Overlapping Speech, Utterance Duration and Affective Content in HHI and HCI – an Comparison

Ingo Siegert, Ronald Böck, Andreas Wendemuth
Cognitive Systems Group
Otto von Guericke University Magdeburg, Germany
{firstname.lastname}@ovgu.de

Bogdan Vlasenko
Idiap Research Institute
Martigny, Switzerland
bogdan.vlasenko@idiap.ch

Kerstin Ohnemus
davero Dialog Gruppe
91058 Erlangen, Germany
kerstin.ohnemus@davero.de

***Abstract*—In human conversation, turn-taking is a critical issue. Especially if only the speech channel is available (e.g. telephone), correct timing as well as affective and verbal signals are required. In cases of failure, overlapping speech may occur which is in the focus of this paper. We investigate the davero corpus a large naturalistic spoken corpus of real callcenter telephone conversations and compare our findings to results on the well-known SmartKom corpus consisting of human-computer interaction. We first show that overlapping speech occurs in different types of situational settings – extending the well-known categories cooperative and competitive overlaps –, all of which are frequent enough to be analyzed. Furthermore, we present connections between the occurrence of overlapping speech and the length of the previous utterance, and show that overlapping speech occurs at dialog instances where certain affective states are changing. Our results allow the prediction of forthcoming threat of overlapping speech, and hence preventive measures, especially in professional environments like call-centers with human or automatic agents.**

I. Introduction

Human communication consists of several information layers, of which the factual layer is the most important. But beyond the pure textual information other relevant information such as affective state, self-revelation, and appeal are transmitted [1]. These different pieces of information are provided to support human conversations and to increase the likelihood of a fluent conversation.

One important requirement for a fluent and successful conversation is an efficient turn-taking, which has to be organized by specific “underlying mechanisms”, such as intonation, semantic cues, facial expressions, eye contact, breathing, and gestures [2], [3], [4]. In the organization of turn-taking and to evaluate the conversation, overlapping speech has a major role. Based on the turn-taking model by Sacks et al. [3], conversational partners aim to minimize overlaps and gaps in their conversations. From to this model, overlaps occurs at places of possible turn ends, either as “terminal overlaps” or “simultaneous start”. Thus, overlapping speech is explained as a result of turn-taking principles. This explanation is extended for different situational settings e.g. by [5], [6], where short feedback signals confirming the statement of the current speaker, are seen as “response token” overlaps. Furthermore, several studies also analyses competitive overlaps, in which the conversational partners compete for the turn [7], [8].

Many recent studies analyzed the phonetic structure of overlapping speech and found that fundamental frequency, intensity, speech rate and rhythm are important features characterizing the overlaps as either being cooperative or competitive [9], [8], [10]. Most of these studies concentrate on local analyses investigating the acoustic characteristics next to or directly at the overlap. Only a few studies incorporate for example information on the duration of turns [11]. But the relation of the length of utterances with the situational type of overlap is not analyzed.

Former studies on overlapping speech concentrate to seek an explanation of how overlapping speech works. They do not analyze which consequences lead to an overlap or which consequences the overlap has for the progress of the interaction. The analyses especially disregard the length of utterances where an overlap occurs (consequence lead to an overlap) and which influence the affective state could have. Especially, in [12] it is emphasized that affective states influence the turn-taking behavior. Thus, problems in turn-taking can also be traced back to changes in the affective state.

Furthermore, these studies are conducted only on human-human interaction (HHI). Investigations on human-computer interaction (HCI) do not consider overlapping speech as an informative signal so far. But, to reach the target of a more naturalistic interaction, future systems have to be adaptable to the users’ individual skills, preferences, and current emotional state [13], [14]. Lot of progress have been made in the area of affect detection from speech, facial expression and gesture [15]. For a fully naturalistic HCI, it is necessary to capture as many human abilities as possible. Thus also linguistic features gain considerable importance [16]. In an earlier study, we could show that discourse particles, exchanged among the interaction partners and used to signalize the progress of the dialogue, are also used in naturalistic HCI, although the system was not able to properly react to them [17]. The usage of these cues is influenced by the user’s age and gender [18].

