Towards automatic estimation of conversation floors within F-formations

Raman, Chirag; Hung, Hayley

DOI
10.1109/ACIIW.2019.8925065

Publication date
2019

Document Version
Final published version

Published in
2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos, ACIIW 2019

Towards automatic estimation of conversation floors within F-formations

Chirag Raman
Delft University of Technology
c.a.raman@tudelft.nl

Hayley Hung
Delft University of Technology
h.hung@tudelft.nl

***Abstract*—The detection of free-standing conversing groups has received significant attention in recent years. In the absence of a formal definition, most studies operationalize the notion of a conversation group either through a spatial or a temporal lens. Spatially, the most commonly used representation is the *F-formation*, defined by social scientists as the configuration in which people arrange themselves to sustain an interaction. However, the use of this representation is often accompanied with the simplifying assumption that a single conversation occurs within an F-formation. Temporally, various categories have been used to organize conversational units; these include, among others, *turn*, *topic*, and *floor*. Some of these concepts are hard to define objectively by themselves. The present work constitutes an initial exploration into unifying these perspectives by primarily posing the question: can we use the observation of simultaneous speaker turns to infer whether multiple conversation floors exist within an F-formation? We motivate a metric for the existence of distinct conversation floors based on simultaneous speaker turns, and provide an analysis using this metric to characterize conversations across F-formations of varying cardinality. We contribute two key findings: firstly, at the average speaking turn duration of about two seconds for humans, there is evidence for the existence of multiple floors within an F-formation; and secondly, an increase in the cardinality of an F-formation correlates with a decrease in duration of simultaneous speaking turns.**

***Index Terms*—free-standing conversational groups, conversation floors, speaking turns**

I. Introduction

Imagine a social scenario like a mingling or networking event. Interactions in such a setting involve multiple dynamic conversations which are a medley of ever evolving topics and partners. And yet, humans can instinctively navigate the complexities of such encounters. How do we do this? We regulate our exchanges both spatially and temporally using implicit social norms or explicit behavioural signals [1]. Furthermore, these cues could be either verbal or non-verbal, expressed visually, vocally, or verbally through spoken language.

A deeper understanding of these group dynamics constitutes a natural objective towards the realisation of machines with social skills. For instance, consider a social robot approaching a group of people in a public space, or the use-case of evaluating attendee experience at a conference poster session. In these and other cases, having an understanding of the dynamics, and where channels of social influence lie, would enable the artificial agent to develop increasingly sophisticated policies for interaction or inference. Conversation groups have been of importance in the application domains of social robotics [2]–[5], activity recognition [6], [7], social surveillance [8]–[10], and social signal processing [11], [12].

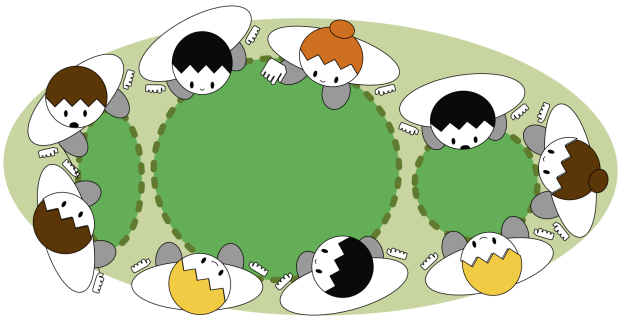

Fig. 1. Depiction of a single F-formation with multiple conversation floors. The darker green regions within dotted lines represent distinct simultaneous conversation floors. Most works representing a conversing group as an F-formation make the simplifying assumption that a single conversation occurs within an F-formation with a joint focus of attention for all members.

Fundamental to the study of such conversations is defining the notion of a free-standing conversational group (FCG). While it is easier to objectively conceptualize an FCG in spatial terms in a scene of multiple interacting groups, delineating the boundary of conversations poses a greater technical challenge. We could think of separating conversations on the basis of topics, but this is challenging if audio data is unavailable due to privacy concerns. We could operationalize a conversation as a set of participating members, but this membership is challenging to infer visually for non-speaking participants. This often leads to the simplifying assumption in some literature that the focus of an FCG is a single conversation. As we illustrate in Fig. 1, and discuss in the following sections, this may not always be the case.

