Vocal turn-taking patterns in groups of children performing collaborative tasks: an exploratory study

Jaebok Kim, Khiet P. Truong, Vicky Charisi,
Cristina Zaga, Manja Lohse, Dirk Heylen, Vanessa Evers

Human Media Interaction, University of Twente, Enschede, The Netherlands
{j.kim, k.p.truong, v.charisi, c.zaga, m.lohse, d.k.j.heylen, v.evers}@utwente.nl

Abstract

Since children (5-9 years old) are still developing their emotional and social skills, their social interactional behaviors in small groups might differ from adults' interactional behaviors. In order to develop a robot that is able to support children performing collaborative tasks in small groups, it is necessary to gain a better understanding of how children interact with each other. We were interested in investigating vocal turn-taking patterns as we expect these to reveal relations to collaborative and conflict behaviors, especially with children behaviors as previous literature suggests. To that end, we collected an audiovisual corpus of children performing collaborative tasks together in groups of three. Through automatic turn-taking analyses, our results showed that speaker changes with overlaps are more common than without overlaps and children seemed to show smoother turn-taking patterns, i.e., less frequent and longer lasting speaker changes, during collaborative than conflict behaviors.

Index Terms: nonverbal behaviors, children speech, social signal processing

1. Introduction

As robots are becoming increasingly intelligent, it becomes feasible to design social robots that can get along with children and facilitate social interactions among children in social situations [1]. In order to create socially normative interactions between children and the robot, and to provide support to the children, the robot should be able to interpret social states among the children. Moreover, since children develop social behaviors in a small group [2], it is necessary to understand how they behave and interact in different social contexts in a small group.

Although there has been some research on the (automatic) analysis of small group interactions, which discussed fundamental topics such as social roles, engagement, dominance, and affective states, most of these studies targeted adults [3, 4, 5, 6]. These studies revealed the significance of nonverbal behavioral cues, e.g. gaze, turn-takings, body postures, head gestures among others, in communication. Consequently, much of the research in automatic human behavior understanding in multiparty interactions has used these cues to detect social states.

However, it is unlikely that these previous findings obtained from groups of adults transfer to groups of children who have not fully developed their social and communication skills yet. More significantly, children do not have the same social context as adults have. Particularly, educational and psychology research have found that *social interaction*, *collaboration* and *play* are important features in the child's development [7, 8, 9]: collaborative plays in small groups can have benefits on learning and development, especially in primary education [10, 11].

In order to develop a social robot that interacts with groups of children, we need to gain a better understanding of the situated nonverbal cues expressed by these children, and automatically process and understand these cues. Although some previous studies have explored conversational analysis of collaboration or social interaction among children [12, 13, 14, 15], patterns of vocal interaction in a small group of children involving collaborative tasks remain unexplored. The lack of research in this area might be due to the fact that performing research on children raises many challenges and consequently, there are not many corpora available for the research.

Vocal turn-taking is a highly comprehensive phenomenon which incorporates various nonverbal cues: e.g. speech, silences, overlaps, and interruptions, and it is known to be associated with a context of conversation: e.g. competition, collaboration, and cognitive load [16, 17, 18]. Particularly, turn-taking patterns of children are known to be different from those of adults [19]. Moreover, turn-taking patterns in (non-)collaboration among children remain unveiled.

In this paper, we present our multimodal corpus of small groups of children interactions. Also, we present an exploratory study into vocal nonverbal interactions among 3 children who are performing collaborative tasks together. Following an inductive approach as commonly applied by social scientists in observational studies, we identified distinctive behaviors: collaboration and conflict in groups of children working together on a 3D puzzle. We analyzed how turn-taking-related nonverbal features relate to these distinctive behaviors in collaborative tasks. Specifically, we aimed to address the following questions: 'What are the most dominant turn-taking patterns in children's speech in small group interaction?', 'How are these turn-taking patterns correlated with collaborative and conflict behaviors?', and 'What turn-taking features are distinctive between collaborative and conflict behaviors?'

This paper is structured as follows. In section 2, we describe more details of previous work related to nonverbal analysis of children speech. In section 3, we present how we collected an audiovisual corpus from children using the 3D puzzle task. In section 4, results of turn-taking feature analysis are presented. Lastly, we summarize and discuss our findings.

