



Multimodal Dynamics of Hand Gestures and Pauses in Multiparty Interactions

Delphine Charuau, Naomi Harte

Sigma group, School of Engineering, Trinity College Dublin, Ireland

charuau@tcd.ie, nharte@tcd.ie

Abstract

This study examines multimodal patterns linking hand gestures and pauses in multiparty interactions, distinguishing between within-speaker pauses and between-speaker pauses. Using the MULTISIMO corpus, which includes annotated audiovisual recordings of collaborative dialogues, we analysed, for each category of gesture, their distribution, pause duration, and the timing of gesture onset and offset relative to the pause. Results showed distinct gestural timing patterns involving within-speaker and between-speaker pauses. Self-adaptors were associated with longer pauses, possibly reflecting increased cognitive demands or speech planning. While syntactic position influenced pause duration, with shorter pauses at utterance endings, it did not impact gesture onset and offset. Additionally, most pauses containing gestures occurred within utterances, regardless of gesture type.

Index Terms: Multimodality, hand gestures, pauses, timing, multiparty interaction

1. Introduction

Speech and gesture form an integrated system in human communication, where hand gestures often align with speech rhythms and prosodic structures [1][2]. Studies have demonstrated that beat gestures are closely synchronised with pitch accents [3][4], while representational gestures tend to coincide with semantically relevant lexical content [5]. Moreover, the presence of representational gestures in questions facilitates faster turn-taking, likely by providing additional semantic and pragmatic cues that help interlocutors anticipate and prepare their responses [6]. While these gestures are tightly coupled with speech, self-adaptors have traditionally been considered non-communicative, reflecting cognitive or emotional states rather than coordinating with speech production [7][8]. However, recent work showed that self-adaptors are more frequent around turn boundaries in conversation [9], suggesting a possible role in the regulation of turn-taking. Their distribution may therefore not be entirely random and could reflect underlying mechanisms that contribute to turn coordination.

Silent pauses fulfill multiple functions in conversation, including discourse structuring, speech planning, and turn-taking regulation. Within a speaker's turn, they contribute to structuring utterances by marking syntactic and semantic units [10], signal hesitation or speech planning [11]. In turn-taking mechanisms, their duration and location influence speaker transitions and floor retention [12][13]. Unlike filled pauses, which actively manage listener attention, silent pauses are more tightly linked to syntactic and prosodic constraints.

Although gestures are tightly coupled with rhythm and prosody, their dynamics during pauses has received less atten-

tion. While some studies suggest that gestures may persist in certain forms during disfluencies, especially through gestural holds [14], recent research has shown that speech and gesture can both be suspended simultaneously during disfluencies [15]. These findings point to disfluency as more than just a verbal phenomenon, highlighting its multimodal nature involving both vocal and visual elements. More broadly, the coordination between gestures and pause remains less explored, particularly in interactions where pauses not only serve cognitive and structuring functions, but also act as transition points in turn-taking dynamics.

This study investigates whether gestures exhibit distinct timing patterns depending on whether they occur within a speaker's turn or at a turn transition. Additionally, we examine how discourse structure influences this coordination. By analysing multimodal patterns, we aim to identify temporal relationships that reveal how hand gestures and speech coordinate during silences, depending on the communicative context of the pause and its position in the discourse structure.

2. Methods

2.1. Corpus

This study is based on the MULTISIMO Corpus [16], an open-access multimodal dataset designed for the analysis of multiparty interactions. The corpus consists of video and audio recordings of in-person triadic interactions, where two participants engage in a collaborative task under the supervision of a monitor. All participants are seated around a table, allowing for natural communication and multimodal exchanges. The data collection was designed to capture communicative behaviours, including speech, gestures, and other nonverbal signals.

The open-access dataset includes recordings from 39 participants (average age: 30 y.o.), of whom 27 were non-native and 12 native English speakers. The interactions were recorded using high-resolution cameras and high-quality microphones to ensure precise capture of both verbal and nonverbal cues. The dataset contains 18 recorded sessions, all of which were fully transcribed for speech and manually annotated for hand gestures. These 18 interactions were used for this study. Each session, lasting approximately 10 minutes, comprises an introduction phase, a collaborative task that consists of responding to three structured questions, and a conclusion phase. For this work, only segments involving active collaboration were analysed, while introductory and concluding parts were excluded due to limited interaction [17]. Although the task was structured around predefined questions, participants engaged in spontaneous discussion, allowing for naturalistic multimodal interactions. Annotations provided with the corpus include speech transcriptions, dialogue structures, pauses, and gestures.

