Automatic analysis of children's engagement using interactional network features

Jaebok Kim, Khiet P. Truong

Human Media Interaction, University of Twente, The Netherlands
{j.kim, k.p.truong}@utwente.nl

Abstract

We explored the automatic analysis of vocal non-verbal cues of a group of children in the context of engagement and collaborative play. For the current study, we defined two types of engagement on groups of children: harmonised and unharmonised. A spontaneous audiovisual corpus with groups of children who collaboratively build a 3D puzzle was collected. With this corpus, we modelled the interactions among children using network-based features representing the centrality and similarity of interactions. The centrality measures how interactions among group members are concentrated on a specific speaker while the similarity measures how similar the interactions are. We examined their discriminative characteristics in harmonised and unharmonised engagement situations. High centrality and low similarity values were found in unharmonised engagement situations. In harmonised engagement situations, we found low centrality and high similarity values. These results suggest that interactional network features are promising for the development of automatic detection of engagement at the group level.
Index Terms: children, engagement, social network, non-verbal

1. Introduction

The state-of-the-art in social signal processing has contributed to the development of social robots facilitating engagement among a group of people [1]. For example, a robot could play the role of a side-participant and support interactions of a participant who is not engaged with others, called "weak engagement" in triadic interactions [2]. In child-child interactions, this weak engagement problem could often be observed since children (6–9 yrs) are still developing social skills at their own pace. Moreover, children learn social interactions in collaborative play in which "harmonised and unharmonised engagement" [2] can often occur. "harmonised engagement" is defined as the situation where children interact substantially and keep their connections during play. On the other hand, in "unharmonised engagement", a child is left out of the interaction (weak engagement). However, due to the great heterogeneity and temporal dynamics of engagement in a group of children [3], it might be challenging to point out who is harmonised or unharmonised among a group. Hence, as a first step, we explore features and characteristics on a group-level: how can we model characteristics of engagement in group members' interactions in the context of collaborative play?

Engagement types are characterised by the way children interact with each other in a group. Hence, we focus on a feature representation capturing these group interactions rather than individual behaviours. In social network analysis (SNA), centrality and similarity measures were introduced to characterise interactions among nodes [4]. SNA has been previously applied in other applications: bioinformatics and conversational analysis [5, 6, 7]. For instance, turn-taking patterns were modelled in a social network to predict social traits automatically [6], and the centrality of turn-taking was used to measure social verticality [7]. We employ SNA to analyse engagement types of a group since SNA characterises interactional flows which could be utilised to model turn-taking, i.e. maintenance of connection in engagement [1].

Although SNA achieved reasonable performances for modelling interactions in a large group of adults, it remains unknown if SNA is feasible for modelling spontaneous social behaviours in small groups of children who may display unpredictable behaviours compared to adults. To the best of our knowledge, the study of the automatic analysis of engagement types of children in the context of collaborative play still remains largely unexplored. In this study, we explore automatic analysis of engagement types in small groups of children using two network-based features: centrality and similarity to take the group interaction of each child into account.

This paper is structured as follows. In Section 2, related studies are introduced, and we present an audiovisual corpus of groups of children and identified engagement types in Section 3. Section 4 defines network-based features modelling interactions among a group. Analysis results are presented and discussed in Section 5, and conclusions are drawn in Section 6.

2. Related work

Modelling of engagement has been extensively studied in the field of Human Robot Interaction (HRI) and Social Signal Processing (SSP) [1, 8, 9, 10, 11]. In [8], various visual cues, e.g. gaze and gesture, were utilised to detect individual and group engagement in 500 ms long segments. However, their features were limited to hand-coded labelling while we are aiming for automatic engagement detection. More importantly, turn-taking between children was not studied although turn-taking is strongly associated with social behaviours [12].

To model speaker-changes between more than two participants, centrality of interactions was employed in social role and dominance detection tasks [7]. While the centrality achieved limited performances, there is still room for improvement of modelling turn-taking. Moreover, social network analysis has been employed to capture structural information about interactions [6]. In particular, speaker traits were clustered using similarity of their turn-taking styles. However, the usage of similarity for modelling small groups remains unexplored.

The aforementioned studies investigated child-robot or child-computer interactions. More significantly, important aspects of social behaviours among small groups, i.e. temporal dynamics and relative levels of behaviours, were often ne-

glected. First of all, in a similar way as for the temporal dynamics of speech, temporal patterns of social behaviours vary occasionally [3]. For instance, it is rare to observe the same level of engagement from the beginning to the end of each task or play. Since the engagement of children alters dynamically over time, it is not desirable to point out who is highly engaged for each session [3] although such a rough approach was often employed [1, 13].