2. Related Work

Vocal nonverbal behaviors include all spoken cues conveying not only explicit messages but also underlying messages [5]. So far, voice activity, voice quality, (para) linguistic vocalisations, silences, interruptions and turn-taking features have been

investigated as major cues related to certain types of social interaction or context [20, 18, 21, 5, 22].

There have been several studies targeting children's vocal interactions in social contexts, focusing on affective states and engagement [23, 24, 25, 26, 27]. In [26, 27], low level acoustic features and high level nonverbal features (e.g. overlaps and voice activity) were used to classify engagement levels. Moreover, affect bursts, speech duration, reaction time, and intensity were investigated as social markers [25]. However, in these studies, children interacted either digital pets or psychologists. Furthermore, some studies focused on more acoustic features rather than nonverbal features.

Only few studies investigated social behaviors in child-to-child nonverbal interactions in a small group, see for example [28, 15]. In [15], they built individual and group engagement models to classify disengagement levels using nonverbal features (e.g. smiling, leaning, and backchanneling). However, the use of vocal nonverbal features was limited to hand-coded speech activity and backchannels.

For children at early ages, vocal turn-taking is a significant social skill to learn among others and they often show speaker changes with overlaps and interruptions [19, 29] while 'no-gap-no-overlap' is socially normative for adults [30, 17]. In contrast to previous works, we focus on an automatic analysis of turn-taking patterns of children in a context of (non-)collaborative behaviors. Given previous literature, we believe that these behaviors are also visible in the way children interact with each other, in particular in the way children exchange speaking turns.

3. Data

3.1. Corpus collection

In order to observe distinctive interactional behaviors related to collaboration, we designed a 3D puzzle task in such a way to support children's collaboration, i.e. building a given structure (e.g. a shape of an animal) together. In our corpus design, we adopted conceptualization of collaboration defined as "a coordinated, synchronous activity that is the result of a continued attempt to construct and maintain a shared conception of a problem" [31, 32]. We expect children to learn, share ideas, and reach given common goals. For this study, Dutch children (9 female and 12 male, $n = 21$) aged 5 - 8 ($6.95 \pm .95$) were recruited from a primary school. Children were first clustered according to their age and then assigned randomly to a group of three for each session. We believed that the size of a group would be appropriate to trigger collaboration as regarded to be the smallest possible social group [2]. We recorded video and audio using 3HD cams, 1 microphone-array, and 3 microphones in a room of the school. Each child's voice was separately collected by using a lapel microphone. We recorded 10 sessions totally, but 3 sessions were excluded because of privacy issues and malfunctioning of equipment. Eventually, 7 sessions are available, totaling approximately 3 hours long.

3.2. Annotation

Two annotators coded the children's behaviors using the ELAN tool [33]. In general social signal processing studies, the ground truth of social behaviors relies on human interpretations. However, a high level of social behaviors such as collaboration might be arduous to code since it requires coders to interpret multiple cues from a group of subjects simultaneously, and it might be biased to a particular cue depending on subjective observations and interpretations. In our study, we did not ask coders to interpret interactional behaviors directly. Instead, we defined 21 so-called low-level behaviors, which are mostly related to task-engagement (e.g. giving blocks and receiving blocks), and asked the annotators to code the start and end times of these behaviors. We believe that emergence of collaboration is observable when we have clear clues of how children perform tasks and interact with each other [34].

Category	i-behavior	r-behavior	n	μ	σ
C	giving blocks	receiving blocks	86	2.62	1.72
F	grasping blocks from others	stopping others	48	2.14	2.23
N	giving blocks	observing others	446	5.34	4.56

Table 1: *Summary of interactional behaviors (i(r)-behavior: initiator(responder)'s behavior, n: number of segments, μ: mean duration (sec), σ: standard deviation of duration)*