2.2. Annotation of pauses and gestures

To investigate the relationship between pauses and gestures, a refined annotation scheme was applied. Since the original corpus labeled silent phases as 'No speaker' without distinguishing specific pauses types, additional annotations were automatically generated and manually verified. A total of 1103 pauses were extracted.

2.2.1. Annotation of pauses

In this study, we distinguished between within-speaker pauses, which occur within a speaker's turn, and between-speaker pauses, which are silences marking speaker transitions (turn-taking), considering only those longer than 200 ms. Additionally, pauses were automatically labeled according to their position within the discourse structure, using complete speech transcription. Thus, we distinguished pauses occurring in three locations: (1) end of a question, (2) end of an utterance, and (3) within the utterance (mid-utterance).

2.2.2. Annotation of gestures

For each pause (within-speaker pause and between-speaker pause), we identified whether it was accompanied by a gesture produced by the speaker preceding the pause (speaker 1). Pauses with gestures were then categorised based on the associated gesture type.

Gestures were initially annotated into five categories: beat, deictic, iconic, symbolic, and N/A (gestures where speakers touch their face, hair or scratch themselves). For this study, we reclassified them into four categories: symbolic gestures, representational gestures (grouping deictic and iconic gestures), beat gestures, and self-adaptors (formerly N/A). This grouping was due to the low frequency of deictic gestures during pauses. Additionally, as all three convey semantic content related to speech, combining them under representational gestures aligns with previous research on these gestures [6]. No symbolic gestures were observed during the pauses.

2.3. Measurements

For each pause type, we calculated the distribution of gestures by category. We then determined their distribution across different positions within the discourse structure. The duration of each within-speaker pause was measured as the interval between the end of speech and its resumption within the same turn. The duration of each between-speaker pause was measured as the interval between the end of one speaker's turn and the beginning of the next speaker's turn. These measurements were taken both with and without the presence of gesture.

To analyse the temporal relationship between gestures and pauses, we extracted the onset and offset of each gesture relative to the boundaries of the within-speaker pause and between-speaker pause in which it occurred. Given the variability in pause duration, a temporal normalisation was applied. For gestures initiated within the pause, the onset was normalised relative to the total duration of the pause: 0 represents the onset of the pause, while 1 corresponds to the pause offset. For gestures starting before the pause, the onset was normalised relative to the duration of the preceding interpausal speech unit. In this case, negative values indicate gestures initiated before the pause, with -1 representing the onset of the preceding interpausal unit (i.e., the speech segment between two pauses).

A similar normalisation was applied to determine the position of the gesture offset relative to the pause boundaries. In

this system, the values 0 and 1 represent the beginning and the end of the pause, respectively, while gestures extending beyond the pause received values between 1 and 2, where 2 represents a gesture ending one full pause duration after the pause itself.

2.4. Statistical analyses

To test the significance of our results, we used generalised linear mixed-effects models (GLMMs) implemented using the R `lme4` package. Given the continuous and bounded nature of gesture onset and offset, a Gamma distribution was applied. The dependent variables included pause duration, normalised gesture onset and offset, while the independent variables were pause type (between-speaker pause vs. within-speaker pause), gesture category, and discourse structure position. To account for individual variability and conversational context, we included random effects for speaker 1 linguistic background (native vs. non-native), their role (moderator vs. participant) and the section of the interaction. Yet, these random effects did not show significant variance, indicating these factors did not substantially impact the analysed timing variables. The significance threshold was set at 0.05, with pairwise comparisons adjusted using Bonferroni correction.

3. Results

3.1. Analysis of pauses

Our dataset contains 1103 pauses, with 707 classified as between-speaker pauses and 396 as within-speaker pauses occurring within the speaker's turn. Of the total 1103 pauses, 219 were accompanied by one gesture: 82 during between-speaker pause (11.6% of all between-speaker pause) and 137 during within-speaker pauses (34.6% of all within-speaker pauses).