In this paper we will extend our analysis of linguistic cues to overlapping speech and analyze the meaningfulness of overlapping speech regarding the utterance length and the user’s affective state change. We conducted a contrasting study using human-human interaction (HHI) as well as HCI. This will be the first step towards an automatic evaluation of overlapping speech and could help future “Cognitive Infocommunication” systems to understand the human better [14]. Technical systems that use this extended recognition of linguistic cues adapt to their users and thus become his attendant and ultimately his *companion* [13], [19].

Based on these considerations, we investigate the following three research questions in this paper.

Q1 Is overlapping speech occurring frequently enough in our material to be analyzed in a dyadic conversation?

Q2 Is there any connection between overlapping speech and the length of the previous utterance?

Q3 Is overlapping speech occurring at points where the affective state is changing?

The remainder of the paper is structured as follows: In Section II the utilized datasets are shortly described and specific differences are emphasized. Afterwards, in Section III, we describe the preparation of the date in terms of types of overlap and affective annotation. In Section IV the results are presented and discussed. Finally, in Section V, a conclusion of our investigations and an outlook for further research is given.

II. Utilized Datasets

A. Davero Corpus of Telephone-based Conversations

The dataset is described in detail in [20]. It is created within a research project aiming to develop a technical system that supports callcenter employees to respond appropriately to the current affective state of the caller and was recorded in a callcenter collecting real and authentic phone calls in German. The calls embrace various topics, like information talks, data change notifications, and complaints.

In order to allow a complete analysis of the conversation both, agent and caller, were recorded acoustically. To gain realistic and high-quality recordings as well as to avoid disturbing background noise, a separate recording place had been set up. In total, 49 days $*$ 7 hours have been recorded. Since the recorded phone conversations are real customer interactions they had to be anonymized first, blanking out all personal information. Furthermore, the start and end-times of each dialog and overlapping speech segments were marked and each utterance was assigned to its corresponding speaker (agent or caller). To date, this dataset contains 1,600 dialogs with 27,000 individual utterances. The dialogs have an average length of about 5 minutes with a standard deviation of ± 2 minutes.

B. SmartKom multi-modal Corpus

The SmartKom multi-modal corpus contains naturalistic affects within a HCI [21]. The system responses were generated by a Wizard-of-Oz (WOZ) setup. For our evaluations we use German dialogs concerning a technical scenario, recorded in a public environment. The database contains multiple audio channels and two video channels (face and body in profile posture). The primary aim of this corpus was the empirical study of HCI in a number of different tasks. It is structured into several sessions. Each session contains one conversation and is approximately 4.5 minutes long.

This corpus has several annotation levels, of which for our investigation the turn segmentation and an affective annotation based on the acoustic channel is used [22]. The considered set of the SmartKom corpus contains 438 emotionally labeled dialogs with 12,076 utterances in total and 6,079 user utterances. The utterances are labeled in seven broader affective states: neutral, joy, anger, helplessness, pondering, surprise and unidentifiable episodes. Unfortunately, the turn segmentation is not time-aligned with the affective annotation.

III. Preparation of Datasets

A. Analysis of Overlapping Speech

We analyze overlapping speech as an additional pattern of an interaction, as we assume that a valuable contribution to the assessment of interactions is provided. Overlapping speech refers to the case, when both speakers are talking simultaneously.

1) Davero Corpus: By listening to examples of the Davero corpus, four different situations (S) can be identified where overlapping speech occurs:

S1 Short feedback, no interruption of the speaker
S2 Premature turn-taking at the end of the speaker's turn
S3 Simultaneous starting after longer silence
S4 Barge-in, aiming to take the turn over

These situations are based on the descriptions of [11], distinguishing response tokens (S1), terminal overlaps (S2), simultaneous starts (S3) and competitive overlaps (S4). A prototypical illustration is given in Figure 1.

