Anticipation and its applications in human-machine interaction

Stanislav Ondáš, Matúš Pleva
Department of Electronics and Multimedia Communications,
Faculty of Electrical Engineering and Informatics, Technical University of Košice,
stanislav.ondas@tuke.sk, matus.pleva@tuke.sk

Abstract. *Behinds the capability of a person to be an interlocutor of a conversation there lies many human capabilities, many of which are carried out unconsciously and very naturally in childhood. Human-like turn-taking in the human-machine interactions (HMI) can be seen as a critical issue to achieve natural conversational interaction. The production of the listener response starts before the speaker turn is finished, which means, that listener is often able to anticipate the remaining content of unfolded speaker turn. The ability to anticipate can be identified as a very important human capability, which supports rapid turn-taking. This anticipation process, which occurs during listening relates to sentence comprehension and turn-taking. The proposed paper study this phenomenon on the small corpus of Slovak interviews, where the attention is focused on overlapping segments and discussed possible applications, where anticipatory behavior on the machine side can bring benefits.*

1 Introduction

Nowadays spoken communication between human and machine has become obvious, what relates to the new types of devices, which start to be a part of our everyday life. Good examples are Amazon Echo, Google Home, TV with voice control, Voice search applications, social robots.

We can classify that form of communication mostly as simple dialogues, what means question-answer scenarios and task-oriented domain-specific dialogue.

Unlike mentioned scenarios, in case of human-human spoken interaction, the situation is significantly more complex. Here, information exchange is often performed very quickly, a lot of information is omitted, but supplemented by the listener based on the common background and history (short-term, long-term, topic-related, speaker-related, situation-related). Behind a capability of a person to be an interlocutor of a conversation lies a lot of important human capabilities, many of which are carried out unconsciously and very naturally. People are easily able to track the content, to detect a speech act (dialogue act) behind a speaker's turn, to perform effective and rapid turn-taking, to provide feedback in a role of listener, or incrementally construct their turns.

In the proposed work, we focus on the three related aspects of the spoken communication - turn-taking, sentence comprehension and anticipation.

"A turn is the time when a speaker is talking, and turn-taking is the skill of knowing when to start and finish a turn in a conversation. It is an important organizational tool in spoken discourse." [1] Turn-taking can be described as a process in which one dialogue participant talks, then stops and gives the floor to another participant. It is a human skill, which we learn without any effort in childhood.

Stivers et al. in [7] observed, that human-human conversations are characterized by rapid turn-taking often with a minimal gap lower than 200 msec. Several other studies (e.g. [8], [9]) indicate that utterance production can take more than 600msec (see [10], [12]). It indicates, that utterance productions usually start during listening. This finding is supported by measurements of EEG signals. Magyari et al. in [2] observed changes in EEG which relates to the fact, that human brain is able to estimate turn duration. This estimation is based on anticipating the way the turn would be completed. They founded a neuronal correlate of turn-end anticipation and a beta frequency desynchronization as early as 1250 msec, before the end of the turn. They suggest that anticipation of the speaker utterance leads to accurately timed transitions in everyday conversations. The ability to anticipate the content of the speaker turn and turn end was researched and confirmed in several papers (see e.g. [2], [3], [4], [6]).

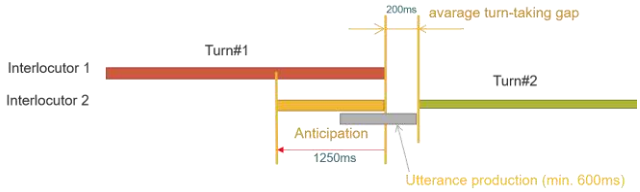

Figure 1 Turn-taking timing

Anticipation relates to the sentence comprehension. It can be inferred, that instead of the word sequence, the meaning is anticipated. We can identify, that in the moment, when a person is able to anticipate remaining part of the speaker turn, there is partial or complete comprehension, which can be signalized using one of the following behavior: providing a feedback (backchannel signals) or attempt to take the floor.

The attempt to take the floor can result into the overlapping speech. Overlapping speech is a segment of the conversation, where both interlocutors speak simultaneously. The speaker tries to finish his turn, but the listener starts his own turn. Several studies quantified a significant occurrence of overlaps in dyadic human-human spoken interactions (e.g 18.9% in [26]).

We supposed that overlaps are the best place to observe anticipatory behavior of the listener.

The reason, why we decided to study the relation between rapid turn-taking, comprehension and anticipation is the lack of fluentness in case of human-machine spoken interaction. We can observe that machines are still not enough skilled in turn-taking.