Examining the types of gestures associated with pauses, we observed that representational gesture was the most frequent across both between-speaker pauses and within-speaker pauses. Among gestures observed in within-speaker pauses, 45.1% were associated with a representational gesture, 28% with a self-adaptor, and 26.8% with a beat gesture. Statistical analysis confirmed that representational gestures occurred significantly more than the other categories in pauses. For between-speaker pauses, a similar distribution emerged, with representational gesture remaining the most common (representational gesture = 37.3%; self-adaptor = 33.9%; beat gesture = 28.8%), but the differences between gesture types were not significant.

Figure 1 displays the distribution of gesture-accompanied pauses in relation to the discourse structure for both between-speaker pauses and within-speaker pauses. Gestures were significantly more frequent in mid-utterance compared to those occurring at the end of a question or the end of an utterance. Representational gestures occurred significantly more often than beat gestures, while self-adaptors were evenly distributed across all positions. Statistical analysis revealed that the type of pause did not have a significant effect on gesture occurrence, regardless of its position in the discourse.

3.2. Pause duration

Figure 2 illustrates pause duration according to the types of associated gesture, as well as pauses without gestures. The presence or absence of a gesture within a pause did not significantly impact its duration. However, when examining pauses that include a gesture, we found that within-speaker pauses with self-

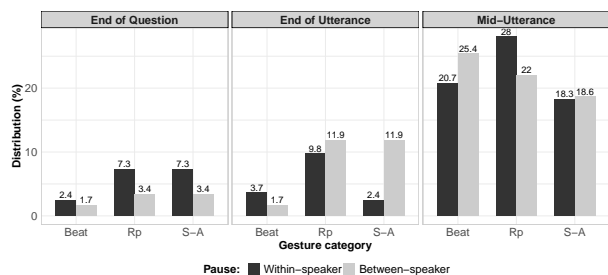


Figure 1: Distribution of gestures in between-speaker pause and within-speaker pauses according to the position in the discourse structure. Gesture categories are Beat, Rp = Representational, S-A = Self-adaptor

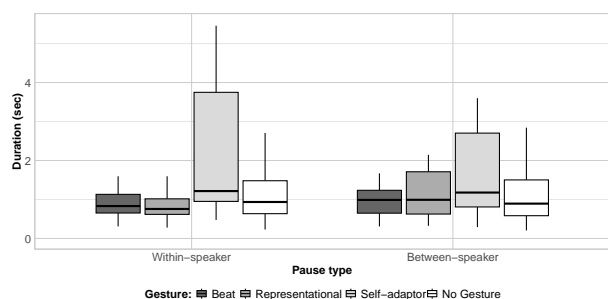


Figure 2: Duration of pauses with gestures in between-speaker pause and within-speaker pauses according to the type of gesture.

adaptors were significantly longer than those containing a representational gesture or beat gesture (median duration = 1.22 sec vs. 0.756 sec and 0.830 sec, respectively; $p = 0.037$). A similar trend was observed in between-speaker pauses, where pauses accompanied by a self-adaptor tended to be longer than those associated with other gesture types (self-adaptor = 1.18 sec; representational gesture = 0.991 sec; beat gesture = 0.988 sec), though this difference was not significant. Additionally, within-speaker pauses containing self-adaptors showed high variability in duration, indicating potential influences from inter-speaker variability or specific cognitive processing demands. In contrast, the duration of pauses containing beat gestures or representational gestures was relatively similar to that of pauses without gestures, indicating that their occurrence did not affect pause length.

Regarding location in the discourse structure, no overall significant effect was found on within-speaker pause and between-speaker pause duration. However, when examining specific interactions, we observed that for within-speaker pauses occurring at the end of a question, those containing a beat gesture or self-adaptor were significantly longer than those with representational gestures (median duration: beat gesture = 3.73 sec; self-adaptor = 3.80 sec; representational gesture = 0.770 sec; $p = 0.015$).

Finally, there was no significant difference in duration between between-speaker pauses and within-speaker pauses, with median durations of 0.92 sec and 0.931 sec, respectively. Thus, speaker transitions did not appear to influence pause length.

