Predicting and Regulating Participation Equality in Human-robot Conversations: Effects of Age and Gender

Gabriel Skantze
KTH Speech Music and Hearing
Lindstedtsv. 24
10044 Stockholm, Sweden
+46-8-7907874
skantze@kth.se

ABSTRACT
In this paper, we investigate participation equality, in terms of speaking time, between users in multi-party human-robot conversations. We analyse a dataset where pairs of users (540 in total) interact with a conversational robot exhibited at a technical museum. The data encompass a wide range of different users in terms of age (adults/children) and gender (male/female), in different combinations. Overall, the analysis indicates that demographically heterogeneous pairs are more imbalanced, especially pairs of adults and children, where children are less prone to self-select in the turn-taking. The analysis also indicates that it is possible for the robot to reduce the imbalance by addressing the least dominant user and asking directed questions. However, for children to respond, it is important to seek mutual gaze and switch addressee often. Finally, we show that it is possible to predict the imbalance at an early stage in the interaction – in order to increase the participation equality as early as possible – and that knowledge about the users' age and gender helps in this prediction.

Keywords
Human-robot interaction; Turn-taking; Speech; Multi-party interactions; Age; Gender; Children; Participation equality

1. INTRODUCTION
Conversational robots are becoming increasingly common in areas such as education and elderly care. The physical appearance of the robot allows it to not only convey both verbal and non-verbal signals (in the form of prosody, gaze and gestures), but also to take part in interactions where the physical situation is of importance, such as discussions about objects in the shared space. This also allows robots to take part in multi-party interactions with several users, something which is much harder with, for example, conversational agents on 2D displays [19]. Compared to a dyadic setting, where only one user interacts with the robot, multi-party interaction allows users to not only talk to the robot, but also to each other. In an educational setting, this could potentially increase the engagement and learning outcomes. However, multi-party interaction also allows for more imbalanced participation. If one of the interlocutors is more dominant, it is possible that the others will not have a chance to contribute as much. Therefore, it is very important for a robot in such settings to be aware of the *participation equality* between the speakers, that is, the relative amount of speaking time for the participants. If a robot is able to predict any imbalance in speaking time, and if it can take actions to regulate the turn-taking and balance the interaction, this would be an argument for involving robots in conversational settings. Unlike humans, such a robot could use a completely unbiased model, purely based on statistics, to make these decisions.

HRI '17, March 06 - 09, 2017, Vienna, Austria

ACM 978-1-4503-4336-7/17/03…$15.00
DOI: http://dx.doi.org/10.1145/2909824.3020210

In this paper, we investigate three research questions. Firstly, there are several studies that find important differences in turn-taking behaviour between adults and children [9, 16], and between males and females [1, 6, 7]. Such differences could lead to different levels of participation equality. Thus, we want to study the effects of these variables on turn-taking behaviour and participation equality in multi-party human-robot interaction. Secondly, we want to study the effects of the robot's behaviour on the users' turn-taking behaviour. To what extent can the robot regulate the turn-taking by actively selecting the next speaker? Related to this, we also want to understand whether different user groups respond differently to the robot's turn-taking regulation signals. Thirdly, we want to study whether it is possible to predict the imbalance in participation at an early stage, and whether knowledge about age and gender help in this prediction. Such a prediction could potentially allow the robot to start shaping the turn allocation as early as possible in the dialogue.

The data we analyse comes from a system that was exhibited at the Swedish National Museum of Science and Technology for nine days [26]. As illustrated in Figure 1, two visitors at a time played a collaborative game together with the robot head Furhat (see Figure 2), in a fully automated setup. On the touch table between the players, a set of cards are shown. The two visitors and the robot are given the task of sorting the cards according to some criterion. For example, the task could be to sort a set of inventions in the order they were invented, or a set of animals based on how fast they can run. This is a collaborative game, which means that the users have to discuss the solution together with the robot. However, the robot does not have perfect knowledge about the solution. Instead, the robot's arguments are motivated by a randomized belief model. This means that the users have to determine whether they should trust the robot's belief or not, just like they have to do with each other. Thus, the robot's role in the interaction is similar to that of the users – it is not intended to be perceived as a tutor. During the nine days of the exhibition, we recorded hundreds of interactions with users from the general public, including adults and children. Since the robot's turn-taking behaviour was randomly selected for each turn, it is possible to study the effects on the users' behaviours.

In a previous analysis of the recorded data [26], we have shown that the robot's verbal and non-verbal behaviours – including filled pauses, breath and gaze – affect the user's turn-taking behaviours. However, we did not analyse whether these behaviours differed between adults and children, or males and females, as done here.

