Modeling of interruptions in human-agent interaction

Liu Yang, Catherine Achard, Catherine I Pelachaud

Modeling of interruptions in human-agent interaction

Liu YANG
Institut des Systèmes Intelligents et de Robotique, Sorbonne University
yangl@isir.upmc.fr

Catherine ACHARD
Institut des Systèmes Intelligents et de Robotique, Sorbonne University
catherine.achard@upmc.fr

Catherine PELACHAUD
CNRS, Institut des Systèmes Intelligents et de Robotique, Sorbonne University
catherine.pelachaud@upmc.fr

ABSTRACT

Turn management is one of the skills necessary for social interactions. In human-human interactions, the exchange of turns is naturally completed by interruption, a communicative act connoting “cooperation” or “competition”. Interruptions, which at first glance can be considered as discourteous, are inherent in interaction and must therefore be modelled to create new interfaces such as Embodied Conversational Agent. A challenge is then to manage these agents, represented graphically so that they communicate autonomously with humans both verbally and nonverbally.

This article presents ongoing work whose goal is to animate an Embodied Conversational Agent capable of managing the interruptions in an interaction. In particular, we are interested in when and how to appropriately interrupt humans without offending them and how to respond appropriately when interrupted. In order to achieve this goal, we start by analyzing human-human interaction data.

CCS CONCEPTS

Human-centered computing → human-agent interaction (HAI).

KEYWORDS

• Nonverbal behaviour • Conversational interruption • Turn-taking • Embodied conversational agent (ECA)

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 (ECAs) are particularly appreciated as they allow natural interactions using verbal and nonverbal cues. However, there are still many difficulties in their development, such as interruption management for example.

In face-to-face conversation, interlocutors exchange quickly the role of speaker and listener in turns. Exploring the structure of human conversation is an important part of human interaction research. In interactions, humans adapt and adjust their behaviour according to that of their interlocutors. Partners exchange speaking turns which can give rise to interruptions, overlaps or silence. So, the management of turn-taking during conversation is one of the skills necessary for ECA development. Even with cultural and gender variants [3, 38], researches show remarkably commonalities across the languages, such as the avoidance of overlapping or the minimum gap between turns [2]. Actually, in natural interaction, the listener predicts roughly the duration and content remaining before the end of a sentence [1]. This ability allows to minimize pauses and speech overlaps between partners, in particular, because the planning time necessary for the production of the next turn (600ms to 1500ms) [2] is much longer than the duration of gaps between turns (100ms to 300ms) [14].

In most cases, during an interaction, the turn exchanges smoothly and the transitions with no gap and no overlap are common [4].

The coordination is smooth when the listener waits for his/her turn or sends signals to specify s/he wants to take the next speaking turn. On the other hand, the listener may not want to grab the turn signalled by the current speaker giving rise to silence or take the turn before the current speaker finishes leading to an interruption.

An Embodied Conversational Agent (ECA) is a human-like character that is able to communicate autonomously with human beings and the environment, using verbal and nonverbal communication. With the growing interest in human-agent interactions, it is desirable to make these interactions more natural and human-like. In this context, an ECA needs to be able to manage turn-taking mechanisms including interruptions, to handle human's interruptions and to react appropriately both verbally and nonverbally. Some researchers in the ECA field are also working on interruption [39, 40], but they are more focused on verbal content, and the impact of nonverbal behaviour has not attracted enough attention. However, recent theoretical studies and experimental results show that the information conveyed by nonverbal behaviours also greatly affects interaction[41, 42].

Nonverbal communication is the transmission of messages or signals through nonverbal platforms. It includes the use of visual cues such as body language, distance and physical environment/appearance, voice and touch. It can also include the use of time and eye contact as well as the movement when talking and listening, gaze behaviour, dilated pupils and blink rate.