In the present work, we dive beyond the geometric bounds of an FCG to gain a deeper understanding of the conversations occuring within it. In this initial approach, we focus specifically on speaking participants as the most decisive indicator of the existence of a conversation. Concretely, we pose the following broad research questions:

RQ 1. Can we use observed speaker turns to infer the con-

This research was partially funded by the Netherlands Organization for Scientific Research (NWO) under project number 639.022.606.

978-1-7281-3891-6/19/$31.00 ©2019 IEEE

versation floors within an F-formation?

RQ 2. How does the cardinality of an F-formation affect the conversation floors developed within it?

The ground truth for speaker turns in this work comes from manual annotations of video data, mimicking use-cases where audio data might be unavailable due to privacy concerns. Concretely, our contributions are as follows: conceptually, we provide an indicator of distinct conversation floors that uses speaking turns alone, and situate this indicator in schisming literature [13]–[15]; analytically, we provide evidence that multiple conversation floors exist within an F-formation, and show that the cardinality of an F-formation correlates negatively with turn duration of simultaneous speakers.

The rest of this paper is organized as follows. We describe some of the spatial and temporal perspectives used to study FCGs in Section II. In Section III we provide a review of literature involving the use of these spatial or temporal notions, motivating the need to consider both of these aspects in unison. In Section IV, we propose an operationalization of an indicator of distinct conversation floors, building upon the concepts of conversation schisming. The dataset we use and the experiments performed for answering the research questions are described in Section V and Section VI respectively. Finally, Section VII summarizes our findings and concludes the paper.

II. Background

Spatial Factors. One of the most common proxemic notions to describe an FCG is Adam Kendon's *Facing Formation*, or *F-formation*, originally defined as:

> An F-formation arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access. [16, p. 210]

Kendon argues that activity is always located, and denotes the space in front of a person that is used for the activity as the person's *transactional segment*. When two or more people come together to perform some activity, they are liable to arrange themselves such that their individual transactional segments overlap to create a joint transactional space. This joint space between the interactants is called an *o-space*. As we discuss in the next section, many computational works involving the automatic detection of FCGs from video focus on the detection of F-formations, often assuming that the transaction involves a single conversation.

Temporal Factors. The conversation of focus in an FCG, however, is dynamic in nature. If conversations change over time, what are the temporal units that describe their underlying structure? Some of the terms used in early literature to organize conversational units include *turn*, *topic*, *gap*, and *floor*. Edelsky provides an excellent review of these concepts in [17], stating that most of these units were defined on the basis of some technical or mechanical structure such as signals of speakers or auditors, ignoring the intention of the participant. Using inferred participants' meanings rather than technical definitions, Edelsky defines turns and floors as follows:

> The floor is defined as the acknowledged what's-going-on within a psychological time/space. What's going on can be the development of a topic or a function (teasing, soliciting a response, etc.) or an interaction of the two. It can be developed or controlled by one person at a time or by several simultaneously or in quick succession.[17, p. 405]

III. Related Work

Detecting Conversational Groups. In most works, a conversational group is operationalized as an F-formation. Early work on the task of detecting FCGs in video data developed concurrently from two perspectives: those that estimate the location of the o-space using a Hough-voting strategy [8], [18]; or those that view an F-formation as a set with individuals being assigned exclusive membership [12], [19]. There has also been considerable work focused on incorporating temporal information for the same task of detecting conversational groups [7], [20]–[22]. Notably, these approaches utilise the head pose as a proxy for Visual Focus of Attention (VFoA) [9] in addition to the body pose to model F-formation membership, and assume a single conversation within an F-formation. The assumption that members in a group have a single joint focus of attention is seen in other works as well. Hung, Jayagopi, Ba, *et al.* [23] model a single joint focus of visual attention of participants to estimate dominance in groups.Vazquez, Steinfeld, and Hudson Vazquez, Steinfeld, and Hudson also assume a single conversation within an F-formation while developing a policy for a robot to be aware of a single focus of attention of the conversation.

Estimating involvement. In a conversation, the floor is typically held by a single participant at a time [13]. What then characterizes the silent participants in a conversation group? The following works demonstrate that the task of estimating participant involvement is subjective in nature, and that gaze behaviour and turn-taking patterns can be informative.