Next, we defined roles of children as the **initiator** who triggers social interactions first and the **responder** who responds. We looked at the interaction between the intention of the initiator to collaborate and the response of the responder. If the responder accepts the intention (e.g. giving blocks) and responds to it by a certain form of responding behaviors (e.g. receiving blocks), we concluded that an actual collaboration happens and denoted it as **collaboration** (**C**). On the other hand, the initiator and the responder also showed different aspects of social interactions such as a form of conflicts (e.g. grasping blocks from others - stopping others). We denoted these contrast behaviors as **conflict** (**F**). Lastly, if the responder rejects the intention of collaboration or does not pay attention to it, we regarded it as **neglected collaboration** (**N**). To measure an inter-coder agreement level on the categories, 15% of the data was double coded by the two coders. By considering high prevalence, we computed Gwet's AC1 [35], resulting in .701 ($p < .01$).

Based on our definitions, we derived sequences of interactional behaviors. The sequences are limited to cases where the initiating and responding behavior are observed consecutively, which means that the responding behavior begins at least before the initiating behavior ends. We do not know what happens between two consecutive behaviors if there is a time gap between them and it is hard to decide how long the time gap should be in order to call it an 'interactional sequence'. Eventually, we excluded sequences which have a time gap between the initiating behavior and the responding behavior. Moreover, all sequences are mutually exclusive of each other. Table 1 presents the categories, examples, and descriptive statistics.

4. Analysis

In this section, we describe the vocal turn-taking features used in our study. We present the analysis and results addressing general turn-taking patterns in children's speech, correlations between turn-taking patterns and **C**, **F**, **N**, and distinctive turn-taking features among **C**, **F**, **N**.

4.1. Turn-taking related features

Based on previous research on between-speaker-intervals and interruptions [17, 18], we categorized turn-taking patterns into two cases depending on whether speech overlaps occur during speaker changes or not as Table 2 describes. The two cases are illustrated in Figure 1. The first case (a) describes a 'clear'

Features (short notation)	Description
speech	Voice activity (e.g. words, phrases, and sentences) detected by VAD
self-silence	Silence between speech of each speaker
overlap	Overlap of speech between two speakers
speaker change without overlaps (change)	Speaker change only if there is no overlap between all speech right before or after the change
speaker change with overlaps (change-ov)	Speaker change if there is overlap between any speech right before or after the change
successful interruption (s-int)	Speaker starts speaking while another is speaking and the speaker ends her/his turn after the another does
unsuccessful interruption (u-int)	Speaker starts speaking while another is speaking and the speaker ends her/his turn before another does

Table 2: *Description of vocal nonverbal features*

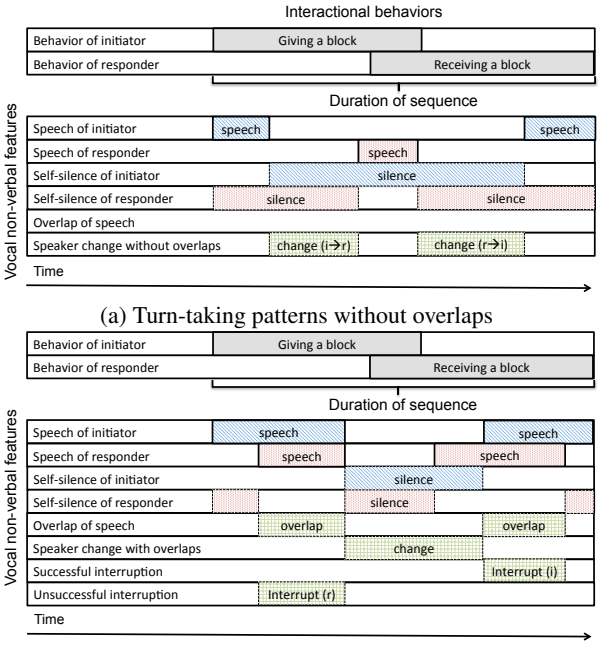
speaker change without overlaps, for example, a change from the responder to the initiator, i.e. '**change (r→i)**' or a change from the initiator to the responder, i.e. '**change (i→r)**'. Note that the speech of the responder possibly starts prior to receiving a block, which means that we extract all features from the beginning of the initiating behavior to the end of the responding behavior. The second case (b) describes an 'unclear' speaker change with overlaps. We further distinguish between successful (when the interrupter talks longer than the interruptee), unsuccessful (when the interrupter stops talking before the interruptee) interruptions and speaker change with overlaps: '**change**' feature that denotes the interval between speech (or overlap) right before and after the change.