Giving the agent the opportunity to interrupt or respond to an interruption helps to improve the effectiveness of communication and increase engagement. Thus, the agent should be able to make decisions about the timing and the type of interruption using different modalities of human signals (acoustic, linguistic, facial expression, motion, and emotion). Likewise, when faced with a human interruption, the agent will have to decide whether or not to cede its speaking turn and immediately replan its own behaviour to ensure that communication runs smoothly.

In Section2, we describe different interruption types and their effects by introducing the interruption taxonomy used in our research. Then we quickly review the research results on interruption characterization and prediction models. The corpus, objective and expected result will be presented in Section 3, before a short conclusion in Section 4.

2 Related works

In this section, we present some previous research on interruption classification, interruption prediction, and human interruption management models.

2.1 Interruption

Interruptions are natural and frequent in real interactions. They occur when one person interrupts to speak while the other person is still speaking and can be regarded as a deviation from the simple turn-taking model. Interruptions act to mediate the content and redirection of a conversational exchange. They can be broadly divided into two strategies: competitive and cooperative interruptions [7]. Both interruption strategies are very similar in their local discourse characteristics, but their global roles in helping interlocutors to exchange information are quite different [8].

Competitive disruption occurs when the listener interrupts to control the interaction, usually disrupting the flow of dialogue between the partners and can be seen as a conflict. A competitive interruption could be:

- Disagreement: The listener disagrees with what the current speaker is saying and wants to express their opinion immediately.
- Floor taking: The switch does not intend to change the topic of the current speaker and usually expands on the current topic by speaking from the current speaker.
- Topic change: to accomplish the task of changing the subject.
- Tangentialization: the listener summing up information from the current speaker to prevent listening to unwanted information [8].

On the opposite, a cooperative interruption could be:

- Agreement: Show agreement, compliance, understanding or support.
- Assistance: The switch provides the current speaker with a word, a phrase or an idea.
- Clarification: to understand the message sent by the speaker. The purpose is to ask the current speaker to clarify or explain information about which the listener is not clear [8].

Beattie [9] has defined a taxonomy of interruptions as shown in Figure 1. In our work, we focus on simple, silent, butting-in interruptions.

Studies in [3, 14] show that the interrupter starts the preparation of his speech around 600ms before speaking, and needs 200 ms for pre-articulation preparation (i.e. breathing, vocal track) [24, 25]. During this time needed for interruption preparation, the interrupter exhibits nonverbal behaviours such as gesture [26], head movement [27], etc.

Thus, the nonverbal behaviour of the listener has been studied to predict interruption or recognize its type as presented in the following sections.

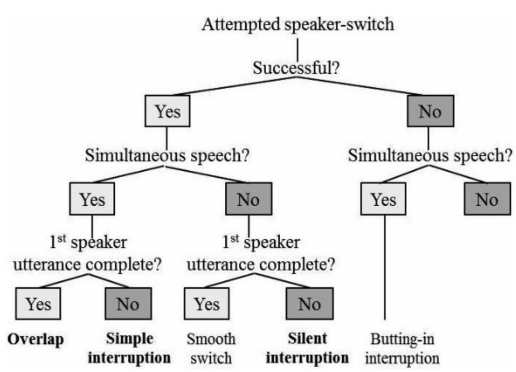

Figure 1: Classification of interruption and smooth speaker exchange [9]

2.2 Models

Differently from the previous studies [44, 45] based on turn-by-turn systems that implement the generation process only after the human user finishes the utterance, interruptions request the system to understand the dialogue incrementally [46, 48] to give the human user an immediate response. Incremental dialogue systems can prevent the user from speaking for a long time without being understood by the system.

To give the response at an appropriate timing, the agent must be able to predict the ending time of the current turn [47] to decide when to interrupt or to stop when an interruption comes up.

In order to apply the interruption model into real-time human-agent interaction, it should be continuous at each timestep, not only at specific event points [50].

To predict the future interruption on a multimodal dyadic interaction corpus, Lee et al. [28] used both the foreground speaker's acoustic cues and the listener's gestural cues.