Zhang and Hung [24], [25] study the task of detecting associates of an F-formation; members that are attached to an F-formation but do not have full status [16]. They argue that the labeling of conversation groups is not an objective task. Collecting multiple annotations of perceived associates, they demonstrate how detecting them can improve initial estimates of full-members of an F-formation. Oertel, Funes Mora, Gustafson, and Odobez [26] characterize silent particiants into multiple categories (attentive listener, side participant, bystander) from audiovisual cues. Oertel and Salvi [27] also show that it is possible to estimate individual engagement and group involvement in a multiparty corpus by analysing the participants' eye-gaze patterns. Bohus and Horvitz [28] propose a self-supervised method for forecasting disengagement with an interactive robot using a conservative heuristic. The heuristic is constructed by leveraging features that capture how close the participant is, whether a participant is stationary or moving, and whether a participant is attending to the robot.

Some works also used turn-taking features to estimate some notion of involvement. Pentland, Madan, and Gips [29] measured engagement by the z-scored influence each person has on the other's turn-taking for a pair of participants. Hung and Gatica-Perez [30] found that the pause duration between an individual's turns, aggregated at group level, is highly predictive of cohesion in small group meetings.

Schisming. In a conversation with at least four participants, the conversation sometimes splits up into two or more conversations. This transformation is referred to as a *schism* [13] or *schisming*. One of the earliest allusions to the phenomenon of schisming based on anecdotal evidence occurs in the work of Goffman, who suggested that a gathering of two participants *exhausts* an encounter and forms a *fully-focused gathering* [31, p. 91]. With more than two participants, there may be persons officially present in the situation who are not themselves so engaged. These *bystanders* change the gathering into a *partly-focused* one. If more than three persons are present, there may be more than one encounter carried on in the same situation, resulting in a *multifocused* gathering.

In subsequent work, Sacks, Schegloff, and Jefferson [13] and Goodwin [14] both indicated that the co-existence of two turn-taking systems is the most decisive characteristic of schisming. This view was supported by Egbert, who demonstrated that although schisming is a participation framework with two simultaneous conversations, each with its own turn-taking system, there is an interface between them during schisming [15]. She also makes a systematic differentiation between overlap and simultaneous talk during schisming. In overlap, simultaneous speakers compete for the floor, an event usually resolved by returning to *one-speaker-at-a-time*. In schisming by contrast, simultaneous speakers orient to one of two distinct floors, an event which if resolved successfully, results in the establishment of two floors [15, p. 43]. Overlapping speech is therefore expected to occur throughout the lifespan of all conversation floors within an F-formation.

IV. Methodology

In this section we build upon the previously discussed concepts to propose using simultaneous speakers in an F-formation as an initial conservative indicator of the existence of distinct conversation floors.

A common concern with observing groups of conversing people is the potential violation of privacy. In our experience with collecting group interaction datasets, participants often regard having their microphone data recorded and transcribed as being more invasive than being captured on video. In these situations, the lack of verbal information makes it extremely challenging to infer the topics being discussed. How can we then investigate the existence of distinct conversations? Two observations could prove useful:

Inferring schisms without audio data. The relationship between body movements such as gestures and speech has been long established in literature [32]. Some works have shown promising results in estimating the presence of voice activity from automated gestural analysis or accelerometer data [33]–[35]. It therefore seems feasible that speaker turns can be automatically estimated without audio data. Combined with the observation that the co-existence of two turn-taking systems is the most decisive characteristic of schisming, we argue that it is in turn reasonable to explore the inference of schisms without audio data through speaking turns.

Linking schisming to floors and F-formations. While Egbert does explicitly use the term *floor* to describe the conversations resulting from a schism, it is useful to observe how this relates back to Edelsky's view of floors. Edelsky defined floors in terms of the acknowledged *what's-going-on* within a psychological time space. The object of focus here could either be a topic or some other function. To borrow Goffman's terms, a schism effectively changes a gathering into a *multifocused* one, where each object of focus can be viewed to correspond to a floor in Edelsky's definition. However, if the participant's lower bodies remain configured such that their transactional segments overlap to produce a common o-space, they would still remain in the same F-formation even if the conversation has undergone a schism into two or more distinct floors. Fig. 1 depicts this situation conceptually.

Combining these two broad observations, we argue that it is feasible to explore the existence of distinct conversation floors within an F-formation without audio data, whilst capturing speaker turns from visual observations. We propose to start with the following metric. Given a sliding window w of speaking duration d, we consider a *speaker* to be a participant who speaks for the entire duration d. The number of simultaneous *speakers* thus defined corresponds to the number of distinct conversation floors at that position of w, since they correspond to speaking turns in distinct floors.