Affect bursts [25], laughter [36], and backchannels [37] are considerable nonverbal features, but we excluded them because of the limited automatic extraction. Since we focus on nonverbal features of relatively long term behaviors, we do not have such a dynamic model of low level acoustic features yet. Eventually, we excluded these too. Moreover, some previous studies suggested the threshold to decide reasonable duration of nonverbal features [17]. However, we did not limit duration since we do not have pre-knowledge of proper maximum duration for children turn-taking.

We applied Voice Activity Detection (VAD) to find the individual speech parts (which is available since we recorded each voice separately) and subsequently extracted all features automatically. All features were normalized by the use of mean and frequency which we will discuss in each section separately.

4.2. Overall turn-taking patterns

In order to analyse overall patterns of turn-taking and to investigate which speaker change patterns are dominant, we collected all features for each child regardless of the behaviors or roles. We calculate frequency (total count of the feature/total duration in sec) and mean duration (total duration of the feature/total count of the feature). Table 3 summarizes the averaged results. Apparently, children show more speaker changes with overlaps than without overlaps (Welch two sample t-test for *change* and *change-ov*: $t = -7.62, df = 25.46, p < .00001$). It seems that speaker changes with overlaps are more natural for children in collaborative contexts. As [19, 29] suggests, this type of turn-taking behavior emerges from their still immature communication skills in small groups.



(a) Turn-taking patterns without overlaps

(b) Turn-taking patterns with overlaps

Figure 1: *Illustrations of nonverbal features*

Features	Frequency (count/sec)	Mean duration (sec)
speech	$.127 \pm .035$	1.927 ± 1.721
overlap	$.118 \pm .050$	$1.058 \pm .862$
self-silence	$.109 \pm .038$	3.256 ± 2.381
change	$.138 \pm .031$	2.308 ± 2.809
change-ov	$.293 \pm .084$	2.719 ± 2.684
s-inter	$.067 \pm .027$	$.921 \pm .819$
u-inter	$.051 \pm .024$	$1.248 \pm .902$

Table 3: *Descriptive statistics of nonverbal features*

4.3. Correlations between behaviors and turn-takings

We investigated how (non-)collaborative behaviors correlate with turn-taking patterns. First, we calculated the frequency and mean duration of each **C**, **F**, and **N** segment (frequency: total count of the segments/total session duration, mean duration: total segment duration / total count of the segments). Subsequently, we used the turn-taking features as described in Section 4.2 and transformed these into z-scores to calculate Spearman's rank correlations between these features and the occurrence and mean duration of **C**, **F**, and **N**. Table 4 summarizes our findings. Correlations for **N** were not significant.

In Table 4 (a), we observe that the frequency of **F** significantly correlates with the frequency of all features investigated. It seems that children that show more conflict behaviors also are more active and competitive in their vocal interaction patterns. With respect to **C**, we find mostly negative correlations which are significant for **overlap** and **u-inter**: the more collaborative children are, the fewer competitive behaviors they show.

For correlations between durations of {**C**, **F**, **N**} and turn-taking features, we only find two significant ones, namely **change-ov** and **change**, see Table 4 (b). Children showing longer durations of conflict tend to show shorter speaker changes.

In addition, we studied correlations between frequency of

	(a) frequency			(b) mean duration		
	C	F	N	C	F	N
speech	$-.08$	$+.39^{*}$	$+.03$	$+.06$	$-.05$	$+.18$
self-silence	$-.20$	$+.40^{*}$	$-.05$	$+.04$	$-.38$	$-.04$
overlap	$-.42^{*}$	$+.58^{**}$	$-.25$	$-.30$	$+.11$	$+.08$
change	$+.28$	$+.42^{*}$	$+.14$	$+.13$	$-.47^{*}$	$+.02$
change-ov	$-.19$	$+.59^{**}$	$-.10$	$+.10$	$-.62^{**}$	$+.27$
s-inter	$-.33$	$+.57^{**}$	$-.16$	$-.22$	$+.07$	$+.22$
u-inter	$-.43^{*}$	$+.59^{**}$	$-.21$	$-.22$	$+.12$	$+.06$

