Now or When? Interruption timing prediction in dyadic interaction

Liu Yang, Catherine Achard, Catherine Pelachaud

Now or When? Interruption timing prediction in dyadic interaction

Liu YANG
yangl@isir.upmc.fr
ISIR,CNRS, Sorbonne University
Paris, France

Catherine ACHARD
catherine.achard@upmc.fr
ISIR,CNRS, Sorbonne University
Paris, France

Catherine PELACHAUD
catherine.pelachaud@upmc.fr
ISIR,CNRS, Sorbonne University
Paris, France

ABSTRACT

Interruptions are an important aspect of human-human communication. They help to adjust the conversation flow. Our aim is to equip virtual agents with the ability to handle interruptions, that is to decide when and how to interrupt their human interlocutor. In this paper, we focus on predicting when interruptions may occur during the conversation using multimodal features only from the speaker and propose a model trained on a corpus of dyadic interactions. To assess the model's accuracy, we conduct a perceptual study where we compare different timings (ground truth, randomly chosen or predicted by our model).

CCS CONCEPTS

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

KEYWORDS

Nonverbal Behaviour, Interruption Prediction, Turn-Taking, multimodality, Socially Interactive Agent

ACM Reference Format:
Liu YANG, Catherine ACHARD, and Catherine PELACHAUD. 2023. Now or When? Interruption timing prediction in dyadic interaction. In *ACM International Conference on Intelligent Virtual Agents (IVA '23), September 19–22, 2023, Würzburg, Germany*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3570945.3607293

1 INTRODUCTION

Interruptions are frequent in human communication, occurring in everyday conversations and playing a critical role in shaping the outcome of a conversation. They can either engage in interaction or disrupt the conversation flow, depending on the speaker's intent, the timing of the interruption, and the speaker's response[8]. In natural interactions, speakers switch speaking floors quickly and smoothly. Humans can predict when their partner's turn will end, enabling them to take the speaking turn without breaking the conversation flow [6].

Virtual agents are developed as interaction partners of human users in a variety of applications. To ensure seamless and natural interactions between humans and virtual agents we believe it is important to equip virtual agents with the ability to handle interruptions during interactions. In particular, they ought to predict when human users may interrupt them when they have the speaking turn.

In this paper, we propose a novel approach to find possible interruption initiation timing in dyadic interactions using multi-modal features only from the speaker since this model is to be applied to a virtual agent, of which the behaviour may be different from the real human. Our approach is based on a one-class classification model that has been trained on a corpus of dyadic interactions. We evaluate the model's accuracy through a perceptual study that compares model-predicted interruptions with ground truth data and random interruptions.

2 RELATED WORKS

Interruptions are common, but in most cases speaking turn exchanges smoothly during a conversation, smooth turn exchange is found predictable due to various cues that indicate the end of a turn: Ruth E. Corps et al.[4] proposed a model that predicts turn-ends by using the semantic content and timing of the preceding speech. S.C. Levinson et al.[9] provided insights into the systematic organization of turn-taking and its implications for processing models of language. Skantze [10] proposed a continuous model of turn-taking using LSTM recurrent neural networks, which takes into account contextual information. Crook et al.[5, 11] developed a model for handling user's interruptions when conversing with an embodied conversational agent; this model considers the user's intent and the agent's goals. Chỳlek et al.[3] proposed to use low-level acoustic features to predict interruptions and overlaps with a deep residual learning network. Their method allows for predicting interruption timings using the speaker's acoustic features.

Current studies highlight the importance of effective turn-taking and interruption management in human-agent interactions and focus more on handling the interruptions initiated by the human user, while it is also important that the agent interrupts the human user's floor and adjusts the conversation flow. We propose a one-class classification method using multimodal features such as acoustic features, head movement, and facial expression.