Gravano [29] analysed acoustic features of a telephonic conversation corpus. Compared to other turn transition yields, the analysis result shows significant differences for interruptions in intensity and pitch level, speaking rate and IPU duration (Inter-pausal unit: a maximal sequence of words surrounded by silence).

In [30] a deep residual network is used to model acoustic features and predict the timing of interruption.

Hara [32] presented a turn-taking prediction model by predicting firstly the Transition Relevance Places (TRP), which are important for the interruption. Actually, cooperative interruptions are closer to the TRPs than competitive ones which can possibly occur in the middle of a sentence [31].

Martin Johansson and Gabriel Skantze [51] indicated 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.

All the presented models [28, 29, 30, 32] used the verbal or nonverbal behaviours that arise during a time period just before the interruption point. It means these models use the preparatory action before the interruption to detect whether the interruption will occur or not, while [51] implements only at each IPU.

When a user turn comes up while the agent is still speaking, [50] allows distinguishing between the backchannels from the interruptions.

For the classification of different types of interruption, Lee et al. [21] analyzed the differences in speech intensity, hand motion, and disfluency between the interlocutors during cooperative and competitive interruptions. The discriminant analysis shows that the use of multimodal cues provides a significant improvement in classification accuracy between the two types of interruptions while any individual single modality cue does not show much improvement.

Yang et al. [22] also mentioned acoustic and prosodic differences in both types of interruption. Competitive interruptions have typically higher pitch and amplitude to gain attention while cooperative interruptions often occur at low or medium pitch levels because of their non-competitive nature.

Combining acoustic features, head movement and gaze direction, Truong and Khiet's model [12] is able to distinguish between cooperative and competitive overlaps with a delay of 0.6s after the start of overlaps. Meanwhile, they found that competitive overlaps show higher levels in both intensity and max energy in the mid-range frequency than cooperative overlaps.

2.3 Innovation

Communication is an extremely complex process, related to body language, context, acoustics, language structure, emotions, etc [4, 33, 34, 35]. The models mentioned above only consider a single or a combination of factors, most of which are based on the analysis of speech content and acoustic variables, and do not consider the links between the different factors, especially nonverbal behaviour.

During an interaction, humans adjust their behaviour according to the behaviour of the other party. However, the human interruption management model [39, 40] generates verbal and nonverbal feedback only based on the acoustic characteristics and speech content of humans, without considering the adaptation of the human user's nonverbal behaviour.

We noticed that the predictive models predict only the timing but not the type of interruption.

Moreover, most of the classification models are based on features estimated on a temporal window and thus do not allow to generate non-verbal features in real-time human agent interaction.

The innovation of our work is to apply the interruption model to the human-agent interaction and implement it in real-time conversation. It requires that ECA has the ability to raise an interruption at an appropriate moment and in a proper way. We are more inclined to study how to make an interruption decision indicating when and how to interrupt, then signal the upcoming interruption and meanwhile adapt the agent behaviour to the human's.

Our interruption model will be developed by learning the human-human interaction data. To choose the most useful features, we start first to analyze different modalities of human-human interaction data.

Based on the previous models and experiments, we can see that prosodic features, acoustic features, speech content, gesture, head movement, and TRP playing important role in turn-taking and interruption prediction, classification and response generation, we tend to combine all these aspects and consider emotion, and dialogue act as additional items for analysis.

3 Project

We plan to develop an ECA for Social Skill Training (SST) for a large variety of population facing difficulties interacting with others. Here we present the objectives and methodology of our work.

3.1 Objective & Work in progress

We plan to develop an ECA as a tool for Social Skill Training (SST) for a large variety of populations facing difficulties when interacting with others.

Our goal is to develop an interruption model for human-agent interaction. It requires that ECA has the ability to interrupt its conversational partner and to react when being interrupted. We are interested in designing a decision model that computes on one hand, when and how to interrupt, and on the other hand decide to be interrupted or to keep the speaking turn. We will also develop a behavioural generation model that animates the agent during an interruption. These tools, inserted in a real-time human-agent interaction, aims to improve the quality of the interaction.

