What If I Interrupt You

Liu Yang

What If I Interrupt You

LIU YANG, Institut des Systèmes Intelligents et de Robotique, CNRS, Sorbonne University, France

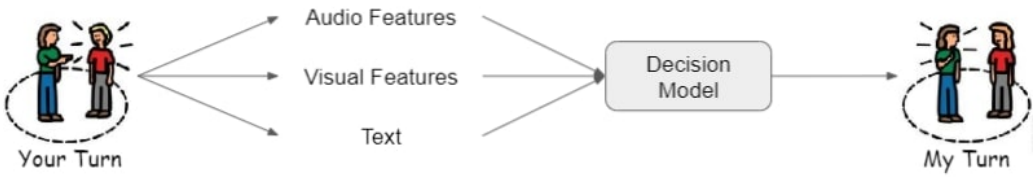

Fig. 1. A general view of interruption management model

Today, with the development of science and technology, we hope to improve human-agent interaction, agents will become more 'human-like', so that they communicate autonomously with humans both verbally and non-verbally. A challenge for the Embodied Conversational Agent is then to handle and manage speaking turn exchanges, and, more particularly interruptions, these are inherent in human-human interaction. We present our ongoing work on modeling interruption management in human-agent interaction. Our research contains two main aspects: 1) when and how should the virtual agent interrupt human users, and 2) how should the virtual agent respond when being interrupted by human users. To achieve this goal, we first started by analyzing human-human interaction data.

CCS Concepts: • **Human-centered computing** → **Human agent interaction (HAI)**.

Additional Key Words and Phrases: Nonverbal behaviour, interruption, turn-taking, embodied conversational agent

ACM Reference Format:
Liu YANG. 2021. What If I Interrupt You. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3462244.3481278

1 INTRODUCTION

With the development of science and technology, human-computer interfaces are more and more involved in daily life. Among them, Embodied Conversational Agents (ECA) are particularly appreciated as they allow 'human-like' interactions using verbal and nonverbal cues. However, there are still many difficulties in their development, such as interruption management.

In face-to-face interaction, interlocutors quickly and frequently exchange the roles of listener and speaker. Turn management is one of the skills needed for social interaction. During this exchange of turns, interruptions, overlaps or silences may occur. Even with cultural and gender variants [18, 23], researches show remarkable commonalities during

Manuscript submitted to ACM

interactions, such as the avoidance of overlapping or the minimum gap between turns [30]. So, exploring the structure of human conversation is important for research on human-agent interaction.

In most cases, during an interaction, speaking turns are exchanged smoothly and the transitions with no gap and no overlap are common [25]. Coordination is smooth when the listener and the speaker are coordinated: the listener waits for her turn or sends signals to indicate that she wants to take the turn. On the other hand, the listener may not want to take the turn signalled by the current speaker, resulting in silences, or may want to take the turn before the current speaker finishes, resulting in an interruption.

Actually, in natural interaction, the listener predicts roughly the duration and content remaining before the end of a sentence [12]. This ability is necessary to minimize pauses and speech overlaps between partners. Indeed, because the planning time necessary for the production of the next turn (600ms to 1500ms) [30] is much longer than the duration of gaps between turns (100ms to 300ms) [14], human should have already made the decision and started preparing for the coming turn exchange before the end of the utterance.

Humans are good at this kind of coordination to exchange the speaking turn with small overlaps or gaps, but for Conversational Agents, we always have the problem of sudden interruptions and long responding delay. Kirchhoff & Ostendorf also mentioned about the challenge about "multi" input/output in the conversation [19].

Giving the agent the capacity to interrupt human users or to respond in time to an interruption helps to improve the quality of interaction and to increase the engagement in the conversation [31]. Thus, the agent should be able to observe human signals and make its own decisions about when and how to interrupt, for example human tends to interrupt more at syntactic and prosodic boundaries [13]. Likewise, when interrupted by human, the agent should be able to recognize the type of interruption and replan its own behaviour on the fly.

2 BACKGROUND & RELATED WORKS

Participants quickly exchange turns during conversations, sometimes even with overlaps, but not all overlaps can be counted as interruptions. Shegloff calls a change of speaking turn an interruption when a participant B intervenes while the current speaker A still holds the floor of the conversation, without letting A finish their turn [28].

2.1 Interruption

Interruptions occur frequently in natural interactions and should be considered when modeling turn-taking. Interruptions in a conversation have different meanings. In order to appropriately interrupt the user or to respond to user's interruption, we should first classify the interruptions according to their purpose, then generate the most reliable reaction for the agent. According to Julia A. Goldberg, we can broadly divide them into two strategies: competitive and cooperative interruptions [11].