In the first situation (S1), the listener just wants to give a feedback. Lacking of other feedback methods (head nodding, eye gaze), the listener has to give the feedback acoustically. Thus no real turn-taking occurs. The second situation (S2) can be seen as a functional turn-taking. The listener knows that the speaker's turn is due to end, but because of the missing visual feedback the listener starts his turn a bit too early. In this case just the alignment of the turn-taking is incomplete. S3 is similar, both speakers start talking coincidentally after a longer silence due to missing cues. S4 shows an disturbed turn-taking. It describes the case where one speaker barges-in while the other is still speaking to deliberately steal the turn from that other speaker.

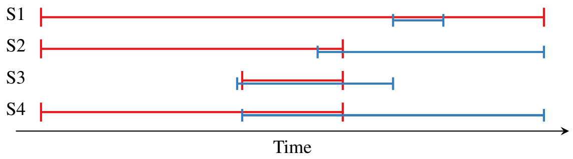

Figure 1. Prototypes of the four different situations of two speakers (denoted as – and –) for overlapping speech in HHI.

To evaluate the overlapping speech according to these descriptions, we employed two labelers with psychological background for the assessment. They could choose between all four situations or describe a situation not covered by the definitions. Of the currently available 27,000 utterances in 1,600 dialogs 5,100 utterances (18.9%, 830 dialogs) are marked to contain overlapping speech.

The final assessment is as follows: S1 has a share of 61.6%, S2 a share of 11.2%, S3 a share of 10.7%, and S4 a share of 16.6%. Furthermore, no additional situation was selected by the annotators. As inter-rater reliability of the crosstalk labelling we calculated a Krippendorff's alpha of 0.63, a substantial reliability according to [23].

2) SmartKom Corpus: The SmartKom Corpus does not have explicitly marked overlapping speech segments. By using the segmentation annotation we could identify two types (T) of overlapping speech (Figure 2):

T1 User interrupts the system.
T2 System interrupts the user.

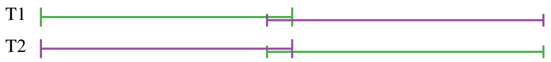

Figure 2. Prototypes of the two different type of overlapping speech between system (–) and user (–) in HCI.

In HCI the system is not seen as an equivalent dialog partner [24], [25]. Therefore, the variety of types is not as big as in HHI. From the 12,640 dialog acts within the SmartKom corpus, we have 817 overlapping speech samples for T1 and 672 samples for T2.

B. Evaluation of the Affective States

We analyzed the affective states in both corpora based on the Geneva Emotion Wheel by K. Scherer [26]. This is an empirically tested instrument for the assessment of affective states including 16 "emotional families", which are arranged on a circle along the axes dominance and valence.

1) Davero Corpus: To conduct the affective assessment, we first employed a few annotators to manually segment the recordings into single dialogs including the speaker turns. We asked four annotators, all of them with psychological background, to assess the affective content of the single utterances. We conducted several training rounds to make them familiar with the used annotation scheme and the affective assessment of acoustic data [27]. To support the annotation process, the program ikannotate was used [28], [29]. This tool supports the annotators by employing a three-step annotation process:

1. The annotator decides if the dominance is high or low.
2. The annotator decides for positive or negative valence.
3. The resulting quadrant of the wheel is displayed with the containing emotion families. The annotator selects one family among them to indicate the perceived emotion.

For the present investigation, we only consider the labels from step 1 and step 2, as we are only interested in a general affective change. For the inter-rater reliability of the affective annotation we calculated a Krippendorff's alpha of 0.20 for dominance and of 0.35 for valence. Although these numbers seem to be quite low compared to other reliability values known from content analysis, they are in line with results from other research groups on affective analyses [27].Considering the annotation results we observe a nearly balanced distribution among the utterances. High dominance has a share of 59.0% and low dominance of 41.0%, positive valence has a share of 53.5% and negative valence of 46.5%.