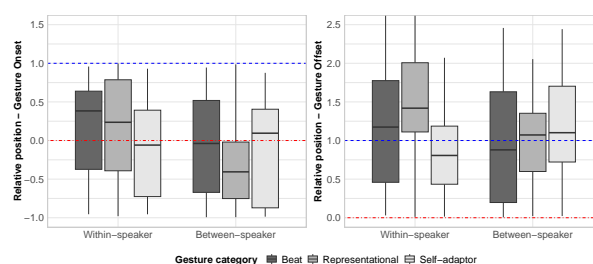


Figure 3: Normalised gesture onset (left) and offset (right) positions relative to pause boundaries (start: red dot-dashed line, end: blue dashed line), according to pause type (between-speaker pause and within-speaker pause) and gesture category.

3.3. Gestural timing in relation to pauses

3.3.1. Relative position of gesture onset

Figure 3 displays the distribution of gesture onset values relative to pause boundaries, according to pause type and gesture category. These values represent relative positions; therefore, no time units are reported. Onset values lower than 0 indicate that gestures began before the pause, during the preceding interpausal unit, while values greater than 0 indicate gestures initiated within the pause itself. Gestures occurring during between-speaker pauses presented different onsets compared to those in within-speaker pauses ($p = 0.036$).

Within-speaker pauses. Beat and representational gestures typically began within the pause, in its first half, with median onset values of 0.24 and 0.38, respectively. Self-adaptors tended to start before the pause, towards the end of the interpausal unit (median = -0.06). We observed significant differences for gestures occurring in mid-utterance ($p = 0.007$). Indeed, representational gestures and beat gestures started later during the within-speaker pause than when they occurred at the utterance or question boundaries, occurring closer to its center. Self-adaptors showed the opposite pattern: while their onset was inside within-speaker pauses located at the utterance or question boundaries, they were initiated before within-speaker pauses occurring within utterances, extending from the prepausal speech unit into the within-speaker pause itself (end of a question = 0.14; end of an utterance = 0.43; end of an utterance = -0.25).

Between-speaker pauses. Representational and beat gestures generally began before the pause, meaning they were initiated during the speaker's turn. Specifically, beat gestures tended to start towards the end of the interpausal unit (gesture onset median = -0.04). Representational gestures began much earlier (median = -0.41), almost at the center of the interpausal unit. In contrast, self-adaptors started during the between-speaker pause (median = 0.10), and showed high variability in timing. The syntactic position of the pause had little effect on beat gestures but strongly influenced self-adaptors ($p = 0.039$). When turn-taking occurred within an utterance, self-adaptors started within the between-speaker pause. However, when turn-taking occurred at question or utterance boundaries, self-adaptors were initiated earlier, before the start of the silence.

3.3.2. Relative position of gesture offset

Gesture offset differed significantly for between-speaker and within-speaker pauses ($p = 0.0015$), suggesting distinct coord-

dination patterns depending on whether the silence marks a speaker transition or occurs within a turn. To better understand these differences, we examined how gestures extended beyond the pause depending on their category and the discourse structure in which they occurred (Figure 3).

Within-speaker pauses. A significant proportion of representational gestures and beat gestures extended beyond the pause into the following interpausal unit, with median offset of 1.42 and 1.17, respectively. In contrast, self-adaptors tended to finish within the within-speaker pause, before its end. Additionally, we observed offset value exceeding 2, indicating that some gestures persisted beyond the next interpausal unit, overlapping with subsequent pauses, interpausal unit or turns. Gesture also tended to extend further when within-speaker pauses occurred at utterance boundaries ($p = 0.001$) or within utterances ($p = 0.001$). The pauses within ongoing speech did not systematically signal a boundary for gestural activity, allowing gestures to continue beyond the within-speaker pause.

Between-speaker pauses. Gesture offset tended to align with the end of the between-speaker pause. Beat gestures concluded at the closure of the silence (gesture offset median = 0.88), just before the next speaker took the floor, while representational gestures and self-adaptors tended to extend slightly beyond the pause, aligning with the beginning of the following speech unit (1.07 and 1.10, respectively). This pattern reflects a stronger synchronisation with the upcoming turn, reinforcing the temporal coordination between gesture offset and speech resumption. Unlike in within-speaker pauses, location of between-speaker pauses did not significantly influence gesture offset in between-speaker pauses, indicating that turn-taking constrained a regulation on gesture timing.