Furthermore, a child can show differing social behaviours depending on who participates in play [14]. To be more specific, individuals vary their non-verbal behaviours depending on whom they interact with and in the situation. For instance, the behaviours of a child low in engagement rely on other people showing more or less engagement in specific contexts. In other words, the context surrounding people greatly affects their behaviours and one of the most significant contexts is the group of people who they interact along with at any given time. Hence, we first focus on modelling engagement on a group-level using centrality and similarity measures rather than modelling individual engagement.

3. Data

3.1. Corpus collection

We designed a 3D puzzle task facilitating children's spontaneous social interactions. Using 3D magnetic cube blocks, children were asked to build given shapes of animals together. 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 age and then assigned randomly to mixed-gender groups of three for each session in order to maximize the diversity in social interactions. Finally, eight sessions were chosen for subsequent analyses, leading to approximately 3 hours of audiovisual data. More details of the setup are presented in [15].

3.2. Annotation

By employing the concept of engagement, the process by which multiple participants maintaining perceived connection during verbal and non-verbal interactions [1], engagement can be observed in various forms characterised by multi-modal interactions. In our work, we identify engagement types by focusing on exchanges of attention among the children during the play. First, we looked into the difficulty of individual engagement coding due to the great variability of their interactions within groups. To prevent judgement biased to speech, we did not give two coders any access of individual speech recordings but the description of engagement and videos. They were asked to code levels of engagement of each child in an absolute manner, i.e. {low, medium, high}. As expected, it was difficult to judge the level of engagement of a group member in the absolute manner, resulting in poor inter-rater agreement (kappa) between two coders (.57). Hence, we designed an annotation based on ranks, which measures a relative level of engagement as follows (from low to high level of engagement):

- **1**: giving relatively less attention to others and receiving relatively less attention from others.
- **2a**: giving relatively less attention to others but receiving attention from others.
- **2b**: giving attention to others but receiving relatively less attention from others.
- **3**: giving attention to others and receiving attention from others.

By using these descriptions, children in a group can be ordered from a low to a high level of engagement. Only if no differences could be observed among the three children, ties were allowed (e.g. {1, 1, 1}, {3, 3, 3}). We treated the classes {2a} and {2b} as equally ranked (in level of engagement) and merged them into one class {2}.

A suitable size of a segment for annotation varies on the context. For instance, 0.5 and 5 seconds (sec) long sized segments were used to predict engagement and roles, respectively [8, 16]. Through several pilot coding sessions, we found, on an empirical basis, that 5 sec long segments were suitable for the annotators to observe various levels of engagement.

Again, to prevent any judgement relying on speech, we provided the coders with only the videos and descriptions. The coders coded every 5 sec segment using the ELAN tool [17]. The agreement level (kappa) was reported as 0.82. Furthermore, we removed speechless segments that do not contain any vocal cues and selected only those segments whose labels were agreed by both coders. From these relative levels of individual engagement, we derive group level engagement: harmonised engagement (HE) and unharmonised engagement (UE). When there is no child who is less engaged than others, i.e. 3-3-3, this would be considered **HE**. All other cases were categorised into **UE**. In **UE**, some children were not engaged in the interactions and often played alone, i.e. 1-1-1. The resulting data set contains a total of 1017 segments (**HE**: 304 and **UE**: 713).

4. Features

As mentioned previously, we do not have pre-knowledge of features for engagement types on groups of children; thus, we investigate sets of vocal non-verbal features and acoustic features based on related works [12, 15]. Based on these, we explored three different feature sets: individual non-verbal (baseline), network-based non-verbal, and acoustic features, as summarised in Table 2.

4.1. Automatic extraction

We aim to develop the automatic analysis of interactions among a group of children. Therefore, we extract our features from every 5s long segment [15] in an automatic way. First, we extract each child's speech segment using voice activity detection from each child's lapel microphone recording. Then, in order to correct errors caused by noise and channel-inferences, we applied automatic speaker identification using iterative model update and manual correction. In a similar way as described in [18], we use Mel-frequency cepstral coefficients (MFCC) features and Gaussian-Mixture-Model (GMM) to detect segments from different speakers. As a result, we obtain "Inter-Pausal Units (IPUs)" (0.5 sec of the minimum length for speech and silence) of speech in each segment [12]. We bridge two speech segments only if there is a short silence (< 0.5 sec) between them. Based on IPUs, we extract features for subsequent analyses. Then, these are normalised into frequency, denoted as FQ (= the number of IPUs / 5 sec), mean duration of IPUs (MD), and PR (= total duration of IPUs / 5 sec) [15]. Note that 5 sec is the duration of each segment for both types: harmonised and unharmonised engagement.