2) SmartKom Corpus: The SmartKom Corpus already has an affective annotation [30]. Unfortunately, this annotation is not on the same time-scale as the dialog act segments. Thus, we have to perform an alignment of both annotation levels, by using the individual timing information of both annotation levels. Unfortunately, the corpus authors only measured the annotation correctness by comparing the results of different annotation rounds rather than calculating an inter-rater agreement measure like Krippendorff's alpha or Fleiss' kappa. Their calculated correctness is 45.52% [31]

Furthermore, in contrast to the Davero corpus, the affective annotation in SmartKom is based on emotional categories. Thus, we first have to deploy a mapping of the categories used in SmartKom to our utilized dimensional categories of dominance and valence. To conduct this mapping, we rely on the Geneva Emotion Wheel [26], as it is also used for the annotation of the Davero corpus. In analogy to similar mappings [32], the assignment of the emotional categories to the valence-dominance space is given in Table I. In contrast to the Davero corpus, neutral is used as a category in SmartKom.

Table I. *Mapping of SmartKom's emotional categories to dominance and valence.*

category	dominance	valence
neutral	0	0
joy	+1	+1
anger	+1	-1
helplessness	-1	-1
pondering	-1	+1
surprise	-1	+1
unidentifiable	0	0

In total we have 14,298 affective segments. Most of these segments (59.7%) are neutral. The distribution on the dominance dimension is 30.7% low and 69.3% high dominance. Positive valence has a share of 39.0% positive and negative valence a share of 61%. Thus, the emotional content within this corpus is shifted towards negative valence and high dominance.

IV. Results

A. Q1: Occurrence of Overlapping Speech

To answer the first question, we calculated the ratio of overlapping speech segments and number of utterances. For the Davero corpus we have 27,000 utterances and 5,100 of them contain overlapping speech. Thus, we have a share of 18.9% for overlapping speech segments. If we consider the dialog level, we have 1,600 dialogs in total of which 830 dialogs contain overlapping speech segments. This results in a share of 51.9%.

The German part of the SmartKom Corpus has a total number of 12,076 utterances with 1,489 occurrences of overlapping speech. This results in a share of 12.3% overlapping speech utterances for this HCI. As we only take into account the overlapping speech, we have 6,347 user utterances and 817 utterances contain overlapping speech.

Thus, we can conclude that overlapping speech is occurring frequently and the first question: "Is overlapping speech occurring frequently enough to be analyzed in a dyadic conversation?" is approved.

B. Q2: Overlapping Speech and Utterance Lengths

This investigation is triggered by the assumption that overlapping speech is occurring because the actual speaker is talking too long and the listener wants to get the turn. To investigate this assumption, we calculated the mean length of the utterance where the overlapping speech occurs ($utt_{overlap}$) in relation to all other utterances (utt_{remain}) of this speaker within a dialog. Afterwards, we averaged over all dialogs and calculated the difference between both averaged mean lengths:

$$\Delta\mathrm{len} = \overline{len}_{utt_{overlap}} - \overline{len}_{utt_{remain}} \qquad (1)$$

This calculation is performed for each of the previous identified different situations separately and averaged afterwards ($\overline{\Delta\mathrm{len}}$). Additionally, we used the non-parametric Mann-Whitney-U-Test, to test the significance of the difference within the utterance lengths. The star denotes the significance level: ** $p < 0.001$.

1) Davero Corpus: From Figure 3 it can be seen, that in two situations, the length of the utterance with overlapping speech is different from other utterances of the same speaker. For S1, the $\overline{len}_{utt_{overlap}}$ is significantly longer than for other utterances. In this situation one speaker gives statements that are just confirmed by the listener without interrupting the speaker. Thus, the speaker can continue his turn. The presence of this type of overlapping speech does not indicate a change in the progress of the dialog. The same statement can be made for S2, where the $\overline{len}_{utt_{overlap}}$ is significantly shorter than for other utterances. The overlapping speech in both situations is just occurring because only the acoustic channel can be used to negotiate the turn-taking. Thus, the length of an utterance together with the information of an occurring speech overlap cannot be used as an indicator for dysfunctional conversation.