4. Discussion

This study examined how hand gestures align with within-speaker and between-speaker pauses in multiparty interactions. Our results highlight differences in gestural timing between these two types of silences.

Among all extracted pauses, only a small proportion contained hand gestures. The majority of within-speaker and between-speaker pauses occurred without gestures. Gestures were more frequent mid-utterance than in utterance or question boundaries, which may reflect their association with pauses linked to ongoing speech planning or cognitive processing [18][19]. Beat gestures, often associated with rhythmic emphasis and discourse structuring [1][20], were less frequent at the end of utterances and questions. Self-adaptors were evenly distributed across all positions, possibly indicating a broader cognitive function.

Additionally, our analysis revealed that pauses containing self-adaptors were longer than those with representational gestures and beat gestures. The duration of within-speaker pauses and between-speaker pauses with representational gestures and beat gestures did not differ from pauses without gestures, suggesting no apparent relationship between these gestures and pause length, unlike self-adaptors. Since longer pauses are often associated with cognitive load, speech planning, or hesitation [11], this reinforces the idea that self-adaptors serve as markers of cognitive effort and speech planning [7][21].

While within-speaker pause and between-speaker pause durations did not differ when accompanied by hand gestures, we observed notable difference in gestural timing. Gestures overlapping with between-speaker pauses tended to begin before the silence, confirming their strong temporal alignment with

speech. In contrast, gestures occurring during within-speaker pauses mostly started within the silence, raising questions about their function: do they emerge as hesitation markers, reinforcing the ongoing discourse, or as anticipatory movements linked to upcoming lexical or conceptual content? Regarding gesture offset, we found that gestures in between-speaker pauses typically ended toward the silence's closure, aligning with the onset of the next speaker's turn. This alignment suggests that gestural termination is tightly coupled with turn transitions. In contrast, gestures in within-speaker pauses often extended beyond the silence into the following speech unit, indicating a looser synchrony with the pause boundary and a potential overlap with upcoming speech.

This study has limitations. The relatively low number of pauses containing gestures constrains our findings. Additionally, the nature of the collaborative task likely increased cognitive load and speech planning demands, potentially affecting fluency and the distribution of pauses and gestures. Future work should consider different interaction contexts to assess the robustness of these patterns. These limitations also highlight the need for more open-access multimodal conversational datasets with rich annotations to support the study of gesture-speech coordination in varied communicative settings.

Nonetheless, this study provides new insights into the temporal coordination of gestures with within-speaker and between-speaker pauses, showing distinct synchronisation patterns depending on the type of silence. A more detailed analysis of gesture phases (e.g., extension, stroke, retraction) relative to pause boundaries could further refine our understanding of multimodal timing mechanisms. While syntactic position had limited influence on gesture timing overall, gestures in mid-utterance pauses exhibited distinct patterns. Future studies could explore different types of mid-utterance pauses — following filled pauses, at syntactic boundaries, or within syntactic phrases — to better capture their role in multimodal coordination. For representational gestures, examining the position of their semantic referent relative to the pause and gesture could clarify whether pauses serve as cognitive planning points for upcoming referents, or if gestures extend into the pause when the referent has already been introduced. Similarly, given their close ties to prosody, beat gestures should be analysed with prosodic features, using them as an additional anchor point in the study of gesture-pause synchronisation.

Understanding how gestures and pauses interact contributes to a more comprehensive model of multimodal communication, shedding light on how speakers manage turn-taking, speech planning, and discourse structuring. These insights are particularly relevant for refining multimodal dialogue systems, which have mainly focused on speech, overlooking the rich information conveyed by gestures and other modalities [22]. Integrating these non-verbal cues can enhance naturalness and efficiency in human-machine interactions.

5. Conclusion

This study highlights how the communicative function of silence shapes multimodal coordination, with gestures adapting differently in within-speaker and between-speaker pauses. Our findings reveal distinct synchronisation patterns, where gestures in between-speaker pauses align more closely with speech transitions, while those in within-speaker pauses show greater flexibility. Future research could explore the interplay between gestures, prosody, and speech content to refine our understanding of gesture-speech synchronisation in multiparty interactions.

6. Acknowledgements

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant number 22/FFP-A/11059.