$A_{f(i,j)}$	C_1	C_2	C_3
C_1	self-silence	$C_2 \rightarrow C_1$	$C_3 \rightarrow C_1$
C_2	$C_1 \rightarrow C_2$	self-silence	$C_3 \rightarrow C_2$
C_3	$C_1 \rightarrow C_3$	$C_2 \rightarrow C_3$	self-silence

Table 1: *Interactional flow matrix for a feature: f*

4.2. Network-based turn-taking features

To model the interactivity among children in a group, we focus on the centrality and similarity of interactions in a group. The centrality measures how interactions among group members are concentrated on a specific speaker while the similarity measures how similar the interactions are. Since we focus on the types of the group conversation instead of individual types, we need to derive overall (i.e. group-level) centrality and similarity of interactions. We expect the overall centrality to show lower values for **HE** than for **UE**: rather than having a relatively centralised interactions, in harmonised engagement, interactions are expected to be more equally distributed among the children. On the other hand, the overall similarity is expected to be higher in the case of **HE** than in that of **UE**.

To model the centrality and similarity, we devise an interactional flow network and matrix as shown in Table 1. Let us denote $A_{f(i,j)}$ as an interactional flow that j'th child precedes i'th child, i.e. $C_j \rightarrow C_i$ in the matrix. Each type of flow is represented by a feature (f) among "clear speaker change (change)", "unclear speaker change with overlap (change-ov)", "successful interruption (s-int)", and "unsuccessful interruption (u-int)", selected in [15]. Note that **change-ov** occurs when there is mutual self-silence between children preceded or followed by speech-overlaps. Moreover, we add "self-silence" [15] (pause), which is regarded as a self-interactional flow $A_{f(i,i)}$ in order to model the maintenance of turns. Eventually, each feature has its own matrix to model interactional flows among a group. In the matrix, a row vector: x_i describes all flows from other children to child i (e.g. $x_1 = [A_{f(1,1)}, A_{f(1,2)}, A_{f(1,3)}]$).

4.2.1. Centrality of interactions

Centrality of interactions can be explained by two terms: frequency and duration of interactional flows, i.e. changes or interruptions ({change, change-ov, s-int, u-int}). For example, if speaker changes from other children to i'th child are frequent or shorter, interactions are highly centralised on the focal child (i.e. i'th child). The centrality of a feature (f) of an i'th child (C_i) among a total K number of children is measured as follows:

$$CT(f_{(C_i)}) = \frac{K-1}{\sum_{j=1}^{K} IF_{j \rightarrow i}}, \forall_i = 1, 2, ...K \qquad (1)$$

where $IF_{j \rightarrow i}$ is the intensity function of the feature (f) representing the interactional flow from j'th child to i'th child, denoted as $A_{f(i,j)}$ in the matrix. The definition of the intensity varies depending on the normalisation and feature types. We use only two types of normalised values for each feature: frequency (FQ) and mean duration (MD) since the proportion of duration (PR) is already modelled in centrality. We have two intensity functions (IF): one for speaker changes and the other for interruptions. Note that centrality should be higher when interactions are centralised on a specific child.

First, for speaker changes: {change, change-ov}, the intensity function (IF) of mean duration is equal to MD of the feature. This function increases the centrality when the duration of change decreases. In other words, the quicker the focal child takes a turn from others, the higher the centrality becomes. For the frequency (FQ), IF should be FQ^{-1}, which increases the centrality when the focal child takes turns from others more frequently.

Second, for interruptions: {s-int, u-int}, the intensity function (IF) of mean duration is MD^{-1} of the feature. This function increases the centrality when the duration of interruptions increases. In other words, the longer the focal child interrupts others, the higher the centrality becomes. For the frequency (FQ), IF is equal to FQ^{-1} as the same as speaker changes.

Based on these features, we calculate the overall centrality (OCT) of each feature (f) in the network using Freeman's centrality [19] defined as follows:

$$OCT(f) = \frac{\sum_{i=1}^{K} [CT' - CT(f_{(C_i)})]}{H}, \forall_i = 1, 2, ...K \quad (2)$$

where CT' is the maximum among the centrality of children and H is the normalising factor, which varies on the topology of the network. For simplicity, we use $(K-1)$ for H.