Competitive interruption occurs when the listener aims to control the interaction. This type of interruption usually disrupts the flow of dialogue between the interlocutors and can be seen as a conflict. A competitive interruption can be classified into four sub-types [22]:

- `Disagreement`: the listener disagrees with the speaker and expresses immediately her own opinion.
- `Floor taking`: the listener grabs the turn and continues its development.
- `Topic change`: it involves changing completely the current topic and starting a new one.
- `Tangentialization`: the listener sums up information from the current speaker to end the turn and avoid unwanted information.

On the opposite, cooperative interruptions aims to help to complete the current turn [22]:

- `Agreement`: the listener shows agreement, compliance, understanding or support with the speaker.
- `Assistance`: the listener provides the current speaker with a word, a phrase or an idea to complete the turn.
- `Clarification`: the listener asks the current speaker to clarify or explain to understand the message expressed by the speaker.

The two strategies are very similar in the local discourse characteristics, but they are very different in the way information is exchanged between the partners [22] and the social attitudes they conveyed [3].

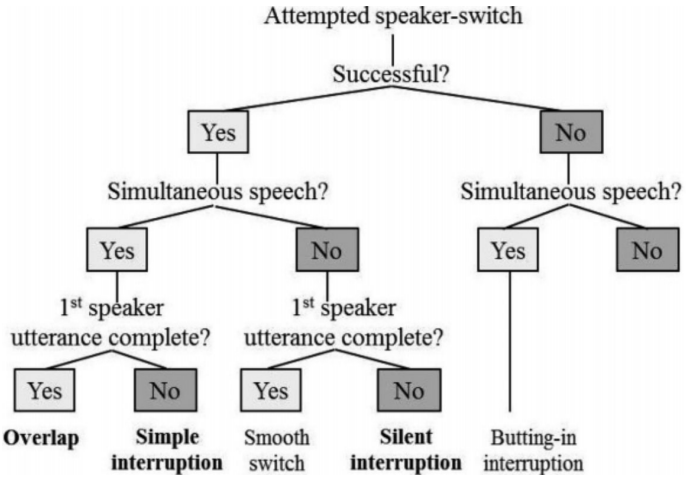

Fig. 2. Classification of interruption and smooth speaker exchange [2]

As we just said, an interruption could start with an overlap (simple interruption), or not (silent interruption), we also consider if the interruption is successful, or not (butting-in interruption).

Figure 2 shows the taxonomy of interruptions defined by Beattie [2]. We focus on simple, silent and butting-in interruptions in our research.

2.2 Models

Several computational models to predict when an interruption may occur or to simulate agent's behaviors in response to being interrupted have been proposed. They relied on analyzing human-human data.

Researchers show that one needs time to generate one's own speech before taking the turn [14, 23]; even a single word would take 600 ms, and another 200 ms for pre-articulation preparation such as breathing and vocal track [6, 27]. Some nonverbal behaviours such as head nod may be produced. They are faster to output than speech[15].

These nonverbal cues have been studied to predict upcoming interruptions or turn exchanges. A deep residual network is used in [5] to model acoustic features and predict the timing of interruption. Lee et al. [21] choose both speaker's acoustic cues and listener's gestural cues to predict the interruption timing in a dyadic conversation. Martin Johansson and Gabriel Skantze [16] highlight the importance of syntactic and semantic completeness, gaze direction and head pose for predicting the turn-taking opportunity in a collaborative multi-party human-robot interaction.

In previous works mentioned above [5, 16, 21], predictive models predict only the interruption timing while the interruption type is neglected.

To respond appropriately when being interrupted by a human, the agent should be able to distinguish different types of interruption. Lee et al. [20] compare the speech intensity, hand motion, and disfluency differences between different types of interruptions. Yang et al. [35] mention that competitive interruptions have typically higher pitch and louder amplitude than cooperative ones in order to gain attention, while cooperative interruptions are usually at lower or medium pitch levels. G.Skantze also gave a discussion about handling user interruptions [29].

Combining acoustic features, head movement and gaze direction, Truong's model [32] is able to distinguish between cooperative and competitive overlaps with a delay of 0.6 second after the start of overlaps. Troung also finds that competitive overlaps show higher levels in both intensity and energy in the mid-range frequency than cooperative overlaps.

Most of the interruption classification models mentioned above are based on long temporal windows before and after the interruption points, which lacks immediacy to apply in real-time human-agent interaction.