7. References

- [1] D. McNeill, *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press: Chicago, 1992.
- [2] P. Wagner, Z. Malisz, and S. Kopp, “Gesture and speech in interaction: An overview,” *Speech Communication*, vol. 57, pp. 209–232, 2014.
- [3] D. P. Loehr, “Temporal, structural, and pragmatic synchrony between intonation and gesture,” *Laboratory phonology*, vol. 3, no. 1, pp. 71–89, 2012.
- [4] T. Leonard and F. Cummins, “The temporal relation between beat gestures and speech,” *Language and Cognitive Processes*, vol. 26, no. 10, pp. 1457–1471, 2011.
- [5] S. Kita and A. Özyürek, “What does cross-linguistic variation in semantic coordination of speech and gesture reveal?: Evidence for an interface representation of spatial thinking and speaking,” *Journal of Memory and language*, vol. 48, no. 1, pp. 16–32, 2003.
- [6] M. ter Bekke, L. Drijvers, and J. Holler, “Hand gestures have predictive potential during conversation: An investigation of the timing of gestures in relation to speech,” *Cognitive Science*, vol. 48, no. 1, p. e13407, 2024.
- [7] P. Ekman and W. V. Friesen, “The repertoire of nonverbal behavior: Categories, origins, usage, and coding,” *Semiotica*, vol. 1, no. 1, pp. 49–98, 1969.
- [8] P. Ekman, “Emotional and conversational nonverbal signals,” in *Language, knowledge, and representation: Proceedings of the sixth international colloquium on cognitive science (ICCS-99)*. Springer, 2004, pp. 39–50.
- [9] P. Żywiczyński, S. Wacewicz, and S. Orzechowski, “Adaptors and the turn-taking mechanism: The distribution of adaptors relative to turn borders in dyadic conversation,” *Interaction Studies*, vol. 18, no. 2, pp. 276–298, 2017.
- [10] I. Grosman, A. C. Simon, and L. Degand, “Variation de la durée des pauses silencieuses: impact de la syntaxe, du style de parole et des disfluences,” *Langages*, no. 211, pp. 13–40, 2018.
- [11] F. Goldman Eisler, *Psycholinguistics: Experiments in Spontaneous Speech*. Academic P.
- [12] M. Heldner and J. Edlund, “Pauses, gaps and overlaps in conversations,” *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.
- [13] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. De Ruiter, K.-E. Yoon *et al.*, “Universals and cultural variation in turn-taking in conversation,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10 587–10 592, 2009.
- [14] M. Graziano and M. Gullberg, “When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons,” *Frontiers in psychology*, vol. 9, p. 879, 2018.
- [15] L. Kosmala, M. Candea, and A. Morgenstern, “Synchronization of (dis) fluent speech and gesture: A multimodal approach to (dis) fluency,” in *Gesture and Speech in Interaction*, Sep. 2019.
- [16] M. Koutsombogera and C. Vogel, “Modeling collaborative multimodal behavior in group dialogues: The multisimo corpus,” in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018, pp. 2945–2951.
- [17] —, “Speech pause patterns in collaborative dialogs,” in *Innovations in Big Data Mining and Embedded Knowledge*. Springer International Publishing, 2019, pp. 99–115.
- [18] A. Piolat, “Localisation syntaxique des pauses et planification du discours,” *L’année psychologique*, vol. 83, no. 2, pp. 377–394, 1983.
- [19] H. Zhang and M. Liberman, “The syntactic, semantic, topic and socioeconomic effects on silent pause distribution,” in *Proceedings of 19th International Congress of Phonetics Sciences*, Aug. 2019, pp. 3378–3382.
- [20] L. M. Morett, S. H. Fraundorf, and J. C. McPartland, “Eye see what you’re saying: Contrastive use of beat gesture and pitch accent affects online interpretation of spoken discourse,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 47, no. 9, p. 1494, 2021.
- [21] M. Neff, N. Toothman, R. Bowmani, J. E. Fox Tree, and M. A. Walker, “Don’t scratch! self-adaptors reflect emotional stability,” in *Intelligent Virtual Agents: 10th International Conference, IVA 2011*, Sep. 2011, pp. 398–411.
- [22] A. Addelese, A. Eshghi, and I. Konstas, “Current challenges in spoken dialogue systems and why they are critical for those living with dementia,” *arXiv preprint arXiv:1909.06644*, 2019.