3 APPROACH

We use the French part of the NoXi corpus[2] for our study. NoXi is a multimodal database that contains free dyadic conversations(21 videos, 7 hours). According to the schema described in [12], 859 interruptions are annotatedObtaining the ground truth of positive samples (occurrence of interruptions) is thus possible but obtaining negative samples (where interruptions could have occurred but have not occurred) is more challenging. Even if an interruption did

IVA '23, September 19–22, 2023, Würzburg, Germany

ACM ISBN 978-1-4503-9994-4/23/09.
https://doi.org/10.1145/3570945.3607293

not occur at a given moment, it does not mean that it could not have happened: how to determine that it is not possible to interrupt at a given moment?

For this issue, Chỳlek et al.[3] assumed that the current speaker was purposefully not interrupted before a real interruption ($t-0.7$s), and consider this point as the negative sample. To overcome the limitation of missing negative samples we use a one-class classification model. Such a model does not need negative samples. It learns to detect interruptions based on only existing positive samples. In this article, we compared our approach with the method proposed by Chỳlek et al.[3] using multimodal features extracted on 1s length temporal window. We leveraged acoustic features extracted from openSmile[7]: fundamental frequency, loudness, and 12 mel-frequency cepstral coefficients (MFCC). We also extracted facial expressions, gaze, and head movements from OpenFace[1], including Action Units (AU) 01, 02, 04, 05, 12, and 15, gaze direction, as well as head position and rotation. For all multimodal features, we calculated their average values on the corresponding temporal window length (0.7s for the approach of Chỳlek et al.[3], 1s for our method) and used this feature vector as input to both models. The works of Chỳlek et al.[3] used a deep residual learning network (ResNet-152). Data is augmented by offsetting each moment by 1 to 3 samples. The method we proposed is a one-class SVM with specific hyperparameters $\gamma = 0.1 and \nu = 0.3$. The output of the one-class SVM is a score representing the similarity of the input features vector to the targeted class, which in our case is interruptions. We manually set a threshold on the output score based on the frequency of interruptions on the validation data.

To compare our one-class SVM model with the method presented by Chỳlek et al.[3], we followed a similar approach to use the annotated interruption onset moments as positive samples, and defined the moment -0.7s as negative samples, with an offset of 3 frames. For both methods, the model is trained on 19 conversations (validation set: 2 half videos from the 19) and tested on the 2 remaining ones. The results are presented in Table 1. It is important to note that we do not exactly obtain the results presented in [3] as we use another database where participants spoke another language and discussed other topics in a different interaction setting. The comparison showed that the inclusion of facial expressions and head motions enhances the prediction accuracy of interruptions. Moreover, our proposed one-class SVM model performs slightly better than the neural network model.

Table 1: Accuracy & F1-score for Deep residual learning network and One-class SVM models with different modality combinations.

	Accuracy	F1-score
Deep residual learning network [3] (acoustic only)	0.56	0.56
Deep residual learning network [3] (all modalities)	0.59	0.58
One-class SVM (ours)	0.61	0.61

4 SUBJECTIVE EVALUATION

We conduct a perceptual study to evaluate our timing prediction model. We compare ground truth annotations, predicted interruptions, and randomly selected ones. We consider 4 independent variables: interruption timing (ground truth, model predicted, randomly chosen), interrupter speech (ground truth, scripted), interrupter audio voice (natural human audio or synthesised voice), interruption type (agreement, clarification, disagreement) and we added one more variable to be tested: interruption turn (ground truth turn, during which interruptions were annotated in the real conversation, or false-positive turn, during which interruptions were predicted to occur but didn't occur in real conversation). The value of these variables is explained below. Thus we obtained 8 conditions we referred to as group 1...8 (see Figure 1). Each group was evaluated by 30 participants using a questionnaire composed of 11 questions. All the participants were fluent in speaking French.