Of course, the metric is inextricably tied to the duration d being considered; too short a duration, and the concurrent turns might capture either backchannels or the overlapping speech within the same floor as described in Egbert's work. However, a reasonably long duration would capture the speaking turns of participants holding distinct floors. This leads to the question: what qualifies as a reasonable choice for d to differentiate overlaps within a floor from turns in distinct floors? In our

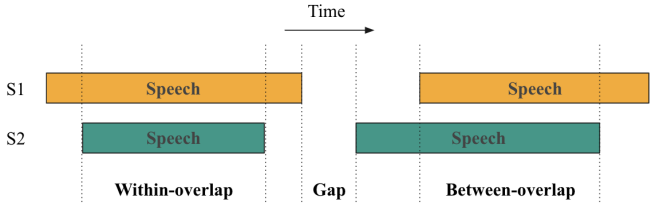

Fig. 2. Illustration of gaps, within-overlaps, and between-overlaps for two speakers (S1 and S2) within the same floor. The scheme was originally proposed by Heldner and Edlund[36] and adopted by Levinson and Torreira in their analysis[37].

experiments, we set the lower bound of d at one second. Here we provide evidence from literature to justify this choice.

Choice of speaking window duration. In a study of gaps and overlaps in conversations, Heldner and Edlund report that on average 40% of the speaker transitions in their corpora involved overlaps (including any overlap of over 10 ms) [36]. These represent overlaps for competing for the floor. As for the duration of these overlaps, their histogram makes clear that the duration follows a mode of 50 ms in the Spoken Dutch Corpus, with a mean of 610 ms, and median of 470 ms, all under one second. In a follow-up detailed statistical analysis, Levinson and Torreira differentiate between types of overlaps: *between-overlaps*, that refer to overlaps where the floor was transfered without a silent gap between speakers; and *within-overlaps*, where overlapping speech occured in between a speaking turn and did not result in a transfer of floor [37]. Fig. 2 illustrates these types of overlaps. They used the Switchboard Corpus of English telephone conversations for their analysis, and found that only 3.8% of the signal corresponded to simultaneous speech of both speakers. This fits well with Sacks and colleagues' observations that "overwhelmingly, one party speakes at a time" [13, p. 700], for physically situated embodied social interactions. As for the duration, *between-overlaps* had a modal duration of 96 ms, a median of 205 ms, a mean of 275 ms. On the other hand, *within-overlaps* exhibited an estimated modal duration of 350 ms, a median of 389 ms, a mean of 447 ms. Further, of all the overlaps annotated, 73% involved a backchannel. These statistics indicate that choosing a lower bound for d would reasonably capture simultaneous speech that does not belong to the same floor.

As for the upper bound, a reasonable value should be at least greater than the average turn duration of a speaker. Using the same operationalization proposed in [36], Levinson and Torreira report that contiguous speech delimited by a silent interval of at least 180 ms had a mean duration of 1680 ms, and a median of 1227 ms.

V. Dataset

For this study, we use the publicly available *MatchNMingle* dataset [38] that records in-the-wild interactions of 92 people during speed-dates followed by a cocktail party. Three sessions of speed-dates and mingling were recorded in all across three days. We specifically focus on the cocktail party recordings that capture free standing conversations between participants. Fig.3 shows the video recordings from five cameras on the last day of data collection. The participants were not given a script to follow and were free to choose the participants they wished to interact with. This allows us to study naturally evolving F-formations and conversation floors in an in-the-wild setting.

Dataset Statistics. The dataset consists of a total of 92 single, heterosexual participants (46 women: 19-27 years with a mean age of 21.6 years and standard deviation of 1.9 years; and 46 men: 18-30 years with a mean age of 22.6 years and standard deviation of 2.6 years). Over 45 minutes of free mingling interaction were recorded for each of the three days; 56 minutes on the first, 50 minutes on the second, and 45 minutes on the third, respectively.

Fig. 3. Snapshots of the mingling session (Cameras 1-5) in MatchNMingle.

Annotations. The dataset provides of annotations for both F-formations and a variety of social actions. The F-formations were annotated directly from a video of the interacting participants captured from overhead cameras. The annotations were made for every second for an interval of 10 minutes per day. Each F-formation annotation provides the participant IDs for its members and the start and end times delimiting the lifetime of the F-formation. In all, 174 F-formations were annotated across 30 minutes. Of these, we filtered out those with cardinality less than four, and those for which a participant was found to leave the field of view of the cameras. This left us with 34 F-formations for our experiments.