To reach our goal, we have first analysed a multi-modal human-human interaction corpus and performed automatic feature analysis for each type of interruptions (see Section 3.2). This study concerns both the timing of the interruption and the behavioural multi-modal features before and during the interruption. Based on the analysis results, we are currently choosing the most useful features for the decision models. They should be applicable to real-time interaction, which contains two main aspects:

- Interruptions raised by the ECA: when and how to interrupt the human user, including the decision of interruption timing, interruption type (cooperative / competitive) and the decision after interruption, whether to grab the turn or to abandon the interruption, depending on the human user's reaction.
- ECA interrupted by human user: how to respond to user interruptions, including the decision of whether to ignore the interruption and continue with the current turn or quit the current turn and yield it to the human user, as well as the quit timing.

We also consider the agent's behavioural adaptation. So matching with the decision models, we aim to develop the behavioural generation models to adapt the agent's interruption behaviour (verbal \& non-verbal) to the human user's one.

We will integrate the models on the Greta platform to schedule interruptions and agent's behaviours during human-agent interaction.

Each step of this work will be evaluated, both quantitatively and qualitatively. While the first will be based on performance measurements, the second one will rely on perceptual studies, based on face-to-face interaction with the agent. The evaluations will cover the credibility of the agent, the quality of the interaction, the acceptability of the agent and its behaviour.

3.2 Corpus

We searched for corpora with human-human interaction and found 6 available databases: IEMOCAP, CCdB [15], French-Spanish(French part only) [16], DUEL (French, German and Chinese part) [17], CreativeIT [18], MHHRI (Human-Human part only) [19]. We compared the corpora along with different criteria that are the availability of the spoken languages and of the different annotations, as well as the total duration of the available data. In particular, we are interested in the annotation of speech transcription, speaker turn, emotion as well as various multimodal signals (prosody, posture, facial expression, head movement, etc).

The table in APPENDIX gives an overview of the comparison. First, based on language (French or English) and duration, we can choose IEMOCAP or the French part of DUEL. They show not much difference in almost all criteria, but DUEL does not record the facial expressions for the interlocutors, which is a very important factor for interruption analysis. Moreover, as DUEL only uses a single camera on the side to record the conversation video, it is not possible to extract facial expressions using Openface [20] for example. So finally we choose to use IEMOCAP for our study.

IEMOCAP corpus [10] 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, 12 hours of dyadic conversations in total.

The spontaneous part is close to five hours in total, the average duration of a single video is about 3.8 minutes (Minimum 1.5 minutes, maximum 6.8 minutes).

For each pair of actors, 16 scenarios have been filmed with 61 markers attached to one of the two participants (53 on the face, 2 on the head, and 3 on each hand). These markers were attached to record the (x, y, z) positions.

The corpus was manually transcribed and segmented at the utterance level. It was annotated with the dialogue act and emotion labels [11].

After simple statistics on these available annotations, the spontaneous part of IEMOCAP has 4704 turns in total, female actors have 2336 turns, and male actors have 2374 turns. Among the turns, there are 3627 overlaps and 1069 silences over all the videos.

Our next step is to annotate the interruptions with their type, be cooperative or competitive.

To identify the interruption, we use alignment annotations (corresponding to the start frame and end frame for each word in the utterance) and extract all interruptions. Then, each interruption will be manually classified into three classes: cooperative interruption, competitive interruption and backchannel or others. We note the timestamp of the beginning point for each interruption, with its accomplishment (success/failure) and its type (cooperative/competitive) regarding the speech content.

TRP will also be annotated manually based on the transcription with Part-of-Speech annotation from Stanford Log-Linear POS Tagger [37], with audio and video data as support. Lastly, acoustic features will be extracted by Opensmile [36].