4.2.2. Similarity of interactions

To model the similarity of interactions, the normalised feature sets are the same as those for centrality. Again, in the matrix, a row vector: x_i describes all interactional flows (of feature f) from others to child i. Then, we measure similarity between row vectors: x_i and x_j by using Gaussian kernel which is defined as follows:

$$K_f(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \qquad (3)$$

The parameter γ is used to control the kernel bandwidth:

$$\gamma = \frac{\widetilde{\gamma}}{(\frac{1}{K^2} \sum_{i=1}^{K} \sum_{j=1}^{K} A_{f(i,j)})} \qquad (4)$$

which means that we normalise the parameter by dividing it by the average value of all interactional flows in the matrix. We could set a bandwidth parameter, i.e. $\widetilde{\gamma}$ using cross-validation but use 1 for a practical reason. In addition, we added a mean value of $K_f(x_i, x_j)$ from possible combinations: $\{K_f(x_1, x_2)$, $K_f(x_1, x_3)$, $K_f(x_2, x_3)\}$ as overall similarity (OS).

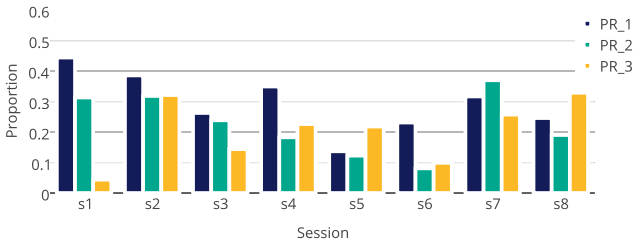
Note that self-silence regulates high sensitivity for cases: self-flow $A_{f(i,i)}$. For example, if we measure the similarity of interactional flows for the first and second child: $K_f(x_1, x_2)$ without self-silence, which means $A_{f(1,1)} = 0$ and $A_{f(2,2)} = 0$, the similarity becomes highly sensitive to $A_{f(1,2)}$ and $A_{f(2,1)}$ (since it calculates a distance between vectors: $[0, A_{f(2,1)}, A_{f(3,1)}]$ and $[A_{f(1,2)}, 0, A_{f(3,2)}]$).

4.2.3. Acoustic features and baseline

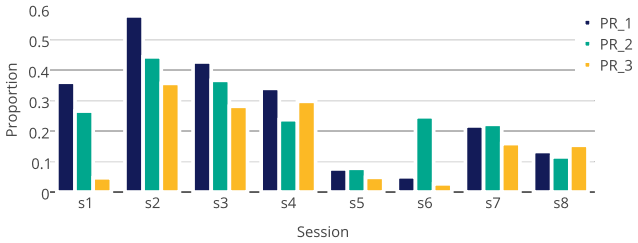
We extracted F0, energy, HNR, ZCR, jitter, and shimmer and added their $\{\Delta, \Delta\Delta\}$ by using openSMILE [20] as a representative set of acoustic cues of social behaviours [11]. For these features, we cannot use the centrality or similarity measures since they do not have interactional flows. Instead, we simply obtained mean and standard deviation (SD) values of the feature vectors for each child as individual features. Finally, for

Category	Features
non-verbal (baseline)	individual features (18)
	speech (9), self-silence(9)
network features	centrality of features (32)
	speaker change (8), speaker change with overlaps (8), successful interruptions (8), unsuccessful interruptions (8)
	similarity of features (32)
	speaker change (8), speaker change with overlaps (8), successful interruptions (8), unsuccessful interruptions (8)
acoustic	SD of features (18)
	F0 (3), energy (3), ZCR (3), HNR (3), jitter (3), shimmer (3)
	Mean of features (18)
	F0 (3), energy (3), ZCR (3), HNR (3), jitter (3), shimmer (3)

Table 2: Feature sets (number of features)



(a) Proportion in unharmonised engagement (UE)



(b) Proportion in harmonised engagement (HE)

Figure 1: Proportion of speaking time of each child

the baseline set, we used normalised frequency (FQ), mean duration (MD), and proportion (PR) of speech and self-silence, which are all widely used in engagement detection [11].

5. Analysis

In this section, we analysed interactions of children in harmonised and unharmonised engagement situations. First, we investigate the individual features known to be associated with engagement [21]. Next, we look into the network-based features modelling turn-taking between children and present statistical significances of differences of feature values between harmonised and unharmonised engagement.