3 RESEARCH QUESTION

We aim to develop an ECA able to handle interruptions. We consider two aspects:

- `ECA interrupts human user`: the agent must decide when (timing) and how (interruption type) to interrupt the human user, and according to the human user's reaction, whether to continue to grab the turn or to abandon the interruption.
- `ECA interrupted by human user`: the agent should be able to recognize different types of interruption and to decide how to respond to user's interruption. It can ignore the interruption and continue the current turn, or can quit and yield the current turn to the human user.

4 APPROACH

To develop such a model, we started by analysing two multi-modal human-human interaction databases. We annotated them using an annotation schema for interruptions (see Section 6) and performed automatic acoustic and visual feature analysis for each interruption. This first study concerns both the timing of the interruptions and the multi-modal features before and during the interruption. In a next step, we aim to choose the most relevant features to build a decision model so the agent can be an interrupter and an interruptee.

When conversing, participants exchange multimodal signals, adapt to each other behaviours through imitation or synchronization [7, 25, 34]. We aim to also consider the agent's behavioural adaptation. Our idea is to generate a human-agent interactive loop, which takes current human and agent behaviours as input, and predicts the next agent's behaviour that will be used, in turn, as input for a future decision. This interactive loop between the agent and a human user, will allow the agent to adjust continuously its behaviour to adapt to the human's one and to better predict and react to interruption.

- `ECA interrupts human user`: We will develop a reinforcement learning model which has shown to be successful to model decision making process. The model will take as input multimodal data from both user and ECA: facial expressions (Action Units), body movement (hand gesture, body rotation, head movement), gaze direction and acoustic features (F0, Energy, MFCC). It will take also dialog acts extracted from human user's speech.
- `ECA interrupted by human user`: We propose to develop a Transformer [33] based multimodal model taking as input the nonverbal behavior of the human user: facial expression, body movement, gaze direction, acoustic

features. Speech content of the agent and of the human, described with dialog act and keywords, will also serve as input to the model. For these last features we will rely on incremental dialog processing technology [17].

Our work will be evaluated both quantitatively and qualitatively. In the first instance, performance measurements will help in the development of the model. Then, the model will be integrated into the GRETA platform [24] for a perceptual study to assess its credibility, quality and acceptability through face-to-face interaction with human users.

5 DATA

For this work, we make use of the IEMOCAP corpus [23]. IEMOCAP was collected to study different modalities in expressive speech. The database is made of five dyadic sessions. Each session consists of a pair of male-female actors acting 7 scripted plays and 8 spontaneous dialogues in predefined scenarios. We focus on the spontaneous part, which is close to five hours in total. The corpus was manually transcribed and segmented at the utterance level. It was also annotated with the emotion labels (happiness, anger, sadness, frustration and neutral state) [26]. We also use the NoXi corpus [4], which consists of free conversations in 45 given topics, in seven languages. We consider only the French part for our research (21 dyadic conversations), which is close to 7 hours. All videos are manually transcribed and segmented.

We have extracted visual information for both corpora. With Openface [1] we have extracted head rotation in 3 axes, and 17 action units (AU) coded with Facial Action Coding System (FACS) [8]. With Alphapose [10] we can get the movement of 15 key points (except the two points on the feet that we don't have on the video). In the NoXi database we have well separated audio sources that allow us to extract the acoustic features for each participant. This is not the case for the IEMOCAP database. So, for the NoXi database, acoustic features such as fundamental frequency, loudness, energy and MFCCs coefficients, are extracted using Opensmile [9] after a denoising process.

6 INTERRUPTION ANNOTATION

We manually annotated the interruptions for both IEMOCAP and NoXi databases with the annotation schema presented in Figure 3.

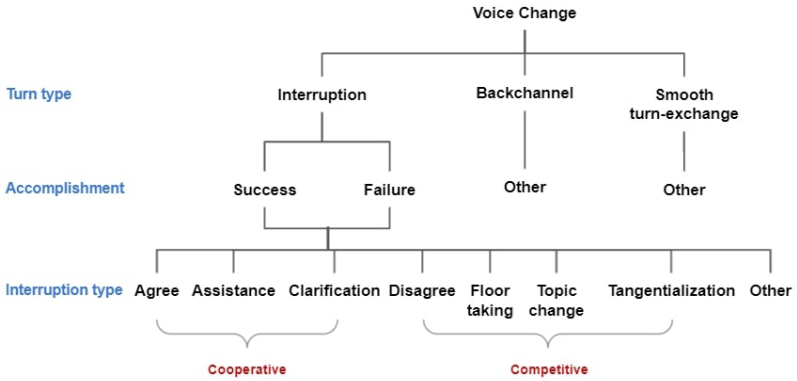

Fig. 3. Interruption annotation schema

According to different turn type, we first classify each voice track exchange into interruption, back-channel or smooth turn-exchange. Back-channels such as "hmm", "yeah" and "ok", who is not grabbing the turn, are considered as

