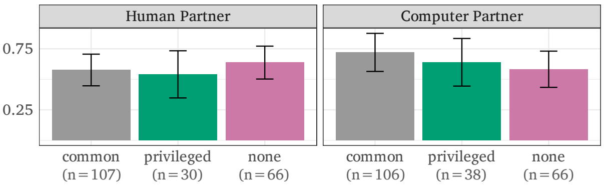

Figure 3: *Mean difference of peak f0 (ERB) in modifier and noun. Error bars indicate the 95% confidence intervals.*

Table 1: *Likelihood ratio test results with χ^2 and (p) values*

Noun	Factor	df	PC1	PC2	PC3
1 Syl	COMPETITOR	2	0.09 (.96)	4.25 (.12)	2.02 (.36)
	PARTNER	1	0.46 (.50)	2.43 (.12)	0.32 (.57)
	COMP*PART	5	1.95 (.86)	7.62 (.18)	3.49 (.62)
2 Syl	COMPETITOR	2	5.51 (.06)	1.39 (.50)	0.90 (.64)
	PARTNER	1	0.003 (.96)	0.20 (.66)	0.43 (.51)
	COMP*PART	5	6.09 (.30)	4.07 (.54)	1.50 (.91)

$2.47, p = .04$), but not for the privileged ground vs. common ground ($-0.09, SE = .07, t = -1.42, p = .33$) or no competitor condition ($0.05, SE = .07, t = 0.66, p = .78$).

3.2.2. *Functional analysis of the f0 contour*

We extracted f0 (Hz) in the adjective-noun phrase using the ESPS *get_f0* function [34] (sampling rate: 5 ms for female voices, 10 ms for male voices). All measurements beyond 1.5 standard deviations from the mean within each utterance were removed to reduce errors. Missing values inside the utterance were linearly interpolated, while missing values at the edges were extrapolated using a constant value of the nearest extreme. Resulting f0 contours were converted to semitones to reduce excursion variation. The mean semitone value within each utterance was subtracted from each contour to reduce speaker-dependent variation. After these pre-processing steps, cases with mono- and bisyllabic nouns were analysed separately. Utterances were time-aligned with respect to their beginning (L1), the adjective-noun boundary (L2), and their end (L3), using a landmark registration process (*landmarkregUtils* R package). We conducted multidimensional functional principal component analysis for the registered f0 curves and their associated time distortions, and fitted LMMs to the coefficients of the resulting three principal components (PCs).[3] The number of PCs was determined so that each would explain at least 10 % of the variance. Neither including COMPETITOR CONDITION nor PARTNER TYPE or their interaction in the base model improved the model fit for any of the three PCs in the mono- or bisyllabic noun cases (see Table 1). Figure 4 illustrates the f0 contours per condition as predicted by PC1. Durations were predicted across all conditions to be 47 s for the adjective and 56 s for the noun in the monosyllabic case, vs. 48 s and 67 s in the bisyllabic case.

4. Discussion and Conclusion

The analysis of modifier use demonstrated a clear influence of the information status manipulation. As expected, most adjectives were used when a competitor was present and visible to both interaction partners, while the fewest were used without a competitor present. The number of adjectives used when a competitor was present yet not visible to the partner fell between the two conditions. It differed statistically significantly from both, but leaned towards the condition without a competitor, indicating that the perspective of the partner was taken into account. This provided the basis for analysing f0 use.

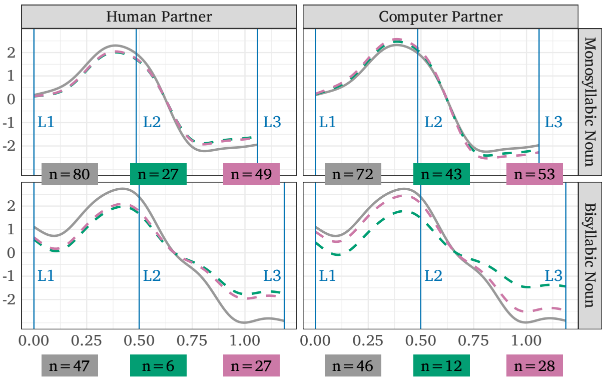

Figure 4: *PC1 predicted f0 contour (normalised semitones) over registered time in common ground* ■, *privileged ground* 🟩, *and no competitor* 🟪 *conditions; modifier (L1-L2), noun (L2-L3).*

The coarse analysis of peak f0 difference showed a slight distinction between the conditions in which both partners had the same information, with the adjective having a higher maximum f0 when there was a competitor in the common ground, as expected. However, taking the entire f0 contour into consideration, we could not substantiate a systematic difference between the three conditions of information status. The predicted f0 contours showed a clear peak towards the end of the adjective. However, peak height and temporal position did not differ significantly across conditions. In particular, the information that was to be concealed was not emphasised more than the information that was available to both partners. Even though the speakers were reminded to conceal the information at each turn, it remains unclear how aware they really were of this requirement throughout the interaction.