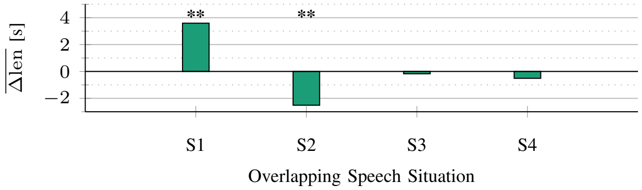

Figure 3. Difference of the utterance length for the four defined overlapping speech situations and the average utterance length in the Davero corpus, stars indicate a significant difference.

2) SmartKom Corpus: From Figure 4 it can be seen that for type T1 the system utterance with the overlapping speech segment ($\overline{len}_{utt_{overlap}}$) is significantly longer than the other system utterances. Thus, it can be assumed that for a positive interaction outcome and fluent conversation the system prompts shouldn't be too long.

For the second type, where the system interrupts the users, the users' utterance length containing overlapping speech is not significantly different from other user's utterances. Therefore, we assume that these interruptions of the system are caused by operator errors of the WOZ-system.

Regarding our second research question, we can state that there is a significant correlation between overlapping speech and the length of the previous utterance in both HHI and HCI.

C. Q3: Overlapping Speech and Affective Changes

To investigate the affective change at the point where overlapping speech occurs, we take into account the observed affective states in two preceding utterances and compare it to the observed affective states in the two succeeding utterances. We distinguish between high (+1) and low (-1) dominance and positive (+1) and negative (-1) valence. Utterances that do not have an affective label or are labeled as neutral are assigned a 0. Thus, we can calculate the difference between the affective states of the preceding utterances and the succeeding utterances for an overlapping speech segment:

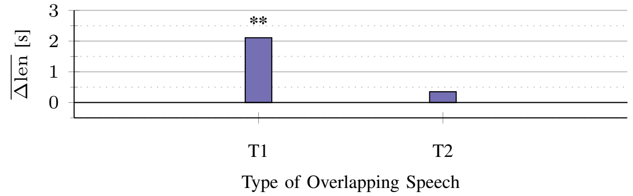

Figure 4. Difference of the utterance length for the two types of overlapping and the average utterance length in SmartKom. The stars indicate a significant difference.

$$\Delta\mathrm{Affect} = \mathrm{Affect}_{\mathrm{before\,overlap}} - \mathrm{Affect}_{\mathrm{after\,overlap}} \qquad (2)$$

Afterwards, we average over all segments ($\overline{\Delta\mathrm{Affect}}$). The significance of the affective change is tested by using the Mann-Whitney-U-Test. The stars denotes the significance level: * $p < 0.01$ and ** $p < 0.001$.

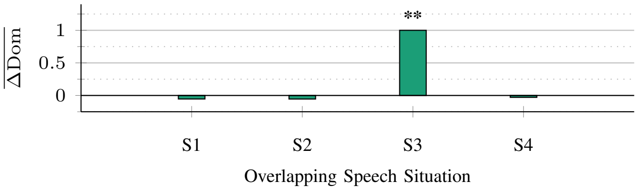

Figure 5. Dominance change at the point where overlapping speech occurs, stars indicate a significant affective state change

1) Davero Corpus: Analyzing the change of affective states in connection with overlapping speech only in S3 (simultaneous starting after longer silence) a significant change in the affective state can be observed, see Figure 5. A possible interpretation for this observation is that the dominance level is dropping. Having a deeper analysis of the data, we can state that the dominance level of the interrupter is raising, while the dominance of the speaker whose turn is interrupted is slightly decreasing. In this case, the overlapping speech event could be a good marker for identifying changes in dominance. For all other situations of overlapping speech, the dominance of the two speakers is not influenced by overlapping speech.

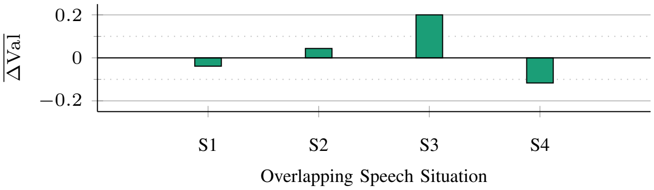

Figure 6. Valence change at the point where overlapping speech occurs in the Davero corpus.