Of the social actions annotated, we only use the Speaking Status—defined as whether or not a person is speaking. The social actions were annotated for a 30 minute segment for each day, by eight annotators hired for the task and trained by an expert. The annotations were made at the frame level using a tool that allowed for interpolation across frames. In all, 20 annotations per second for each social action are provided. Further, the speaking status is estimated from video alone, by observing lip movements or inferring from the participants' head and body gestures.

VI. Experiments

We perform two sets of experiments: first we identify the number of simultaneous speakers in an F-formation using the methodology described in Section IV, and then evaluate whether the number of members in an F-formation (cardinality) affects the speaking duration of simultaneous speakers.

Simultaneous Speakers in an F-formation. The purpose of this experiment is to evaluate the following—can we infer the existence of distinct conversation floors within an F-formation from simultaneous speaker turns? To recap, this intuition build upon early work on schisming indicating that the co-existence of two turn-taking systems is the most decisive characteristic of distinct conversation floors [13], [14]. Here we consider F-foramtions of cardinality four and above, since the possibility of distinct conversations occurs only for those F-formations.

We slide a window w of duration d across the lifetime of the F-formation in steps of one second. For every position of w,

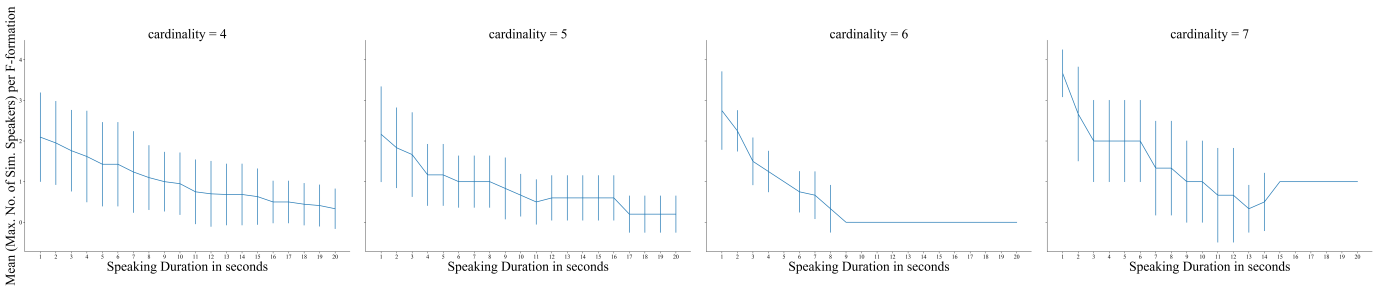

Fig. 4. Plotting the effect of varying the speaking duration threshold d on the number of simultaneous speakers per cardinality of F-formation. To aggreagate the data from each F-formation, the maximum of the number of simultaneous speakers is considered over all the sliding window positions across the lifetime of the F-formation. The y-axis plots the Mean (Maximum number of simultaneous speakers over window positions) over F-formations.

we count the number of participants with a positive speaking status for the entire duration d. We plot the maximum number of simultaneous speakers over all positions of w. Following the formulation described in Section IV, this represents the maximum number of distinct conversation floors that were observed during the life-time of the F-formation. We vary d from 1-20 seconds to guard against the possibility that the smaller values of d might capture co-narration or overlaps within the same floor. The upper bound of 20 seconds was chosen as sanity check; we expected to see very few speakers have a speaking turn that long.

The max operator was chosen to aggregate the number of simultaneous speakers across all window positions into the most conservative measure for what this experiment seeks to evaluate. A value of one for the maximum number speakers over all positions of w would indicate that only a single conversation floor existed within the F-formation. Therefore, observing values greater than one for the max metric would indicate the presence of distinct floors with more certainty than other choices of summarizing statistics.

Fig. 4 plots the mean number of distinct conversation floors per F-formation against varying values of d, per cardinality of F-formation. Cardinality here refers to the number of members in an F-formation. As a sanity check, we would expect the numbers upper-bounded by the number of people in the F-formation; at worst, every person in the F-formation speaks simultaneously to compete for the floor they are a part of. On the same note, we observe that the starting mean values all seem reasonable: about 2 for cardinalities four and five, about 3 for cardinality six, and about 4 for cardinality seven. Assuming that it is common for speakers to have at least one conversing partner, we would expect about half the number of simultaneous speakers as members in an F-formation. Our minimum choice of d was chosen to be greater than the modal duration of overlaps found in previous work [37], so it is less likely that the lower turn durations capture competing overlaps for the same floor. Moreover, at the average turn length of about two seconds observed by Levinson and Torreira [37], we observe that the maximum number of simultaneous speakers is greater than one at all cardinalities considered. This suggests that the simplifying assumption from previous research of a single conversation within an F-formation is insufficient.