Having a human or computer partner did not make a difference to language production in this study. Where information status had an influence (modifier use), it did so for both partners, i.e., the speakers took the perspective of the computer into account as much as that of the human partner. One limitation of the present study is the credibility of the simulated partner. We showed that participants perceived the computer as less humanlike and flexible, but this may differ from actually believing to be speaking to a live partner. We acknowledge that especially the prosodic production depends on the actual intention to address an interlocutor. We therefore assume that live remote interactions may lead to different results.

While the role of the computer as an "at-risk" listener could have resulted in utterances already being produced more prominently overall, thereby eliminating an additional effect of information status, the remote interaction as a whole may have been interpreted as an "at-risk" context, placing the human partner in a similar role as the computer.

Additionally, since participants made the recordings for this study at home, their audio quality is generally inferior to recordings from the laboratory. The required inter- and extrapolation of missing values may have impacted the reliability of the f0 analysis. In the context of HCI, however, it is precisely this scenario that may be of interest: Could a voice assistant derive information about the user's privileged ground from their prosodic realisations on the go? The present study suggests that this may not be the case if only f0 is considered.

[3] See Appendix Figure A for more details about the principal components: `https://doi.org/10.17605/OSF.IO/VUAJM`

5. Acknowledgements

This research was conducted with the financial support of Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at University College Dublin. It was partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Project-ID 232722074 – SFB 1102. We would like to thank Constantijn Kaland and Marc Swerts for insightful brainstorming sessions. We are very grateful to Michele Gubian for his guidance in applying functional data analysis.

6. References

[1] V. S. Ferreira, "A mechanistic framework for explaining audience design in language production," *Annual Review of Psychology*, vol. 70, no. 1, pp. 29–51, 2019.

[2] G. S. Dell and P. M. Brown, "Mechanisms for listener-adaptation in language production: Limiting the role of the "model of the listener"," in *Bridges between Psychology and Linguistics: A Swarthmore Festschrift for Lila Gleitman*, D. J. Napoli and J. A. Kegl, Eds. Hillsdale, NJ: Erlbaum, 1991, pp. 105–130.

[3] W. S. Horton and B. Keysar, "When do speakers take into account common ground?" *Cognition*, vol. 59, no. 1, pp. 91–117, 1996.

[4] P. E. Engelhardt, K. G. Bailey, and F. Ferreira, "Do speakers and listeners observe the Gricean Maxim of Quantity?" *Journal of Memory and Language*, vol. 54, no. 4, pp. 554–573, 2006.

[5] R. Koolen, E. Krahmer, and M. Swerts, "How distractor objects trigger referential overspecification: Testing the effects of visual clutter and distractor distance," *Cognitive Science*, vol. 40, no. 7, pp. 1617–1647, 2016.

[6] E. N. Tourtouri, F. Delogu, L. Sikos, and M. W. Crocker, "Rational over-specification in visually-situated comprehension and production," *Journal of Cultural Cognitive Science*, vol. 3, p. 175–202, 2019.

[7] L. Wardlow Lane, M. Groisman, and V. S. Ferreira, "Don't talk about pink elephants! Speakers' control over leaking private information during language production," *Psychological Science*, vol. 17, no. 4, pp. 273–277, 2006.

[8] C. Kaland, E. Krahmer, and M. Swerts, "White bear effects in language production: Evidence from the prosodic realization of adjectives," *Language and Speech*, vol. 57, no. 4, pp. 470–486, 2014.

[9] D. M. Wegner, "Ironic processes of mental control," *Psychological Review*, vol. 101, pp. 34–52, 1994.

[10] J. Loy and V. Demberg, "Partner effects and individual differences on perspective taking," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 44, no. 44, 2022.

[11] P. R. Peña, P. Doyle, J. Edwards, D. Garaialde, D. Rough, A. Bleakley, L. Clark, A. T. Henriquez, H. Branigan, I. Gessinger, and B. R. Cowan, "Audience design and egocentrism in reference production during human-computer dialogue," *International Journal of Human-Computer Studies*, vol. 176, p. 103058, 2023.