For the change of the speaker's valence, we can state that there is no significant connection with the occurrence of overlapping speech, cf. Figure 6. This could be expected as overlapping speech is related to the turn-taking behavior of the speakers and the dominance of a speaker is seen as the underlying mechanism to regulate the turn-taking [12].

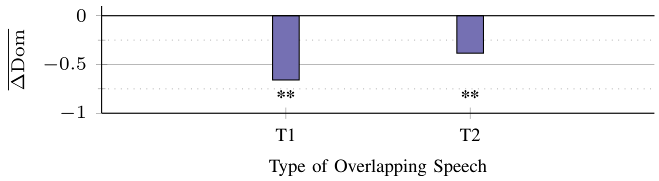

Figure 7. Dominance change at the point where overlapping speech occurs in the SmartKom Corpus. The stars indicate a significant affective state change.

2) SmartKom Corpus: Regarding Figure 7, it can be seen that the dominance for both types is significantly higher after the overlapping speech segment than before. This it quite obvious for the case where the user actively interrupts the system, but when the system interrupts the user this seems quite unintuitive and can only be explained in connection with the valence change. Regarding the valence change (cf. Figure 8), we can state that after the overlapping speech the user is significantly more moved to negative values. This finding in connection with a higher dominance shows that it can be assumed that the user is more angry after overlapping speech, either because he wants to speak and interrupts the system, or he is annoyed because the system interrupts him.

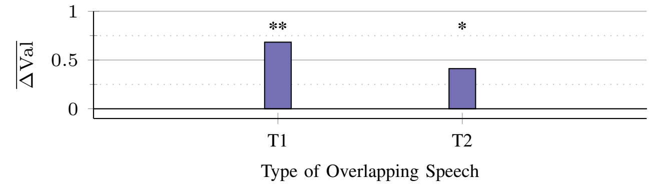

Figure 8. Valence change at the point where overlapping speech occurs in the SmartKom Corpus. The stars indicate a significant affective state change.

Regarding our third research question: “Is overlapping speech occurring at points where the affective state is changing?”, we can conclude that in HHI only changes in the dominance are related to overlapping speech, whereas in HCI significant changes in both affective dimensions, dominance and valence, can be observed.

V. Conclusion

In this paper, we present a first study investigating overlapping speech effects in both HHI and HCI. The analyses are conducted on a dataset of realistic HHI containing telephone based conversations and the well-known SmartKom Corpus of naturalistic HCI. We could show that in both datasets overlapping speech occurs frequently enough to be analyzed with a share of 18.9% for HHI and 12.3% for HCI. For the investigated HHI, this share is in-line with the results of other research groups [11], [33]. The amount of overlap in HCI is a bit lower but still sufficient. For this no numbers of other researchers are to our best knowledge reported in the literature.

Based on the description of situational settings, we first analyzed the correlation between the length of overlap-preceding utterances and the occurrence of overlap. As a result of our first analysis, we could expose significant relations to the length of the spoken utterances and changes in the affective state of the conversational partners. In HHI we could find a significant correlation between overlapping speech as feedback and premature turn-taking. Also in HCI a significant correlation is found between overlapping speech and the length of system utterances. The user’s utterance-lengths did not show significant correlations for the occurrence of overlaps, we assume that these overlaps are just caused by operator malfuntions, as the SmartKom data are recorded in a WOZ-scenario. And now pre-defined design rules are given for the wizards how to use overlap [21].

Secondly, we analyzed the correlation of affective changes at in the surrounding of the overlap in both types of interactions. For this investigation, we showed that overlapping speech goes along with changes in the affective states of dominance and valence in certain situations. In HHI only the situation where both speakers start simultaneously after a longer pause effects a significant change of dominance. For the valence dimension no significant correlation could be found. In the investigated HCI both affective dimensions, dominance and valence, show a significant correlation to overlapping speech in both situation types.

From these results, we are able to derive some rules for the organization of interactions: In telephone based HHI the utterances should not be too long and the listener should be encouraged to give feedback. This avoids competitive barge-in overlaps. For HCI, the system should not talk to long as for all overlapping speech segments, an affective change to higher dominance and negative valence of the speaker can be observed. But this kind of affective change should be avoided. To evaluate these statements for their generality, a broader investigation including additional corpora has to be conducted.