We also observe a decreasing trend for the curves in Fig. 4. This seems intuitive, as it is much less likely that participants would speak for the entire duration of a window as d increases. Interestingly, there is a single example of a speaker speaking for 20 seconds in an F-formation of cardinality seven. On closer inspection, this turned out to be an error in speaking status annotation, and we manually fixed this error for subsequent analysis.

Effect of cardinality on turn duration of simultaneous speakers. Sacks, Schegloff, and Jefferson observed that there is a "pressure for minimization of turn size, distinctively operative with three or more parties" [13, p. 713]. They note that the possibility of a schism introduced by the fourth participant may influence the turn-taking system by 'spreading the turns around' if there is an interest in retaining participants in the conversation. However, they concede that this effect is equivocal, since turn distribution can also be used for encouraging schisming. In this experiment, we explore this effect and pose the question as follows: for a given speaking turn duration d, do we observe a decrease in the maximum number of conversation floors observed over an F-formation's lifetime with an increase in the cardinality of an F-formation?

Qualitatively, this corresponds to the steepness of fall-off of the curves in Fig. 4. It seems that the the curves for cardinality six and seven falloff more steeply than those for cardinalities four and five. To quantitatively test if cardinality has an effect, we fit a Generalized Linear Model (GLM) to the same data as in the previous experiment with an interaction factor between cardinality and the speaking turn duration d. Specifically, we assume the maximum number of simultaneous speakers observed over the lifetime of each F-formation, y_i to

TABLE I
GENERALIZED LINEAR MODEL REGRESSION RESULTS

	Coef (β)	Std Err	z	P>|z|
Intercept	0.0626	0.339	0.184	0.854
Turn-duration	0.0057	0.002	2.296	0.022
Cardinality	0.1869	0.072	2.603	0.009
Turn-duration:Cardinality	-0.0025	0.001	-4.543	0.000006

TABLE II
NOMINAL *P*-VALUES FOR SIX POST-HOC GLM REGRESSION COMPARISONS

Cardinality Pairs	Intercept (β_0)	d (β_1)	c (β_2)	d:c (β_3)
4-5	0.196	0.855	0.794	0.403
4-6	0.364	0.0007	0.010	0.00002*
4-7	0.697	0.428	0.030	0.009
5-6	0.079	0.0008	0.016	0.00016*
5-7	0.434	0.413	0.043	0.052
6-7	0.275	0.006	0.657	0.024

d = turn-duration, c = cardinality, d:c = interaction-factor. βs denote the corresponding regression coefficients. * denotes significance at a threshold of 0.001 after Bonferroni correction for six tests.

be realizations of independent Poisson random variables, with $Y_i \sim P(\mu_i)$ and model μ_i as follows:

$$\log(\mu_i) = \beta_0 + \beta_1 * d_i + \beta_2 * c_i + \beta_3 * d_i * c_i \qquad (1)$$

where d_i refers to the duration of the speaking window, and c_i refers to the cardinality for the ith observation. The βs refer to the regression coefficients. The GLM was fit using the *statsmodels* python package. The results of the GLM regression test are provided in Table I. We conclude that cardinality and the two-way interaction between cardinality and turn duration are statistically significant at a significance level of 0.01. Turn duration is itself significant at a significance level of 0.05.

While the previous test tells us that turn duration and cardinality are significant, we still need to perform post-hoc comparisons to ascertain the differences between the cardinalities. We fit multiple GLMs to each possible pair of cardinalities being considered and correct the corresponding p-values using the Bonferroni correction for multipe testing. Table II provides the corrected p-values for the post-hoc comparisons. From the last column, we find that cardinality and its interaction with turn-duration are significant between the cardinalities $\{4, 6\}$, and $\{5, 6\}$ at a significance level of 0.001.

One potential limitation of this analysis is the imbalance in the number of F-formations of different cardinalities. F-formations of cardinality four were the most common in the data, with reasonable number of samples to infer a pattern. We believe that the intuition of cardinality and its interaction with speaking turn duration being significant is still a sound intuition, although the statistical significance should perhaps be viewed within the context of the number of F-formations we see in the data. Fig. 5 plots the number of observations that contributed to the graphs in Fig. 4.