Rapid turn-taking in HMI is especially important, when we consider the fact, that in case of humanoid robots, social and family robots people tend to expect more natural

dialogue interaction, often called "conversation" due to their human-like embodiment.

Human-like turn-taking in the human-machine interactions can be seen as a critical issue to achieve natural conversational interaction in HMI [11 – 12]. A great view inside this area can be found in [10].

Together with observation of anticipatory behavior in human-human interactions, several questions regarding human-machine spoken interaction arise. E.g.:

Would machines let their "predictable" turns unfinished, when they observe a comprehension on the side of human listener? Can unfinished turns increase a natural character of human-machine interactions?

Or:

Would machines be able to interrupt human speaker turn and if yes, in which situations?

How can be anticipation integrated on the side of machine?

How can machines catch the moment when the listener has enough information to comprehend?

And finally:

Could be helpful if the machine would be able to interrupt the human interlocutor?

To answer proposed questions, we decided to collect human-human interactions to analyze turn-taking, overlaps and anticipatory behavior.

The paper is organized as follows: The second section deals with anticipation in human-human spoken interaction, description of the prepared corpus and results of its analysis. Section 3 provides discussion of applications, where anticipation can bring benefits.

2 Anticipation in human-human spoken interactions

Anticipation play a very important role in the human-human as well as human-machine spoken interaction, because it influences the speed of response in conversation (see [2]) and it is a critical ability to enable "rapid" turn-taking. Anticipation allows the speaker to interpret partial utterances. Sagae et al highlights the importance of anticipation, when they conclude in [19] „*To achieve more flexible turn-taking with human users, for whom turn-taking and feedback at the sub-utterance level is natural, the system needs the ability to start interpretation of user utterances before they are completed.*" and „*it also includes an utterance completion capability, where a virtual human can make a strategic decision to display its understanding of an unfinished user utterance by completing the utterances itself*".

Anticipation is routinely used by human interlocutors in dyadic interactions. Expect the exact results that indicate anticipatory behavior, according our opinion, there can be identified also other indicators of interlocuters anticipation:

- Unfinished turns: We believe that in case of unfinished turns speaker considers that the listener can anticipate remaining part of his turn.
- Overlaps and interruptions: We believe that the listener can try to take floor, when he has come to the comprehension before the end of the speaker turn.

Interruptions differ from overlaps in the timing. Listener usually use a small pause in the speaker turn to interrupt the speaker and to take a floor without causing an overlap. Such pauses are usually marked as "*Transition relevant places*" (TRP) as defined by Sacks et al. in [27]. They can also indicate the place, where the anticipation core [20] or "moment of the maximum understanding" [19] is located. We cannot claim that each TRP is the place, where a listener is able to anticipate. It could be interesting to research, whether places, where anticipation core is located can be marked as TRP.

Unfinished turns and occurrence of overlapping segments can have several other reasons expect comprehension based on anticipation. Therefore, the careful analysis needs to be done.

The mutual understanding or comprehension on the listener side can be indicated by backchannel signals, which are usually produced by human listeners. Backchannel signals was defined by Yngve [14] as an acoustic and visual signals provided during the speaker's turn. Allwood et al. and Poggi in [15, 16], described meaning of acoustic and visual feedback that they provide information about the basic communicative functions, as perception, attention, interest, understanding, attitude (e.g., belief, liking) and acceptance towards what the speaker is saying. Bevacqua et al. in [17] defined some associations between the listener's communicative functions and a set of backchannel signals. They performed an experiment with the 3D Embodied Agent Greta [18], which confirm defined associations. In the described experiment it has been shown, that there exist the association between understanding and following multimodal backchannel signals: *raise eyebrows+"ooh"*, *head nod+"ooh"*, *head nod+"really"*, *head nod+"yeah"* and *head nod*.

2.1 Corpus

Corpus of the investigative interviews of the TV program "Na rovinu" and episodes of TV discussions "Pod lampou" were selected for the analysis of turn-taking mechanisms.

The overall length of the analyzed corpus was approx. 8 hours. There were 9 speakers (8 males, 1 female), two of them play a role of the moderator (male).

Recordings processing have two main levels. In the first level, recordings were transcribed. The first transcriptions were generated by our automatic transcription system [25]. Then, transcriptions were corrected manually. Transcriptions have a form of .trs files, which are generated by Transcriber tool. At the second level, turns and overlaps of both interlocutors were marked in Anvil annotation tool, what enables to analyze turn-taking.

In the first step of the analysis we focused mainly on overlapping segments. To classify different types of overlaps, we designed 13 categories according intent behind the overlaps (see Fig. 2.)

Two basic categories – concurrent and cooperative overlaps are further divided according their function.