4 Conclusion

The objective of our research is to improve the behaviour generation of ECA. Having noticed that interruptions are important during natural interactions, we propose to study the phenomena involved in interactions before modelling them.

This will allow us to provide the ECA with the capacity to interrupt, but also, to correctly react to interruptions.

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] F. Grosjean,(1996) "Using prosody to predict the end of sentences in english and french : Normal and brain-damaged subjects," Language and cognitive processes, vol. 11, no. 1-2, pp. 107–134

[2] 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.(2009), "Universals and cultural variation in turn-taking in conversation," Proceedings of the National Academy of Sciences, vol. 106, no. 26, pp. 10587–10592

[3] D.Maltz, & R.Borker (1983). A cultural approach to male–female miscommunication. In J. Gumperz (Ed.), Language and Social Identity (Studies in Interactional Sociolinguistics, pp. 196-216). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511620836.013

[4] H.Sacks, E.A.Schegloff, and G.Jefferson (1974). A simplest systematics for the organization of turn-taking for conversation. Language 50, 696–735. doi:10.1353/lan.1974.0010

[5] J. Culpeper, D. Bousfield, and A. Wichmann,(2003) "Impoliteness revisited : With special reference to dynamic and prosodic aspects," Journal of pragmatics, vol. 35, no. 10-11, pp. 1545–1579

[6] D.H.Zimmerman & C.West (1975). Sex, roles, interruptions, and silences in conversations. In B. Thorne & N. Henley (Eds.), Language and sex: Difference and dominance(pp. 105-129).

[7] J. Goldberg.(1990) "Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power- and rapport oriented acts", Journal of Pragmatics, 14, 883-903

[8] HZ.Li (2001) Cooperative and Intrusive Interruptions in Inter- and Intracultural Dyadic Discourse. Journal of Language and Social Psychology. 20(3):259-284. doi:10.1177/0261927X01020003001

[9] BEATTIE, GEOFFREY W..(1981) "Interruption in conversational interaction, and its relation to the sex and status of the interactants" , vol. 19, no. 1-2, pp. 15-36. https://doi.org/10.1515/ling.1981.19.1-2.15

[10] C.Busso, M.Bulut,CC. Lee.(2008) IEMOCAP: interactive emotional dyadic motion capture database. Lang Resources & Evaluation 42, 335. https://doi.org/10.1007/s10579-008-9076-6

[11] S. Tulika, P. Aditya. (2020). "Towards Emotion-aided Multi-modal Dialogue Act Classification" DOI: 10.18653/v1/2020.acl-main.402

[12] Truong, Khiet. (2013). Classification of cooperative and competitive overlaps in speech using cues from the context,overlapper, and overlappee. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 1404-1408.

[13] G. W. Beattie,(1981) "Interruption in conversational interaction, and its relation to the sex and status of the interactants" Linguistics, vol. 19, no. 1–2, pp. 15–36.

[14] M. Heldner, J. Edlund,(2010) Pauses, gaps and overlaps in conversations, Journal of Phonetics,Volume 38, Issue 4,2010,Pages 555-568,https://doi.org/10.1016/j.wocn.2010.08.002.

[15] A.J. Aubrey, D. Marshall, P.L. Rosin, J. Vandeventer, D.W. Cunningham, C. Wallraven,(2013) "Cardiff Conversation Database (CCDb): A Database of Natural Dyadic Conversations", V & L Net Workshop on Language for Vision.

[16] L. Terissi, G. Sad, M. Cerda, S. Ouni, R. Galvez, et al.. (2018)A French-Spanish Multimodal Speech Communication Corpus Incorporating Acoustic Data, Facial, Hands and Arms Gestures Information. Interspeech 2018 - 19th Annual Conference of the International Speech Communication Association, hal-01862585f

[17] J. Hough, Y. Tian, L. Ruiter, S. Betz, D. Schlangen, and J. Ginzburg. (2016). Duel: A multilingual multimodal dialogue corpus for disfluency, exclamations and laughter. In the 10th edition of the Language Resources and Evaluation Conference.