5.1. Baseline: proportion of speaking

Among the baseline features, the proportion of speaking time has been associated with engagement [21]. However, as mentioned in [22], one of the participants in a group might be less active in engagement than the others even if the level of speaking activity is high. Therefore, we first looked into how the proportion of speaking time (PR) is distributed in each group of children and whether this could be used as an indicator of the engagement types.

In Fig. 1, PR is shown for each child in a session. Note that PR_1, 2, and 3 are extracted from first, second, and third child in a group, respectively. Based on previous studies, one could assume that in cases of **UE**, the distribution of speaking time for three children might be unequal and show a higher variance than for **HE**. However, we observe that in both **UE** and **HE**, the distribution of speaking time is relatively unbalanced and that no clear patterns can be found. To discover if PR is distinctive between the types, we conducted Kruskal Wallis tests on the standard deviation (SD) of PR_1, 2, and 3 in each group. This revealed that there is no significant difference ($p = .344$) of PR between the types. Unlike in previous studies, the simple individual features (e.g. the proportion of speaking time) did not prove to be discriminative. We conducted the same tests for all other features: "clear speaker change (change)", "unclear speaker change with overlap (change-ov)", "successful interruption (s-int)", and "unsuccessful interruption (u-int)" [15]. However, we could not find any significant differences ($p < .05$). This analysis led us to elaborate on network-based features, to attempt to model engagement types of a group.

5.2. Network-based turn-taking

Types	OCT-change-MD (****)	OCT-change-ov-MD (****)	OCT-s-int-MD (*)	OCT-u-int-MD (.)
UE	+.029	+.034	+.007	+.077
HE	-.069	-.080	-.003	-.033
	OCT-change-FQ (****)	OCT-change-ov-FQ (****)	OCT-s-int-FQ (****)	OCT-u-int-FQ (.)
UE	+.047	+.031	+.016	+.036
HE	-.089	-.058	-.008	-.019

Table 3: *Analysis of overall centrality (OCT)*

To interpret the behaviour of certain features in the **HE** and **UE**, we conducted Kruskal Wallis tests ($df = 1$) to investigate the distinctiveness of these features. Since our focus is on the network-based features, we first address overall centrality (OCT) and similarity (OS) of turn-taking features. We calculated the mean values of z-scores of the features of the **HE** and **UE** ($****$: significance level $p < .0001$).

First, Table 3 shows the overall centrality of turn-taking features. Note that the OCT of all features shows higher values for UE than HE. In other words, in UE, interactional flows are indeed more centralised on a specific child rather than showing equal distribution. In particular, speaker-change related features: {OCT-change-MD, OCT-change-ov-MD} showed significant differences ($p < .0001$).

In Table 4, the overall similarity (OS) indicates lower values for UE rather than HE. We can interpret this finding to indicate that interactions are more equal and similar to each other during HE. In particular, all features {OS-change-MD, OS-change-ov-MD, OS-s-int-MD, OS-u-int-MD} showed significant differences between the types although the overall centrality of unsuccessful interruptions (u-int) did not show significant results. In summary, we found that highly focused interactional

Types	OS-change-MD (****)	OS-change-ov-MD (****)	OS-s-int-MD (****)	OS-u-int-MD (****)
UE	-.123	-.131	-.122	-.120
HE	+.289	+.308	+.288	+.281
	OS-change-FQ (****)	OS-change-ov-FQ (****)	OS-s-int-FQ (****)	OS-u-int-FQ (****)
UE	-.102	-.115	-.102	-.104
HE	+.193	+.218	+.193	+.197

Table 4: *Analysis of overall similarity (OS)*

Types	energy-SD (****)	energy-M (****)	HNR-SD (*)	zcr-SD (****)
HE	-.223	-.224	-.057	-.196
UE	+.117	+.118	+.030	+.103

Table 5: *Analysis of acoustic features, SD (standard-deviation), M (mean)*

flows occur in UE. However, turn-taking patterns between children were similar to each other's in HE. Compared to centrality, similarity showed more significant differences between UE and HE.

Furthermore, we studied differences of acoustic features between UE and HE. We presented significant results in Table 5; energy, HNR, and ZCR related features showed differences. We found that children's speech showed higher variances of the features in UE compared to HE. Moreover, the unweighted average of energy was higher in UE. In future work, we will study modelling group-interactions by using these features, which are expected to be more conclusive.