Concurrent/competitive overlaps represent overlaps, where we can identify an attempt to grab the floor. On the other side *cooperative/non-competitive* category mark overplaps, where listener's goal is to assist the speaker to continue his turn.

Overlap intention category	Overlap type
stop speaker	concurent
obtain confirmation	cooperative
obtain clarification	cooperative
complete information/provide additional info	cooperative
support speaker	cooperative
express noninterest/topic shift	concurent
try to take floor	concurent
floor leaving misunderstanding	concurent
obtain agreement	cooperative
express disagreement	-
express involvement/express aggrement	cooperative
additional question	concurent

Figure 2 Overlaps categories

Overlap intention category	moderator	respondent
	%	
stop speaker	6.19	2.29
obtain confirmation	3.33	0
obtain clarification	6.42	8.04
complete information/provide additional info	20.23	11.49
support speaker	32.38	3.44
express noninterest/topic shift	0	0
Unclear	4.28	2.29
try to take floor	2.85	11.49
floor leaving misunderstanding	0	0
obtain agreement	1.66	0
express disagreement	0.47	2.29
express involvement/express aggrement	3.8	57.47
additional question	18.33	1.15

Figure 3 Overlap categories distribution

2.2 Results

Table 1. shows statistics related to turn-taking obtained from analyzed corpus.

Table 1. Turn-taking statistics

	1. Speaker (moderator)	2. Speaker (respondent)	Total
Num. of turns	741 (45%)	905 (55%)	1646
Num. of overlaps	483 (84.3%)	90 (15.7%)	573 (34.8%)

Total number of both speakers performed turns was 1646. In 34.8% of cases, turns were taken by the listener before the end of the speaker turn, what causes overlapping speech.

More than 84% of overlaps were caused by the moderator, who jumped into respondent speech. Respondent jumped into moderator turns approx. in 16% of all cases. The explanation may be that the moderator has a responsibility for interaction, and he need to maintain topic, topic shifts and timing (duration) of the interaction.

Table in Fig.3 shows the distribution of intentions behind the overlaps, which we detected manually.

According obtained results, we can conclude significant differences between moderator and respondent. On the side of the moderator, the most numerous categories were: supporting the speaker (approx. 32%), request for completing information or providing additional information (around 20%) and asking the additional question (18%).

On the side of the respondent, the most numerous categories imagine overlaps, which occur when a listener expressed involvement and agreement (more than 57%).

Obtained results show that roles, which play interlocutors affects the number of overlaps and also intentions behind attempts to take the floor.

Next issue is that designed categories contain several items, which can be identified as backchannel signals (express involvement/agreement/disagreement, express noninterest). As was concluded above, backchannel signals indicate rather partial understanding, than anticipatory behavior. However, information about this dialogue parts is important too, to analyze backchannel mechanisms in different scenarios and roles, which interlocutor plays.

Designed categories of overlaps correspond wih speech/dialogue acts, which are conveyed through utterances that interrupt the speaker turn. *Confirmation request*, *Clarification request*, *Agreement*, *Disagreement* are typical dialogue acts (see e.g. [24]). On the other hand, categories like "Stop the speaker", "Support speaker", "Move to another topic", "Try to take floor" can be seen as a turn-management commands, which help to manage speaker changing.

3 Anticipation and its practical application

The main reason, why anticipation needs to be considered is to support rapid turn-taking, which enables smooth and fluent human-machine spoken interaction. But anticipation can enable also other human-like capabilities in HMI.

In case of application anticipation in human-machine dialogue interactions, there exists only few works that deals with this phenomenon. We can mention the work of Dominey et al., described in [5], which focuses on next turn anticipation, based on dialogue history, but in this case the anticipation is not focused inside the turn. Sagae et al. focus their work on interpretation of partial utterances, where they are using words prediction – anticipation. They realize an idea of the incremental user utterance processing, which enables to increase speed of turn-taking (switching the speaker).

The anticipatory behavior on the machine side means the ability to find an enough reliable hypothesis of unfold utterance just uttered by human interlocutor. Machines that can communicate with user through spoken language implements a human machine communication chain. Modules of this chain have anticipatory potential, because they use resources with related data, as are recognition network, language models or dialogue model.

3.1 Applications

While, the anticipatory behavior can be considered as non-important for simple task-oriented dialogue systems, it can bring more human-like character of the interaction and advantages in case of systems for multi-party conversations, for human-machine collaboration scenarios or for crisis scenarios, where can be useful or necessary for the machine to be able to rudely interrupt a human speaker(s), to be able to take a floor and to propose ideas or solutions.