A possible application of our investigations in HHI and HCI is the identification of parts where the affective state changes based on the knowledge of overlapping speech and the dialog course: As e.g. situation S3, where both speakers start simultaneously, can be easily identified by duration analysis, this knowledge can be used to find affective material for further emotional analyses.

In our further research activities, we will develop a robust automatic identification of the different types of overlap. Together with the recognition of the user’s affective state, we are a step further to future Cognitive Infocommunication systems acting as a companion towards human users [13], [14].

VI. Acknowledgements

The work presented in this paper was done within the Transregional Collaborative Research Centre SFB/TRR 62 ‘Companion-Technology for Cognitive Technical Systems’ funded by the German Research Foundation (DFG).

References

[1] F. Schulz von Thun, *Miteinander reden 1 - Störungen und Klärungen*. Reinbek, Germany: Rowohlt, 1981.

[2] R. Ishii, K. Otsuka, S. Kumano, and J. Yamato, “Analysis of respiration for prediction of ”who will be next speaker and when?” in multi-party meetings,” in *Proc. of the 16th International Conference on Multimodal Interaction*, ser. ICMI ’14, Istanbul, Turkey, 2014, pp. 18–25.

[3] H. Sacks, E. A. Schegloff, and G. Jefferson, “A simplest systematics for the organization of turn taking for conversation,” *Language*, vol. 50, pp. 696–735, 1974.

[4] C. Peters, C. Pelachaud, E. Bevacqua, M. Mancini, and I. Poggi, "A model of attention and interest using gaze behavior," in *Intelligent Virtual Agents*, ser. LNCS, T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, and T. Rist, Eds. Springer Berlin Heidelberg, 2005, vol. 3661, pp. 229–240.

[5] E. Schegloff, "Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences," D. Tannen, Ed. Washington, D.C.: Georgetown University Press, 1982, pp. 71–93.

[6] T. Stivers, "Stance, alignment and affiliation during storytelling: when nodding is a token of affiliation," *Research on Language and Social Interaction*, vol. 41, no. 1, pp. 31–57, mar 2008.

[7] G. Jefferson, *Notes on some orderlinesses of overlap onset*. Padua, Italy: Cleup Editore, 1983, pp. 11–38.

[8] E. A. Schegloff, "Overlapping talk and the organization of turn-taking for conversation," *Language in Society*, vol. 29, pp. 1–63, 2000.

[9] P. French and J. Local, "Turn competitive incomings," *Journal of Pragmatics*, vol. 7, pp. 701–715, 1983.

[10] E. Kurtić, G. J. Brown, and B. Wells, *Fundamental frequency height as a resource for the management of overlap in talk-in-interaction*. Bingley, UK: Emerald Group Publishing Limited, 2009, pp. 183–204.

[11] ——, "Resources for turn competition in overlapping talk," *Speech Commun.*, vol. 55, no. 5, pp. 721–743, Jun. 2013.

[12] D. Heylen, E. Bevacqua, C. Pelachaud, I. Poggi, J. Gratch, and M. Schröder, "Generating listening behaviour," in *Emotion-Oriented Systems*, ser. Cognitive Technologies, R. Cowie, C. Pelachaud, and P. Petta, Eds. Springer Berlin Heidelberg, 2011, pp. 321–347.

[13] A. Wendemuth and S. Biundo, "A companion technology for cognitive technical systems," in *Cognitive Behavioural Systems*, ser. LNCS, A. Esposito, A. Esposito, A. Vinciarelli, R. Hoffmann, and V. Müller, Eds. Berlin Heidelberg, Germany: Springer, 2012, vol. 7403, pp. 89–103.

[14] P. Baranyi., A. Csapo, and G. Sallai, *Cognitive Infocommunications (CogInfoCom)*. Berlin, Germany: Springer, 2015.

[15] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, pp. 39–58, 2009.

[16] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun*, vol. 53, pp. 1062–1087, 11 2011.