5.2.1. Summary and discussion

In this section, we analysed how distinctive network-based turn-taking and acoustic features are among (un)harmonised engagement types. We observed that interactions are more centralised on a specific child in **UE** (higher overall centrality) than in **HE**. Also, interactions are more similar to each other in **HE** (higher overall similarity) than in **UE**. From our findings, we conclude that network-based turn-taking features are able to capture characteristics of interactions in spite of the small number of participants ($N = 3$). Moreover, acoustic features related to energy, HNR, and ZCR demonstrated considerable differences between UE and HE. Hence, not only the network-based turn-taking features but also acoustic features seem to be promising features for the automatic classification of the types. However, since acoustic features are extracted from shorter frames (e.g. 20ms) compared to turn-taking features (e.g. speaker-change), various methods for integration of all features should be investigated as future work.

6. Conclusions

In this study, we explored the automatic analysis of engagement types among children using network-based turn-taking features. We collected a spontaneous audiovisual corpus with child-child interactions and defined group-level properties: harmonised and unharmonised engagement. To characterise children's interactions in these types, we developed interactional network features to represent levels of both centrality and similarity of interactional flows. In particular, these features modelled turn-taking flows among a group of children.

First, we explored whether the proportion of speaking time of individuals was discriminative of engagement types. However, we could not find any statistically significant differences of features between the types, which means that individual features are not capable of modelling interactions among a group. Next, we found that centrality and similarity of turn-taking features showed significant differences between the types. In the unharmonised engagement type, centrality showed higher values while similarity showed lower values compared to the harmonised engagement type. In other words, children tend to show similar turn-taking patterns when they are engaged in a harmonised way. On the other hand, they demonstrate unbalanced or centralised turn-taking patterns in cases of unharmonised engagement.

In future work, based on the results of our analysis, we will develop statistical models classifying the engagement types. In particular, we will investigate integration of turn-taking features and acoustic features. Moreover, to extend our study to HRI, we will conduct a new data corpus, i.e., collection of not only child-child but also child-robot interactions.

7. 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).

8. References

[1] 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.

[2] Y. Matsuyama, I. Akiba, S. Fujie, and T. Kobayashi, "Four-participant group conversation: A facilitation robot controlling engagement density as the fourth participant," *Computer Speech & Language*, vol. 33, no. 1, pp. 1–24, 2015.

[3] S. Al Moubayed and J. Lehman, "Toward better understanding of engagement in multiparty spoken interaction with children," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 211–218.

[4] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge university press, 1994, vol. 8.

[5] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug–target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.

[6] J. Grothendieck, A. Gorin, and N. Borges, "Social correlates of turn-taking behavior," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2009, pp. 4745–4748.

[7] 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.

[8] 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.

[9] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan, "Detecting engagement in hri: An exploration of social and task-based context," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*. IEEE, 2012, pp. 421–428.

[10] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, 2014.

[11] 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.

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

[13] S. Strohkorb, I. Leite, N. Warren, and B. Scassellati, "Classification of children's social dominance in group interactions with robots," in *Proceedings of the ACM on International Conference on Multimodal Interaction*. New York, NY, USA: ACM, 2015, pp. 227–234.

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

[15] J. Kim, K. P. Truong, V. Charisi, C. Zaga, M. Lohse, D. Heylen, and V. Evers, "Vocal turn-taking patterns in groups of children performing collaborative tasks: an exploratory study," in *Proceedings of the INTERSPEECH*, 2015, pp. 1645–1649.

[16] F. Pianesi, M. Zancanaro, B. Lepri, and A. Cappelletti, "A multimodal annotated corpus of consensus decision making meetings," *Language Resources and Evaluation*, vol. 41, no. 3-4, pp. 409–429, 2007.

[17] 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.

[18] C. Busso, P. G. Georgiou, and S. S. Narayanan, "Real-time monitoring of participant's interaction in a meeting using audio-visual sensors," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2007.

[19] L. C. Freeman, "Centrality in social networks conceptual clarification," *Social networks*, vol. 1, no. 3, pp. 215–239, 1979.

[20] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the international conference on Multimedia*. ACM, 2010, pp. 1459–1462.

[21] N. Campbell and S. Scherer, "Comparing measures of synchrony and alignment in dialogue speech timing with respect to turn-taking activity." in *INTERSPEECH*, 2010, pp. 2546–2549.

[22] K. Jokinen, "Turn taking, utterance density, and gaze patterns as cues to conversational activity," in *Proceedings of ICMI 2011 Workshop Multimodal Corpora for Machine Learning: Taking Stock and Road mapping the Future, Alicante, Spain*, 2011, pp. 31–36.