[12] M. Cohn, K.-H. Liang, M. Sarian, G. Zellou, and Z. Yu, "Speech rate adjustments in conversations with an Amazon Alexa socialbot," *Frontiers in Communication*, vol. 6, 2021.

[13] D. Burnham, S. Joeffry, and L. Rice, "Computer- and human-directed speech before and after correction," in *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*, 2010, pp. 13–17.

[14] S. Oviatt, M. MacEachern, and G.-A. Levow, "Predicting hyperarticulate speech during human-computer error resolution," *Speech Communication*, vol. 24, no. 2, pp. 87–110, 1998.

[15] K. J. Kohler, "Terminal intonation patterns in single-accent utterances of German: phonetics, phonology and semantics," *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK)*, vol. 25, no. 1, pp. 15–185, 1991.

[16] S. Baumann, *The intonation of givenness: Evidence from German*. Walter de Gruyter, 2012, vol. 508.

[17] M. Grice, S. Ritter, H. Niemann, and T. B. Roettger, "Integrating the discreteness and continuity of intonational categories," *Journal of Phonetics*, vol. 64, pp. 90–107, 2017.

[18] S. Roessig, D. Mücke, and M. Grice, "The dynamics of intonation: Categorical and continuous variation in an attractor-based model," *PloS one*, vol. 14, no. 5, p. e0216859, 2019.

[19] S. Roessig, B. Winter, and D. Mücke, "Tracing the phonetic space of prosodic focus marking," *Frontiers in Artificial Intelligence*, vol. 5, p. 842546, 2022.

[20] F. Cangemi, M. Krüger, and M. Grice, "Listener-specific perception of speaker-specific production in intonation," *Individual Differences in Speech Production and Perception*, pp. 123–145, 2015.

[21] S. Ritter and M. Grice, "The role of tonal onglides in German nuclear pitch accents," *Language and Speech*, vol. 58, no. 1, pp. 114–128, 2015.

[22] P. R. Doyle, L. Clark, and B. R. Cowan, "What do we see in them? Identifying dimensions of partner models for speech interfaces using a psycholexical approach," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2021.

[23] P. R. Doyle, I. Gessinger, J. Edwards, L. Clark, O. Dumbleton, D. Garaialde, D. Rough, A. Bleakley, H. P. Branigan, and B. R. Cowan, "The Partner Modelling Questionnaire: A validated self-report measure of perceptions toward machines as dialogue partners," 2023, https://doi.org/10.48550/arXiv.2308.07164.

[24] K. Seaborn, I. Gessinger, S. Yoshida, B. R. Cowan, and P. R. Doyle, "Cross-cultural validation of partner models for voice user interfaces," in *Proceedings of the ACM Conference on Conversational User Interfaces*. New York, NY, USA: ACM, 2024.

[25] Z. G. Cai, Z. Sun, and N. Zhao, "Interlocutor modelling in lexical alignment: The role of linguistic competence," *Journal of Memory and Language*, vol. 121, p. 104278, 2021.

[26] J. R. De Leeuw, "jsPsych: A JavaScript library for creating behavioral experiments in a web browser," *Behavior Research Methods*, vol. 47, pp. 1–12, 2015.

[27] R Core Team, *R: A Language and Environment for Statistical Computing, version 4.3.2*, R Foundation for Statistical Computing, Vienna, Austria, 2023. [Online]. Available: https://www.R-project.org/

[28] M. Gubian, F. Cangemi, and L. Boves, "Automatic and data driven pitch contour manipulation with functional data analysis," in *Proceedings of Speech Prosody*, 2010, paper 954.

[29] M. Zellers, M. Gubian, and B. Post, "Redescribing intonational categories with functional data analysis," in *Proceedings of Interspeech*, 2010, pp. 1141–1144.

[30] M. Gubian, J. Harrington, M. Stevens, F. Schiel, and P. Warren, "Tracking the New Zealand English NEAR/SQUARE merger using functional principal components analysis," in *Proceedings of Interspeech*, 2019, pp. 296–300.

[31] P. Boersma and D. Weenink, *Praat: doing phonetics by computer, version 6.3.10*, 2023. [Online]. Available: http://www.praat.org/

[32] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, 1990.

[33] C. Gussenhoven, B. H. Repp, A. Rietveld, H. H. Rump, and J. Terken, "The perceptual prominence of fundamental frequency peaks," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 3009–3022, 1997.

[34] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech Coding and Synthesis*, pp. 497–518, 1995.