Table 4: *Correlation between features and categories ($df = 19$, $*$ and $**$: significance with $p < .05$ and $p < .01$, respectively)*

{**C**, **F**, **N**} and durations of turn-taking features and correlations between durations of {**C**, **F**, **N**} and frequency of turn-taking features. We do not present all results here, but from significant results: frequency of **change-ov** and mean duration of **F** ($r = +.70, p < .001$) and mean duration of **u-inter** and frequency of **C** ($r = -.57, p < .01$), we could reinforce our previous findings: conflict behaviors are more associated with 'competitive' speaker changes than collaborative behaviors are.

4.4. Distinctive turn-taking features among (non-) collaborative behaviors

We studied how turn-taking features vary among segments of **C**, **F**, and **N**. For this, we collected the features from all segments of each category. Since the optimal size of a window for feature extraction is unknown, we extracted features from the beginning of the i-behavior to the end of the r-behavior (see Fig. 1). We estimated frequency (count of a feature/duration of a segment) and mean duration (averaged duration of the feature in a segment) for each feature in each behavior type. Table 5 presents the averaged results. Note that we extracted individual features: i.e. i(r)-speech and i(r)-self-silence for speech and self-silence of the initiator (responder), respectively since we could specify features by the role of children, in contrast to the previous analyses.

Table 5 (a) reports the averaged frequency of nonverbal features with respect to the categories and their pairwise comparison results. To assess differences of the features among the categories, we conducted Kruskal Wallis test ($df = 2, p < .01$), followed by Nemenyi test for the pairwise comparisons. First, we observe that for both **C** and **F**, frequencies of the features are higher than those in **N**. Also, most of the features indicated higher frequencies in **F** than in **C**. Children seem to speak and take overlapping turns more frequently in the conflict: while **change-ov** yielded a significant difference, **change** did not.

With respect to duration, we observe in Table 5 (b) that children seem to have longer **overlap**s and **u-inter** in **F** than those in **C**. On the other hand, **change**s are longer in **C** than in **F**. These observations seem to imply that collaborative actions can be associated with a more 'relaxed' way of grabbing turns by showing less frequent but longer speaker changes and shorter overlaps, which are different from 'no-gap-no-overlap' of adults. Lastly, we found durations of all features are longer in **N** than any others. Since verbal interactions are less frequent but longer in **N**, children seem to 'loosen' turn-taking when they neglect or ignore the intention of collaboration.

5. Conclusions and Discussion

In this study, we explored characteristics of vocal turn-takings among children with respect to different social interactions: collaboration and conflict. We collected a spontaneous audiovisual corpus with child-child interactions and automatically extracted turn-taking related features from this data. Based on our analyses, we reached three conclusions. First, we found that speaker changes with overlaps are more frequent than those without overlaps in child-child interactions in collaborative contexts. Second, we found significant *positive* correlations between the frequency of conflict and the frequency of turn-taking features such as overlap, speaker changes with overlaps, and (un)successful interruptions. We also found reasonable *negative* correlations between the frequency of collaboration and overlap and unsuccessful interruption. Third, we found significant differences between the collaboration and conflict categories with respect to the frequency of speaker changes with overlap.

	C	F	N	C-F	F-N	C-N
(a) Frequency of nonverbal features						
i-speech	.289	.485	.263	*	**	
r-speech	.304	.574	.181	*	****	
i-self-silence	.313	.389	.212		*	
r-self-silence	.383	.556	.202	*	****	*
overlap	.057	.107	.048			
change	.150	.136	.110			
change-ov	.386	.638	.181	**	****	*
s-inter	.053	.077	.047			
u-inter	.070	.077	.039			
(b) Mean duration of nonverbal features						
i-speech	1.166	.945	1.778		****	**
r-speech	1.016	.792	1.447			
i-self-silence	1.463	1.008	1.876		*	
r-self-silence	1.426	1.130	2.199		***	**
overlap	.462	.623	1.045	****	****	****
change	.970	.651	1.209	****	****	***
change-ov	1.421	.963	1.730			
s-inter	.442	.795	.668			
u-inter	.462	.623	1.045	****	****	****