VII. CONCLUSION

In this study, we presented an initial exploration into unifying the spatial and temporal perspectives of a free-standing conversing group. Specifically, we proposed using simultaneous speaking turns as an indicator for the existence of distinct conversation floors. In the absence of audio data to identify the topics being discussed, our proposed metric can

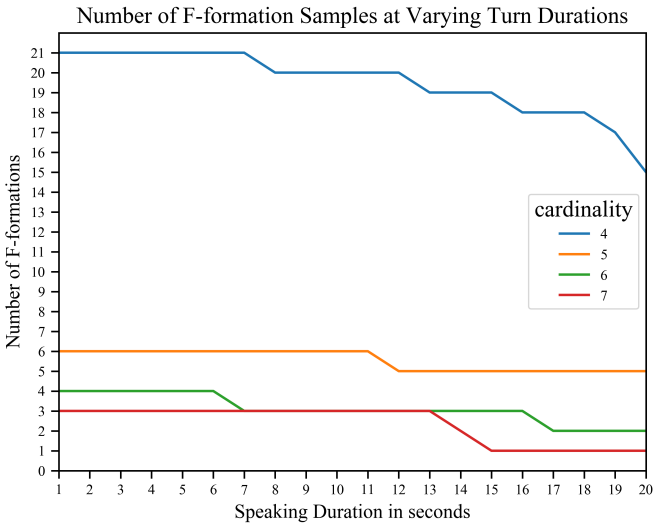

Fig. 5. Number of F-formations at different speaking turn durations.

be used to gain a deeper understanding of the conversation dynamics within an F-formation, since speaking turns can be inferred from visual or wearable-sensor data. Our experiments demonstrate that at an average turn duration of two seconds for humans [37], there is evidence of multiple conversation floors within a single F-formation. Further, we found that an increase in cardinality of an F-formation correlates with a decrease in turn duration of simultaneous speakers, specifically between F-formations of sizes {4,6}, and {5,6} in our data. A deeper analysis would be required to identify whether the differences in F-formations of cardinality six hold across datasets, with preferably more examples of F-fomrations of size six and greater. In this initial approach to the problem, our study does not account for the behaviour of the silent participants, or the evolution of turn taking dynamics within a floor. These remain promising avenues to explore for future works.

ACKNOWLEDGMENT

Chirag Raman thanks Stavros Makrodimitris, Madhumita Sushil, Giovanni Cassani, Erik B. van den Akker, and Yeshwanth Napolean for their time and thoughtfulness.

REFERENCES

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, Nov. 1, 2009.

[2] H. Huettenrauch, K. S. Eklundh, A. Green, and E. A. Topp, "Investigating spatial relationships in human-robot interaction," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Oct. 2006.

[3] M. Vazquez, A. Steinfeld, and S. E. Hudson, "Parallel detection of conversational groups of free-standing people and tracking of their lower-body orientation," IEEE, Sep. 2015.

[4] ——, "Maintaining awareness of the focus of attention of a conversation: A robot-centric reinforcement learning approach," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, IEEE, Aug. 2016.

[5] M. Vazquez, E. J. Carter, B. McDorman, J. Forlizzi, A. Steinfeld, and S. E. Hudson, "Towards robot autonomy in group conversations: Understanding the effects of body orientation and gaze," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, ACM Press, 2017.

[6] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *European Conference on Computer Vision*, Springer Berlin Heidelberg, 2012.

[7] K. Tran, A. Gala, I. Kakadiaris, and S. Shah, "Activity analysis in crowded environments using social cues for group discovery and human interaction modeling," *Pattern Recognition Letters*, Jul. 2014.

[8] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. D. Bue, G. Menegaz, and V. Murino, "Social interaction discovery by statistical analysis of f-formations," British Machine Vision Association, 2011.

[9] M. Farenzena, A. Tavano, L. Bazzani, D. Tosato, G. Paggetti, G. Menegaz, V. Murino, and M. Cristani, "Social interactions by visual focus of attention in a three-dimensional environment," *Expert Systems*, May 2013.

[10] L. Bazzani, M. Zanotto, M. Cristani, and V. Murino, "Joint individual-group modeling for tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Apr. 1, 2015.

[11] G. Groh, A. Lehmann, J. Reimers, M. R. Friess, and L. Schwarz, "Detecting social situations from interaction geometry," in *2010 IEEE Second International Conference on Social Computing*, Aug. 2010.