interjects phatic responses to the speaker. Smooth turn-exchanges and interruptions are not distinguished by gaps and overlaps, one important variable is the semantic utterance completion. Even overlapped, the voice track exchange should be marked as smooth turn-exchange if the speaker completed the utterance. Even with gaps, the voice track exchange should be marked as interruption if the listener grabbed the turn before the speaker complete the utterance.

Smooth turn exchange correspond only to the successful speaking turn exchange, while interruptions can be successful and failed cases. Both cases can be marked as cooperative or competitive, and "other" when the interrupter quit too fast to recognize its type.

953 interruptions in IEMOCAP and 1367 interruptions in NoXi were annotated, in which we found a significant difference between successful interruptions and failed ones (IEMOCAP: 92.4% success vs. 7.6% failure, NoXi: 87.52% success vs. 12.48% failure). Most of the interruptions were initiated with overlaps (87.6% with overlaps). For both corpora, floor taking interruption takes the most part of competitive interruption (43.3% in IEMOCAP, 63.6% in NoXi), and agreement takes the most part of cooperative ones(63.7% in IEMOCAP, 79.02% in NoXi).

7 FUTURE WORK

We have presented our research question, general concept and data preparation with some primary insights. We still need to conduct further work. Our research plan is as follows:

- `2021`: We plan to finish the human-human interaction data analysis, the selection of features, and build the first decision model of "agent -> human interruption", able to decide when to interrupt and the type of the interruption.
- `2022`: We will develop the decision model of an "ECA interrupted by human user", able to recognize and respond to human user's interruption. Then we will integrate this decision model with a behavioural generation model and integrate it into the GRETA platform.
- `2023`: We will conduct perceptual studies with user experimentation, to evaluate the credibility, interaction quality and acceptability of our model.

8 DISCUSSION

Interruptions in human-human conversation occur often. They are part of the conversation dynamism. Depending of their type, interruptions can be perceived as being more or less polite by the current speaker. In human-agent interaction, it is important to understand when an agent can interrupt a user but also would the user tolerates such an act by the agent? Are the interruptions in human-agent interaction perceived as in human-human interaction? Will they have the same effect as expected?

We aim to develop an interruption model based on human-human interaction data. The data we consider are dyadic interactions in rather natural settings. It has been annotated at different levels: interruptions strategies, interruptions success, and nonverbal behaviors of both, interruptee and interrupter.

Another important aspect of our work is to understand how an interrupting agent is perceived at the level of the agent itself and of the intereaction quality.

9 CONCLUSION

Interruptions occur quite often during an interaction. Our aim is to endow an ECA with the capacity to handle them, being both an interrupter and an interruptee. We started by analysing human-human interaction data, and will develop a neuro-network model for interruption management.

ACKNOWLEDGMENTS

This work was performed as a part of IA ANR-DFG-JST Panorama and ANR-JST-CREST TAPAS (19-JSTS-0001-01) project.

REFERENCES

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1–10.

[2] Geoggrey W. Beattie. 1981. Interruption in conversational interaction, and its relation to the sex and status of the interactants. *Linguistics* 19, 1-2 (1981), 15–36.

[3] Angelo Cafaro, Nadine Glas, and Catherine Pelachaud. 2016. The effects of interrupting behavior on interpersonal attitude and engagement in dyadic interactions. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. 911–920.

[4] Angelo Cafaro, Johannes Wagner, Tobias Baur, Soumia Dermouche, Mercedes Torres Torres, Catherine Pelachaud, Elisabeth André, and Michel Valstar. 2017. The NoXi database: multimodal recordings of mediated novice-expert interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. 350–359.

[5] Adam Chỳlek, Jan Švec, and Luboš Šmídl. 2018. Learning to interrupt the user at the right time in incremental dialogue systems. In *International Conference on Text, Speech, and Dialogue*. Springer, 500–508.

[6] Eleanor Drake, Sonja Schaeffler, and Martin Corley. 2014. Articulatory effects of prediction during comprehension: an ultrasound tongue imaging approach. In *Proceedings of the 10th International Seminar on Speech Production (ISSP)*.