Table 5: *Analysis of features respect to the categories (i(r): initiator(responder), *, **, ***, and ****: significance with .05, .01, .001, and .0001 of Nemenyi test, respectively)*

In summary, in collaborative task-based child-child interactions, speaker changes with overlap are more common than without overlap. Children also seem to show less frequent but longer speaker changes with overlaps during collaboration than in conflict. Our findings imply that children generally have competitive turn-takings, but they show relatively relaxed turn-takings in collaboration. These results give us insights for automatic assessment of social states in groups of children: vocal turn-taking features seem to be promising indicators.

For future studies, we need to elaborate on more sophisticated models that are capable of capturing dynamics of low-level features as described in [22]. In addition, we need more detailed turn-taking features and a more rigorous way to annotate social behaviors. Finally, many other aspects such as a child's personality or sociometric status should also be taken into account in the final model. Eventually, our aim is to develop automatic classification of high-level social behaviors, such as collaboration, conflict, or exclusion, in groups of children.

6. Acknowledgements

The research leading to these results was supported by the European Community's 7th Framework Programme under Grant agreement 610532 (SQUIRREL - Clearing Clutter Bit by Bit).

7. References

[1] B. Robins, K. Dautenhahn, R. Te Boekhorst, and A. Billard, "Robotic assistants in therapy and education of children with autism: can a small humanoid robot help encourage social interaction skills?" *Universal Access in the Information Society*, vol. 4, no. 2, pp. 105–120, 2005.

[2] C. Stangor, *Social groups in action and interaction*. Psychology Press, 2004.

[3] R. F. Bales, "A set of categories for the analysis of small group interaction," *American Sociological Review*, vol. 15, no. 2, pp. 257–263, 1950.

[4] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, vol. 27, no. 12, pp. 1775–1787, 2009.

[5] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schröder, "Bridging the gap between social animal and unsocial machine: A survey of social signal processing," *Affective Computing, IEEE Transactions on*, vol. 3, no. 1, pp. 69–87, 2012.

[6] Y. Hayashi, H. Morita, and Y. I. Nakano, "Estimating collaborative attitudes based on non-verbal features in collaborative learning interaction," *Procedia Computer Science*, vol. 35, pp. 986–993, 2014.

[7] L. S. Vygotsky, *Mind and society: the development of higher mental processes*. Cambridge, MA: Harvard University Press, 1978.

[8] J. Piaget, *The psychology of the child*. New York: Basic Books, 1972.

[9] M. B. Parten, "Social participation among pre-school children," *The Journal of Abnormal and Social Psychology*, vol. 27, no. 3, pp. 243–269, 1932.

[10] K. A. Bruffee, *Collaborative learning: Higher education, interdependence, and the authority of knowledge*. ERIC, 1999.

[11] S. Benford, B. B. Bederson, K.-P. Åkesson, V. Bayon, A. Druin, P. Hansson, J. P. Hourcade, R. Ingram, H. Neale, C. O'Malley *et al.*, "Designing storytelling technologies to encouraging collaboration between young children," in *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 2000, pp. 556–563.

[12] C. L. Sidner, C. Lee, C. D. Kidd, N. Lesh, and C. Rich, "Explorations in engagement for humans and robots," *Artificial Intelligence*, vol. 166, no. 1, pp. 140–164, 2005.

[13] A. Moreno, R. Van Delden, R. Poppe, and D. Reidsma, "Socially aware interactive playgrounds," *Pervasive Computing*, vol. 12, no. 3, pp. 40–47, 2013.

[14] L. Tian, D. Duan, J. Cui, L. Wang, H. Zha, and H. Aghajan, "Video based children's social behavior classification in peer-play scenarios," in *Proceedings of the IEEE Asian Conference on Pattern Recognition*, 2013, pp. 770–774.

[15] I. Leite, M. McCoy, D. Ullman, N. Salomons, and B. Scassellati, "Comparing models of disengagement in individual and group interactions," in *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 99–105.