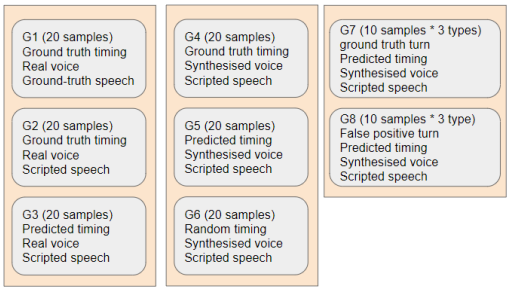

Figure 1: 8 groups with different conditions in interruption timing, interrupter speech, and interrupter audio voice.

Figure 2: Screenshot of generated video for an interruption.

4.1 Stimuli

To assess the predicted interruption timings, we compared them to the ground truth and randomly selected interruption timings. We utilized a static image featuring two stylized young individuals to accompany the interrupter and interruptee audios. As shown in Figure 2, the person on the left represents the interrupter, and the person on the right represents the interruptee. Subtitles were displayed concurrently with the audio for both the interrupter and interruptee beneath their corresponding silhouette.

From the NoXi database, we chose five interruptions categorized into three types: agreement, disagreement, and clarification (e.g. French: 'Ouais c'est ça ouais', English: 'Yeah that's it').

To examine the impact of the interrupter's voice, we used either natural human voice interrupter audios extracted from the videos of the NoXi database or synthesised voice. In all conditions, we used the original audio from the database for interruptees.

Table 2: Evaluation questions.

Do you think the interruption is	1. well placed? 2. acceptable? 3. coherent?
Do you think the interrupter is	4. competitive? 5. cooperative? 6. dominant? 7. friendly?
Do you think the interrupter	8. is trying to control the conversation? 9. intend to take the floor? 10. should let his interlocuter finish what he was about to say? 11. shouldn't have interrupted?

4.2 Comparison & results

For each video, participants answered 11 questions (see Table 2) related to the timing of the interruption, the type of interruption, and their perception of the interrupter, using a 5-point Likert scale. We report the significant differences between groups with a t-test (95% confidence of the p-value), Bonferroni correction was applied for multiple testing.

4.2.1 Interrupter speech. We compared the stimuli of groups 1, 2, and 3 to see how interrupter speech impacts the perception of the interruption. The interruptions of group 1 were evaluated as significantly more coherent (Q3) and cooperative (Q5) than those of groups 2 and 3, even though the interruption timings were the same for groups 1 and 2. There were no remarkable differences in competitiveness (Q4), dominance (Q6), and friendliness (Q7) between the three groups. Furthermore, the interrupters in group 1 were perceived as more likely to grab the turn (Q9) than the interrupters in groups 2 and 3. The interrupters of groups 1, 2, and 3 were rated as should let the speaker finish talking (Q10), but none of them tried to control the conversation (Q8).

4.2.2 Interruption timing. We compared the stimuli of groups 4, 5, and 6 to study how the different interruption timings are perceived. Group 4's interruptions were perceived as more acceptable (Q2) than groups 5 and 6. Compared to group 6, group 4's interruptions were also found to be better placed (Q1); the interrupters were perceived as more cooperative (Q5), friendly (Q7), and less competitive (Q4) and dominant (Q6), but no significant difference was found between the comparisons of group 4 vs. group 5, and group 5 vs. group 6. The interruptions of groups 4, 5, and 6 were perceived as coherent (Q3, score>3) but there was no significant difference between these groups. The interrupters of groups 5 and 6 were perceived as more likely to control the conversation (Q8) and "should not have interrupted" (Q11) than those of group 4. Compared to the interrupters of groups 4 and 5, the interrupters of group 6 were perceived as more likely to grab the floor (Q9) and should let the speaker finish the turn (Q10).