[18] A. Metallinou, C.C. Lee, C. Busso, S. Carnicke, and S.S. Narayanan, (2010) "The USC CreativeIT Database: A Multimodal Database of Theatrical Improvisation," Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality

[19] O. Celiktutan, E. Skordos and H. Gunes, (2017)Multimodal Human-Human-Robot Interactions (MHHRI) Dataset for Studying Personality and Engagement, IEEE Transactions on Affective Computing - DOI 10.1109/TAFFC.2017.2737019.

[20] T. Baltrušaitis, A. Zadeh, Y. C. Lim, and L.P. Morency, (2018) "OpenFace 2.0: Facial behaviour Analysis Toolkit", IEEE International Conference on Automatic Face and Gesture Recognition

[21] C.-C. Lee, S. Lee, and S. Narayanan, (2008)"An analysis of multimodal cues of interruption in dyadic spoken interactions," in Interspeech, Brisbane, Australia

[22] L.-C. Liang,(2001) "Visualizing spoken discourse: prosodic form and discourse function of interruptions," in Second SIGdial Workshop on Discourse and Dialog

[23] F. Yang and P. Heeman, (2007) "Avoiding and resolving initiative conflicts in dialog," in NAACL HLT, Rochester, NY.

[24] E. Drake, S. Schaeffler, and M. Corley (2014). "Articulatory effects of prediction during comprehension: an ultrasound tongue imaging approach," in Proceedings of the 10th International Seminar on Speech Production, Cologne.

[25] S. Schaeffler, J.M. Scobbie, and F. Schaeffler (2014). "Measuring reaction times: vocalisation vs. articulation," in Proceedings of the 10th International Seminar on Speech Production, Cologne.

[26] D. MacNeill,(1992) Hand and Minds: What Gestures Reveal about Thoughts. Chicago, IL: University of Chicago Press.

[27] D. Haylan, (2005)"Challenges ahead. head movements and other social acts in conversation," AISB

[28] C. Lee and S. Narayanan, (2010)"Predicting interruptions in dyadic spoken interactions," 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, pp. 5250-5253, doi: 10.1109/ICASSP.2010.5494991.

[29] A. Gravano, J. Hirschberg,(2012) "A Corpus-Based Study of Interruptions in Spoken Dialogue". Interspeech-2012

[30] C. Adam, J. Švec, and L. Šmídl.(2018) "Learning to interrupt the user at the right time in incremental dialogue systems." International Conference on Text, Speech, and Dialogue. Springer, Cham.

[31] H. Constantin de Chanay and C. Kerbrat-Orecchioni, (2010) "Les interruptions dans les débats médiatiques : une stratégie interactionnelle," Pratiques. Linguistique, littérature, didactique, pp. 83–104

[32] Hara, Kohei & Inoue, Koji & Takanashi, Katsuya & Kawahara, Tatsuya. (2019). Turn-Taking Prediction Based on Detection of Transition Relevance Place. 4170-4174. 10.21437/Interspeech.2019-1537.

[33] B. Wells and S. Macfarlane,(1998) "Prosody as an Interactional Resource : Turn projection and Overlap," Language and Speech, vol. 41, pp. 265–294.

[34] O. Egorow and A. Wendemuth,(2019) "On Emotions as Features for Speech Overlaps Classification," in IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2019.2925795.

[35] Egorow, Olga & Wendemuth, Andreas. (2017). Emotional Features for Speech Overlaps Classification. 2356-2360. 10.21437/Interspeech.2017-87.

[36] F. Eyben, M. Wöllmer, B. Schuller:(2010) "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", Proc. ACM Multimedia (MM), ACM, Florence, Italy,

[37] K. Toutanova and Christopher D. Manning. (2000). "Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger." In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC-2000), pp. 63-70.