[12] H. Hung and B. Kröse, "Detecting f-formations as dominant sets," ACM Press, Proceedings of the 13th international conference on multimodal interfaces, 2011.

[13] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Studies in the organization of conversational interaction*, 1974.

[14] C. Goodwin, "Forgetfulness as an interactive resource," *Social Psychology Quarterly*, 1987.

[15] M. M. Egbert, "Schisming: The collaborative transformation from a single conversation to multiple conversations," *Research on Language & Social Interaction*, Jan. 1997.

[16] A. Kendon, *Conducting interaction: patterns of behavior in focused encounters*. Cambridge University Press, 1990.

[17] C. Edelsky, "Who's got the floor?" *Language in Society*, 1981.

[18] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani, "Multi-scale f-formation discovery for group detection," IEEE International Conference on Image Processing, Sep. 2013, pp. 3547–3551, ISBN: 978-1-4799-2341-0.

[19] F. Setti, C. Russell, C. Bassetti, and M. Cristani, "F-formation detection: Individuating free-standing conversational groups in images," *PLOS ONE*, May 21, 2015.

[20] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe, "Analyzing free-standing conversational groups: A multimodal approach," ACM Press, Proceedings of the 23rd ACM international conference on Multimedia, 2015.

[21] S. Vascon, E. Z. Mequanint, M. Cristani, H. Hung, M. Pelillo, and V. Murino, "Detecting conversational groups in images and sequences: A robust game-theoretic approach," *Computer Vision and Image Understanding*, Feb. 2016.

[22] E. Ricci, J. Varadarajan, R. Subramanian, S. R. Bulo, N. Ahuja, and O. Lanz, "Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos," in *2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Dec. 2015.

[23] H. Hung, D. B. Jayagopi, S. Ba, J.-M. Odobez, and D. Gatica-Perez, "Investigating automatic dominance estimation in groups from visual attention and speaking activity," ACM Press, International conference on Multimodal interfaces, 2008.

[24] L. Zhang and H. Hung, "Beyond f-formations: Determining social involvement in free standing conversing groups from static images," IEEE Proceedings of the Conference on Computer Vision and Pattern Recognition, Jun. 2016.

[25] ——, "On social involvement in mingling scenarios: Detecting associates of f-formations in still images," *IEEE Transactions on Affective Computing*, 2018.

[26] C. Oertel, K. A. Funes Mora, J. Gustafson, and J.-M. Odobez, "Deciphering the silent participant: On the use of audio-visual cues for the classification of listener categories in group discussions," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction - ICMI '15*, ACM Press, 2015.

[27] C. Oertel and G. Salvi, "A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue," in *Proceedings of the 15th ACM on International conference on multimodal interaction - ICMI '13*, ACM Press, 2013.

[28] D. Bohus and E. Horvitz, "Managing human-robot engagement with forecasts and... *um* ... hesitations," in *Proceedings of the 16th International Conference on Multimodal Interaction - ICMI '14*, ACM Press, 2014.

[29] A. Pentland, A. Madan, and J. Gips, "Perception of social interest," in *Proceedings of the 5th international conference on devlopment and learning ICDL 2006*, Department of Psychological and Brain Sciences, Indiana University, 2006.

[30] H. Hung and D. Gatica-Perez, "Estimating cohesion in small groups using audio-visual nonverbal behavior," *IEEE Transactions on Multimedia*, Oct. 2010.

[31] E. Goffman, *Behavior in public places: notes on the social organization of gatherings*, 1. paperback ed., 24. printing. New York, NY: The Free Press, 1966.

[32] D. McNeill, *Language and Gesture*. Cambridge University Press, 2000.

[33] H. Hung and S. O. Ba, "Speech/non-speech detection in meetings from automatically extracted low resolution visual features," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, 2010.

[34] E. Gedik and H. Hung, "Speaking status detection from body movements using transductive parameter transfer," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, ACM, 2016.

[35] ——, "Personalised models for speech detection from body movements using transductive parameter transfer," *Personal and Ubiquitous Computing*, Aug. 1, 2017.

[36] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, Oct. 1, 2010.

[37] S. C. Levinson and F. Torreira, "Timing in turn-taking and its implications for processing models of language," *Frontiers in Psychology*, Jun. 12, 2015.

[38] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. v. d. Meij, and H. Hung, "The MatchNMingle dataset: A novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates," *IEEE Transactions on Affective Computing*, 2018.