3.1.1 Machine in a role of a moderator

Machine in the role of discussion moderator can be the next example of application, where anticipatory behavior can play important role. Emotionally colored multi-party spoken interactions as are e.g. political discussions often require a lot of effort on the side of moderator to lead the interaction and to manage turn-taking for all speakers. Moreover, he often needs to enforce good behavior or compliance with specified time.

3.1.2 Machine in a task of simultaneous interpretation

Anticipating in simultaneous interpreting simply means that interpreters say a word or a group of words before the speaker actually says them [13]. It means that interpreters, familiar with the domain and content, can interpret beforehand thank to the anticipation. Anticipation is a key competence that interpreters need to learn before they can become professionals [13], [21].

Nowadays, machine translation is a common application. There also exist applications, which perform interpretation, but they interpret whole sentences after their pronunciation. But, if we imagine machine as the simultaneous interpreter, it must be able to anticipate, to produce fluent simultaneous interpretations without meaningless gaps. In that case, anticipation can enhance perceived quality and clarity of the translation.

Anticipatory function can help also human interpreters in the way, that it can suggest them hypothesis about next words in the speaker utterance. Architecture of such machine-supported simultaneous interpretation system is sketched on the Fig. 4.

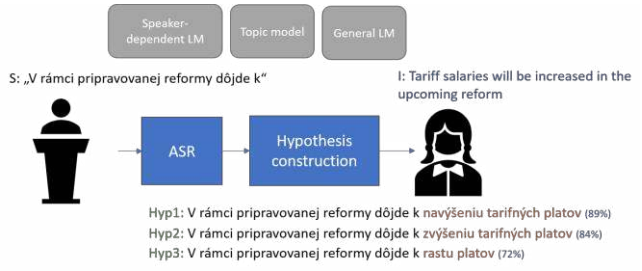

Figure 4 Machine-supported simultaneous interpretation

One of the scenarios of machine-supported simultaneous interpretation is that the system will listen to the speaker and try to provide early predictions of the next words according just recognized partial utterance. Speaker speech will be continually recognized by Automatic Speech Recognition (ASR) module, where recognition hypothesis will be provided as often as possible. Each recognition hypothesis will serve as the input into Hypothesis construction module, in which, next words will be predicted according general, speaker-dependent and topic models. Statistical n-gram language models/DNN networks that are used in ASR module can also serve for final hypothesis generation.

The same scenario can be used also on the side of the interpreter, where the system can suggest next words of his interpretation.

Especially interesting can be an automatic transfer of emotions, which can help interpreter to choose the most appropriate words, which consider emotional coloring of interpreted speech. Work proposed in [23] can be used for desired emotion transfer.

Conclusions

The aim of the paper is to stimulate discussion on the use of anticipation and anticipatory behavior in the HMI and in the practical applications. Certainly, anticipation lies behind the smooth human-human interactions and rapid turn-taking and we believe that it can significantly accelerate human-machine spoken interaction and to support human-like character of the HMI.

We believe that machine-supported anticipation can decrease cognitive load in several applications, e.g. in simultaneous interpretation, where such a system can suggest hypothesis for the interpreter in advance or to support preparation of the interpretation result.

We realize that obtained results are relevant only to the interview scenario and the distribution of analyzed overlaps category will change according roles, relationship, emotions of interlocutors and according discussed topics. The first challenge of our future work will be collecting and analysing of dialogue interactions in other scenarios.

Acknowledgment

The research presented in this paper was supported by the Slovak Research and Development Agency projects

APVV SK-TW-2017-0005, APVV-15-0731, Ministry of Education, Science, Research and Sport of the Slovak Republic under the research project VEGA 1/0511/17 and by Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 009TUKE-4/2019.