4.2.3 Interruption turn & interruption type. We examined the results obtained from the stimuli of groups 7 and 8 to investigate how false-positive turn interruptions were perceived and the impact of interruption types. The only significant difference between the stimuli of groups 7 and 8 is that the interrupters of group 7 were perceived as more friendly (Q7) than those of group 8. We further analyzed the differences between different interruption types. Agreement and clarification interruptions were perceived similarly in all aspects. While compared to disagreement interruptions, the agreement type was perceived as better placed (Q1), more acceptable (Q2), and coherent (Q3), with interrupters perceived as more cooperative (Q5) and friendly (Q7), and less competitive (Q4) and dominant (Q6). Clarification and disagreement interruptions had no significant differences in terms of placement (Q1) and coherence (Q3). Interrupters of agreement and clarification interruptions were more likely to control the conversation (Q8) and grab the turn (Q9) compared to those of disagreement interruptions.

4.2.4 Interrupter audio voice. To figure out the influence on perception when natural human voice or synthesised voice was used, we compared the results between group 2 and group 4, and between group 3 and group 5. The natural human voice group interruptions were perceived as more acceptable (Q2) than the synthesised voice group, and the interrupters of the natural, human voice group were perceived as less competitive (Q4) or dominant (Q6) than the synthesised voice ones, who were perceived as more like interrupting unnecessarily (Q10, Q11). Participants might become more sensitive to interruption timing when there is no longer natural intonation.

5 CONCLUSION

In this paper, we presented a novel approach to predict interruptions during conversations through the use of a one-class classification model with multimodal features from the speaker. To evaluate the effectiveness of our model, we conducted an objective study and a perceptual experiment to gain insights into how interruptions are perceived under different conditions. As a result, randomly chosen interruptions were rated rather similarly to the ground truth and predicted ones. Moreover, when using synthesised voice and scripted sentences, interrupting as in the ground truth or at other moments did not make a significant difference in perception. One possible reason is that interruptions may not have to occur at specific times in a conversation; there seems to be quite a lot of flexibility. But, this result is modulated by other factors (coherence of the interrupter speech sentences and voice quality). Interruption timing may not be the prime factor, rather the quality of the voice (natural vs. synthesized voice) seems more important.

In the near future, we plan to integrate our interruption prediction model into a virtual agent platform to allow the agent to interrupt the user as well as to react to the user's interruption.

ACKNOWLEDGMENTS

This work was performed as a part of the ANR-JST-CREST TAPAS (ANR-19-JSTS-0001) and IA ANR-DFG-JST Panorama (ANR-20-IADJ-0008) 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] 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.

[3] Adam Chỳlek, Jan Švec, and Luboš Šmídl. 2018. Learning to interrupt the user at the right time in incremental dialogue systems. In *Text, Speech, and Dialogue: 21st International Conference, TSD 2018, Brno, Czech Republic, September 11-14, 2018, Proceedings 21*. Springer, 500–508.

[4] Ruth E Corps, Martin J Pickering, and Chiara Gambi. 2019. Predicting turn-ends in discourse context. *Language, Cognition and Neuroscience* 34, 5 (2019), 615–627.

[5] Nigel Crook, Cameron Smith, Marc Cavazza, Stephen Pulman, Roger Moore, and Johan Boye. 2010. Handling user interruptions in an embodied conversational agent. In *Proceedings of the AAMAS International Workshop on Interacting with ECAs as Virtual Characters*. 27–33.

[6] Jan-Peter De Ruiter, Holger Mitterer, and Nick J Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language* 82, 3 (2006), 515–535.

[7] 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.

[8] Adam Kendon. 1990. *Conducting interaction: Patterns of behavior in focused encounters*. Vol. 7. CUP Archive.

[9] Stephen C Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology* 6 (2015), 731.

[10] Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language* 67 (2021), 101178.

[11] Cameron Smith, Nigel Crook, Daniel Charlton, Johan Boye, Raul Santos De La Camara, Markku Turunen, David Benyon, Björn Gambäck, Oli Mival, Nick Webb, et al. 2011. Interaction strategies for an affective conversational agent. *Presence* 20, 5 (2011), 395–411.

[12] Liu Yang, Catherine Achard, and Catherine Pelachaud. 2022. Annotating interruption in dyadic human interaction. In *Thirteenth Language Resources and Evaluation Conference, LREC*.