[16] L. Ten Bosch, N. Oostdijk, and J. P. De Ruiter, "Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues," in *Text, speech and dialogue*. Springer, 2004, pp. 563–570.

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

[18] D. B. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez, "Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues," in *Proceedings of the 10th international conference on Multimodal interfaces*. ACM, 2008, pp. 45–52.

[19] B. Maroni, A. Gnisci, and C. Pontecorvo, "Turn-taking in classroom interactions: Overlapping, interruptions and pauses in primary school," *European journal of psychology of education*, vol. 23, no. 1, pp. 59–76, 2008.

[20] L. Chen, M. Harper, A. Franklin, T. R. Rose, I. Kimbara, Z. Huang, and F. Quek, "A multimodal analysis of floor control in meetings," in *Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 36–49.

[21] A. Delaborde and L. Devillers, "Use of nonverbal speech cues in social interaction between human and robot: emotional and interactional markers," in *Proceedings of the 3rd international workshop on Affective interaction in natural environments*. ACM, 2010, pp. 75–80.

[22] M. Cristani, A. Pesarin, C. Drioli, A. Tavano, A. Perina, and V. Murino, "Generative modeling and classification of dialogs by a low-level turn-taking feature," *Pattern Recognition*, vol. 44, no. 8, pp. 1785–1800, 2011.

[23] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. J. Russell, and M. Wong, "You stupid tin box-children interacting with the aibo robot: A cross-linguistic emotional speech corpus." in *Proceedings of LREC*, 2004.

[24] S. Yildirim, S. Narayanan, and A. Potamianos, "Detecting emotional state of a child in a conversational computer game," *Computer Speech & Language*, vol. 25, no. 1, pp. 29–44, 2011.

[25] M. Tahon, A. Delaborde, and L. Devillers, "Corpus of children voices for mid-level markers and affect bursts analysis." in *Proceedings of LREC*, 2012, pp. 2366–2369.

[26] R. Gupta, C.-c. Lee, D. Bone, A. Rozga, S. Lee, and S. Narayanan, "Accoustic analysis of engagement behavior in children," in *Workship on Child Computer Interaction 2012*, 2012, pp. 1–7.

[27] R. Gupta, C.-c. Lee, S. Lee, and S. Narayanan, "Assessment of a child's engagement using sequence model based features," in *Workshop on Affective Social Speech Signals 2013*, 2013.

[28] J. Sanghvi, G. Castellano, I. Leite, A. Pereira, P. W. McOwan, and A. Paiva, "Automatic analysis of affective postures and body motion to detect engagement with a game companion," in *Proceedings of 6th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 2011, pp. 305–311.

[29] E.-T. Susan, "Children's verbal turn-taking," in *Developmental pragmatics*. New York: Academic Press, 1979, pp. 391–429.

[30] D. C. O'Connell, S. Kowal, and E. Kaltenbacher, "Turn-taking: A critical analysis of the research tradition," *Journal of psycholinguistic research*, vol. 19, no. 6, pp. 345–373, 1990.

[31] J. Roschelle and S. D. Teasley, "The construction of shared knowledge in collaborative problem solving," in *Computer supported collaborative learning*. Springer, 1995, pp. 69–97.

[32] A. Weinberger, K. Stegmann, and F. Fischer, "Knowledge convergence in collaborative learning: Concepts and assessment," *Learning and Instruction*, vol. 17, no. 4, pp. 416–426, 2007.

[33] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: a professional framework for multimodality research," in *Proceedings of LREC*, 2006, pp. 5–8.

[34] L. J. Corrigan, C. Peters, G. Castellano, F. Papadopoulos, A. Jones, S. Bhargava, S. Janarthanam, H. Hastie, A. Deshmukh, and R. Aylett, "Social-task engagement: Striking a balance between the robot and the task," in *Embodied Commun. Goals Intentions Workshop ICSR*, vol. 13, 2013, pp. 1–7.

[35] K. L. Gwet, *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[36] K. P. Truong and D. A. Van Leeuwen, "Automatic discrimination between laughter and speech," *Speech Communication*, vol. 49, no. 2, pp. 144–158, 2007.

[37] N. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in english and japanese," *Journal of pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.