[38] C. Kerbrat-Orecchioni,(1999) "Les cultures de la conversation : Le langage en société." .

[39] Crook, N. Smith, C. Cavazza, M. Pulman, S. Moore, R. Boye, Johan. (2010). "Handling user interruptions in an embodied conversational agent."

[40] Crook, Nigel & Field, Debora & Smith, Cameron & Harding, Sue & Pulman, Stephen & Cavazza, Marc & Charlton, Daniel & Moore, Roger & Boye, Johan. (2012). Generating context-sensitive ECA responses to user barge-in interruptions. Journal on Multimodal User Interfaces. 6. 10.1007/s12193-012-0090-z.

[41] Krauss, R. M., Chen, Y., & Chawla, P. (1996). Nonverbal behaviour and nonverbal communication: What do conversational hand gestures tell us? In M. P. Zanna (Ed.), Advances in experimental social psychology, Vol. 28 (p. 389–450). Academic Press. https://doi.org/10.1016/S0065-2601(08)60241-5

[42] Ross Buck, C. Arthur VanLear, Verbal and Nonverbal Communication: Distinguishing Symbolic, Spontaneous, and Pseudo-Spontaneous Nonverbal behaviour, Journal of Communication, Volume 52, Issue 3, September 2002, Pages 522–541, https://doi.org/10.1111/j.1460-2466.2002.tb02560.x

[43] Poggi I., Pelachaud C., de Rosis F., Carofiglio V., De Carolis B. (2005) Greta. A Believable Embodied Conversational Agent. In: Stock O., Zancanaro M. (eds) Multimodal Intelligent Information Presentation. Text, Speech and Language Technology, vol 27. Springer, Dordrecht. https://doi.org/10.1007/1-4020-3051-7_1

[44] R. Smith (2014) "Comparative error analysis of dialog state tracking," in Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)

[45] N. G. Ward and D. DeVault, (2015) "Ten challenges in highly-interactive dialog system," in 2015 AAAI Spring Symposium Series

[46] DeVault, D., Sagae, K., & Traum, D. (2011). Incremental interpretation and prediction of utterance meaning for interactive dialogue. Dialogue & Discourse

[47] Baumann, T., & Schlangen, D. (2011). Predicting the micro-timing of user input for an incremental spoken dialogue system that completes a user's ongoing turn.

[48] Peldszus, A., Buß, O., Baumann, T., & Schlangen, D. (2012). Joint satisfaction of syntactic and pragmatic constraints improves incremental spoken language understanding.

[49] Buschmeier, H., Baumann, T., Dosch, B., Kopp, S., & Schlangen, D. (2012, July). Combining incremental language generation and incremental speech synthesis for adaptive information presentation. In Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (pp. 295-303).

[50] Skantze, G., & Hjalmarsson, A. (2010, September). Towards incremental speech generation in dialogue systems. In Proceedings of the SIGDIAL 2010 Conference (pp. 1-8).

[51] Johansson, M., & Skantze, G. (2015, September). 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 (pp. 305-314).

APPENDIX

Corpus	Language	Session	Duration	Camera	Audio	Turn annotation	Emotion	Posture	Head	FExpression	Transcription
IEMOCAP	English	80	5h	2 on side	mixed	O	O	O	P & R	O	O
CCdB	English	14	1h10m	2 in front	separate	O	O	N	P & R	Openface	O
French-Spanish	French	24	2h20m	1 in front for one speaker	mixed	N	N	O	O	N, possible with openface	N
DUEL	French	30	8h	1 on side	mixed	O	N	Kinect		N	O
	German	10	7h	2 in front	mixed	O	N			N, possible with openface	O
	Chinese	30	5h	1 on side	mixed	O	N			N	O
CreativeIT	English	17	1h20m	1 on side	separate	O	O	O(full body)	O	N	O
MHHRI	English	34	<30m	2 portables	mixed	N	N	O	N	N	N

Table 1: Comparison of different corpora