[17] I. Siegert, D. Prylipko, K. Hartmann, R. Böck, and A. Wendemuth, "Investigating the form-function-relation of the discourse particle "hm" in a naturalistic human-computer interaction," in *Recent Advances of Neural Network Models and Applications*, ser. Smart Innovation, Systems and Technologies, S. Bassis, A. Esposito, and F. C. Morabito, Eds. Berlin, Heidelberg, Germany: Springer, 2014, vol. 26, pp. 387–394.

[18] I. Siegert, M. Haase, D. Prylipko, and A. Wendemuth, "Discourse particles and user characteristics in naturalistic human-computer interaction," in *Human-Computer Interaction. Advanced Interaction Modalities and Techniques*, ser. LNCS, M. Kurosu, Ed. Berlin, Heidelberg, Germany: Springer, 2014, vol. 8511, pp. 492–501.

[19] P. Baranyi and A. Csapo, "Definition and synergies of cognitive infocommunications," *Acta Polytechnica Hungarica*, vol. 9, pp. 67–83, 2012.

[20] I. Siegert, D. Philippou-Hübner, M. Tornow, R. Heinemann, A. Wendemuth, K. Ohnemus, S. Fischer, and G. Schreiber, "Ein Datenset zur Untersuchung emotionaler Sprache in Kundenbindungsdialogen," in *Proc. of the 26th ESSV*, Eichstätt, Germany, 2015, pp. 180–187.

[21] W. Wahlster, Ed., *SmartKom: Foundations of Multimodal Dialogue Systems*. Heidelberg, Berlin: Springer, 2006.

[22] F. Schiel, S. Steininger, and U. Türk, "The SmartKom multimodal corpus at BAS," in *Proc. of the 3rd LREC*, 2002, pp. 35–41.

[23] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, 03 1977.

[24] M. Skowron, S. Rank, M. Theunis, and J. Sienkiewicz, "The good, the bad and the neutral: Affective profile in dialog system-user communication," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Springer Berlin Heidelberg, 2011, vol. 6974, pp. 337–346.

[25] D. Prylipko, D. Rösner, I. Siegert, S. Günther, R. Friesen, M. Haase, B. Vlasenko, and A. Wendemuth, "Analysis of significant dialog events in realistic human–computer interaction," *Journal on Multimodal User Interfaces*, vol. 8, pp. 75–86, 2014.

[26] K. R. Scherer, "What are emotions? and how can they be measured?" *Soc Sci Inform*, vol. 44, pp. 695–729, 2005.

[27] I. Siegert, R. Böck, and A. Wendemuth, "Inter-Rater Reliability for Emotion Annotation in Human-Computer Interaction – Comparison and Methodological Improvements," *Journal of Multimodal User Interfaces*, vol. 8, pp. 17–28, 2014.

[28] R. Böck, I. Siegert, M. Haase, J. Lange, and A. Wendemuth, "ikannotate – a tool for labelling, transcription, and annotation of emotionally coloured speech," in *Affective Computing and Intelligent Interaction*, ser. LNCS, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Berlin, Heidelberg, Germany: Springer, 2011, vol. 6974, pp. 25–34.

[29] R. Böck, I. Siegert, B. Vlasenko, A. Wendemuth, M. Haase, and J. Lange, "A processing tool for emotionally coloured speech," in *Proc. of the 2011 IEEE ICME*, Barcelona, Spain, 2011, p. s.p.

[30] S. Steininger, S. Rabold, O. Dioubina, and F. Schiel, "Development of the user-state conventions for the multimodal corpus in smartkom," in *Proc. of the 3rd LREC*, 2002, pp. 371–377.

[31] I. Jacobi and F. Schiel, "Interlabeller Agreement in SmartKom Multi-Modal Annotations," Ludwig-Maximilians-Universität München, Tech. Rep., 12 2003.

[32] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic Emotion Recognition: A Benchmark Comparison of Performances," in *Proc. of the IEEE ASRU-2009*, Merano, Italy, 2009, pp. 552–557.

[33] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.