References

[1] https://www.teachingenglish.org.uk/article/turn-taking
[2] L. Magyari, M.C. Bastiaansen, J.P. de Ruiter, and S.C. Levinson, "Early Anticipation Lies behind the Speed of Response in Conversation", Journal of Cognitive Neuroscience, pp. 2530-2539, 2014.
[3] R.S. Gisladottir, S. Bögels, and S.C. Levinson, "Oscillatory Brain Responses Reflect Anticipation during Comprehension of Speech Acts in Spoken Dialog", Front. Hum. Neurosci, 2018.
[4] A.J. Liddicoat, "The projectability of turn constructional units and the role of prediction in listening", Discourse Studies, Vol.6, No.4, pp. 449-469, 2004.
[5] P.F. Dominey, G. Metta, F. Nori, and L. Natale, "Anticipation and initiative in human-humanoid interaction", Humanoids 2008 - 8th IEEE-RAS International Conference on Humanoid Robots, Daejeon, pp. 693-699, 2008.
[6] M. Paľová, E. Kiktová "Prosodic anticipatory clues and reference activation in simultaneous interpretation", in XLinguae 12(1XL), pp. 13-22. 2019.
[7] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, et al. (2009). Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences, U.S.A., 106, 10587–10592.
[8] P. Indefrey, W. J. M. Levelt, "The spatial and temporal signatures of word production components". Cognition, 92, 101–144, 2004
[9] T. T. Schnurr, A., Costa, A. Caramazza, "Planning at the phonological level during sentence production.", Journal of Psycholinguistics Research, 35, 189–213, 2006.
[10] J. Holler, K, H. Kendrick, M. Casillas, S. C. Levinson, eds. (2016). Turn-Taking in Human Communicative Interaction. Lausanne: Frontiers Media. doi: 10.3389/978-2-88919-825-2, 2016
[11] K. R. Thórisson, "Natural turn-raking needs no manual: computational theory and model, from perception to action," in Multimodality in Language and Speech Systems, eds B. Granström, D. House, and I. Karlsson (Netherlands: Springer), 173–207, 2002
[12] M. Heldner, and J. Edlund, "Pauses, gaps and overlaps in conversations." J.Phon. 38, 555–568, 2010
[13] Anticipation in simultaneous interpreting, https://www.languageconnections.com/, March 2018.
[14] V. Yngve, "On getting a word in edgewise", in Papers from the Sixth Regional Meeting of the Chicago Linguistic Society, pp. 567–577, 1970.
[15] J. Allwood, J. Nivre, and E. Ahlsn, "On the semantics and pragmatics of linguistic feedback", Semantics Vol.9, No.1, 1993.
[16] I. Poggi, "Mind, hands, face and body. A goal and belief view of multimodal communication", Weidler, Berlin, 2007.
[17] E. Bevacqua, S. Pammi, S.J. Hyniewska, M. Schröder, and C. Pelachaud, "Multimodal Backchannels for Embodied Conversational Agents", in: Intelligent Virtual Agents, IVA 2010. Lecture Notes in Computer Science, Vol.6356. Springer, Berlin, Heidelberg, 2010.
[18] R. Niewiadomski, E. Bevacqua, M. Mancini, and C. Pelachaud, "Greta: an interactive expressive eca system", in: AAMAS 2009 - Autonomous Agents and MultiAgent Systems, Budapest, Hungary, 2009.
[19] K. Sagae, D. DeVault, and D.R. Traum, "Interpretation of partial utterances in virtual human dialogue systems", Proc. of the NAACL HLT 2010. Association for Computational Linguistics, Stroudsburg, PA, USA, pp.33-36, 2010.
[20] E. Kiktova, J. Zimmermann, "Detection of Anticipation Nucleus using HMM and Fuzzy based Approaches", (in review process) DISA 2018, August, Košice, Slovakia
[21] K. G. Seeber, "Intonation and Anticipation in Simultaneous Interpreting", Cahiers de Linguistique Française, vol.23, 2001, pp. 61-97, ISSN 1661-3171, 2001
[22] S. Ondáš, J. Juhár, M. Pleva, et. al.: "Speech technologies for advanced applications in service robotics", in Acta Polytechnica Hungarica. Vol. 10, no. 5 (2013), p. 45-61., ISSN 1785-8860
[23] M. Mikula and K. Machová, "Combined approach for sentiment analysis in Slovak using a dictionary annotated by particle swarm optimization.", in: Acta Electrotechnica et Informatica, Vol. 18, No. 2, 2018, 27–34, DOI: 10.15546/aeei-2018-0013
[24] S. Ondáš and J. Juhár, "Distance-based dialog acts labeling," 2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Gyor, 2015, pp. 99-103.
[25] Lojka M., Viszlay, P., Stas, J., Hladek, D., Juhar, J.: Slo-vak Broadcast News Speech Recognition and TranscriptionSystem. International Conference on Network-Based Infor-mation Systems. In: Barolli L., Kryvinska N., Enokido T.,Takizawa M. (eds) Advances in Network-Based Informa-tion Systems. NBiS 2018. Lecture Notes on Data Engineer-ing and Communications Technologies - LNDECT, vol 22.Springer, Cham, pp. 385–394, 2019.
[26] I. Siegert, R. Bock, A. Wendemuth, B. Vlasenko and K. Ohnemus, "Overlapping speech, utterance duration and affective content in HHI and HCI - An comparison," 2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), Gyor, 2015, pp. 83-88.
[27] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," Language, vol. 50, no. 4, pp. 696–735, Dec. 1974.