[7] Olga Egorow and Andreas Wendemuth. 2019. On Emotions as Features for Speech Overlaps Classification. *IEEE Transactions on Affective Computing* (2019).

[8] Paul Ekman and Wallace V Friesen. 1976. Measuring facial movement. *Environmental psychology and nonverbal behavior* 1, 1 (1976), 56–75.

[9] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.

[10] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. 2017. RMPE: Regional Multi-person Pose Estimation. In *ICCV*.

[11] Julia A. Goldberg. 1990. Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power- and rapport-oriented acts. *Journal of Pragmatics* 14, 6 (1990), 883–903.

[12] Francois Grosjean. 1996. Using prosody to predict the end of sentences in English and French: Normal and brain-damaged subjects. *Language and cognitive processes* 11, 1-2 (1996), 107–134.

[13] Rebecca Heins, Marita Franzke, Michael Durian, and Aruna Bayya. 1997. Turn-taking as a design principle for barge-in in spoken language systems. *International Journal of Speech Technology* 2, 2 (1997), 155–164.

[14] Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics* 38, 4 (2010), 555–568.

[15] Dirk Heylen. 2005. Challenges Ahead: Head movements and other social acts in conversations. In *Proceedings of the Joint Symposium on Virtual Social Agents*. 45–52.

[16] Martin Johansson and Gabriel Skantze. 2015. Opportunities and obligations to take turns in collaborative multi-party human-robot interaction. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 305–314.

[17] Casey Kennington, Spyridon Kousidis, and David Schlangen. 2014. Multimodal dialogue systems with inprotks and venice. In *Proceedings of the 18th SemDial Workshop on the Semantics and Pragmatics of Dialogue (DialWatt). Posters*.

[18] Catherine Kerbrat-Orecchioni. 1999. Les cultures de la conversation: Le langage en société. *Sciences humaines. Hors série* 27 (1999), 38–41.

[19] Katrin Kirchhoff and Mari Ostendorf. 2003. Directions for multi-party human-computer interaction research. In *Proceedings of the HLT-NAACL 2003 Workshop on Research Directions in Dialogue Processing*. 7–9.

[20] Chi-Chun Lee, Sungbok Lee, and Shrikanth S Narayanan. 2008. An analysis of multimodal cues of interruption in dyadic spoken interactions. In *Ninth Annual Conference of the International Speech Communication Association*.

[21] Chi-Chun Lee and Shrikanth Narayanan. 2010. Predicting interruptions in dyadic spoken interactions. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 5250–5253.

[22] Han Z. Li. 2001. Cooperative and Intrusive Interruptions in Inter- and Intracultural Dyadic Discourse. *Journal of Language and Social Psychology* 20, 3 (2001), 259–284.

[23] Daniel N. Maltz and Ruth A. Borker. 1983. A Cultural Approach to Male-Female Miscommunication.

[24] Isabella Poggi, Catherine Pelachaud, Fiorella de Rosis, Valeria Carofiglio, and Berardina De Carolis. 2005. Greta. a believable embodied conversational agent. In *Multimodal intelligent information presentation*. Springer, 3–25.

[25] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. *Language* 50, 4 (1974), 696–735.

[26] Tulika Saha, Aditya Patra, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards Emotion-aided Multi-modal Dialogue Act Classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4361–4372.

[27] Sonja Schaeffler, James M Scobbie, and Felix Schaeffler. 2014. Measuring reaction times: vocalisation vs. articulation. In *Proceedings of the 10th International Seminar in Speech Production (ISSP 10)*.

[28] Emanuel A Schegloff. 2001. Accounts of conduct in interaction: Interruption, overlap, and turn-taking. In *Handbook of sociological theory*. Springer, 287–321.

[29] Gabriel Skantze. 2020. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language* (2020), 101178.

[30] Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences* 106, 26 (2009), 10587–10592.

[31] Deborah Tannen. 1981. Indirectness in discourse: Ethnicity as conversational style. *Discourse Processes* 4, 3 (1981), 221–238.

[32] Khiet P Truong. 2013. Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee.. In *INTERSPEECH*. 1404–1408.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[34] Bill Wells and Sarah Macfarlane. 1998. Prosody as an interactional resource: Turn-projection and overlap. *Language and Speech* 41, 3-4 (1998), 265–294.

[35] Li-chiung Yang. 2001. Visualizing spoken discourse: Prosodic form and discourse functions of interruptions. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.