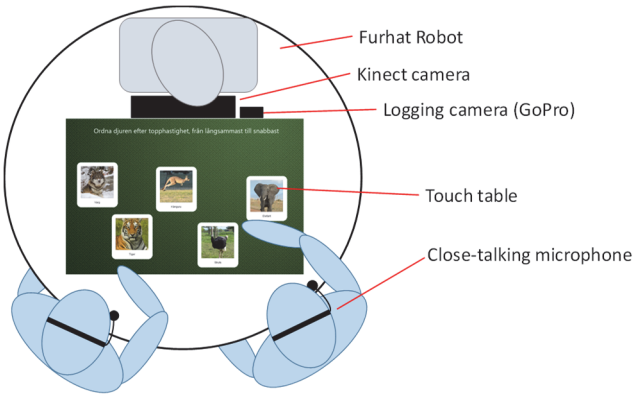

Figure 1: The setup used in the museum.

Figure 2: Left: Partial view from the logging camera. Right: The Furhat robot head.

2. BACKGROUND

In this section, we will provide a background on turn-taking in multi-party interaction, review the literature on effects of age and gender on turn-taking, and discuss related work on how to detect and regulate participation equality in human-robot interaction.

2.1 Turn-taking in Spoken Interaction

Many human social activities require some kind of turn-taking protocol, which regulates the order in which the different actions are supposed to take place, and by whom. This is obvious when, for example, playing a game of chess (where the protocol is very simple), but it also applies to spoken interaction. Since it is difficult to speak and listen at the same time, speakers in dialogue have to somehow coordinate who is currently speaking and who is listening. In a seminal article, Sacks et al. [24] described a protocol for this, which is schematically illustrated in Figure 3. In their view, speakers try to minimize the amount of gaps and overlaps between turns. At certain points in the speech, there are *Transition-Relevance Places* (TRPs), where a shift in turn could potentially take place. At these places, the current speaker may select a next speaker, who then 'has the right and is obliged' to take the next turn to speak, whereas no other participants are supposed to do so. If the current speaker does not select a next speaker, any participant has the opportunity to 'self-select', or the current speaker may continue. In a two-party (dyadic) conversation, the selection of the next speaker is trivial, but if there are several speakers (multi-party interaction), the current speaker typically gazes at the selected next speaker, or may use some other indicator, such as the person's name or a pointing gesture.

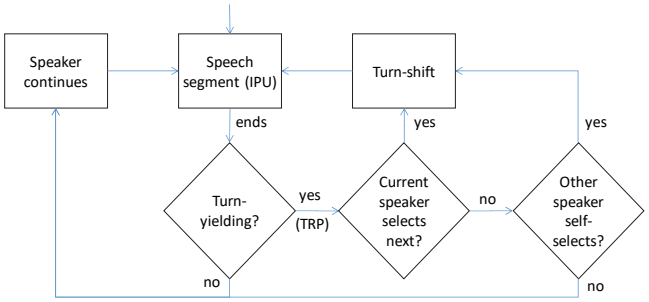

Figure 3: The 'standard model' of turn-taking as proposed by Sacks et al. (1974)

Traditionally, spoken dialogue systems have rested on a very simplistic model of turn-taking, where a certain amount of silence (say 700 ms) is used as an indicator that the user has stopped speaking, and that the turn is yielded to the system. The problem with this model is that turn-shifts are often supposed to be much more rapid than this, and that pauses within a turn may often be longer. Thus, the system will sometimes appear to give sluggish responses, and sometimes interrupt the user. A more accurate model would make a more continuous decision, trying to identify whether the user is *yielding* or *holding* the turn. Several studies have found that speakers use their voice and face to give *turn-holding* and *turn-yielding cues* [8, 11, 15]. For example, an incomplete syntactic clause or a filled pause (such as 'ehm') typically indicates that the speaker is not yielding the turn. Prosodically, a rising or falling pitch at the end of a segment tend to be turn-yielding, whereas a flat pitch is turn-holding. The intensity of the voice tends to be lower when yielding the turn, and the duration of the last phoneme tends to be shorter. Gaze has also been found to be an important cue – speakers tend to not look at the addressee during an utterance, but then shift the gaze towards the addressee when yielding the turn [14]. Studies on human-human dialogue have shown that the more turn-yielding cues are presented together, the more likely it is that the other speaker will take the turn [8, 11, 15]. Thus, one way of deciding whether the robot should take the turn or not after a speech segment, is to use machine learning to combine these different cues [12, 17].

It is important to stress that the diagram shown in Figure 3 is highly simplistic, and humans often 'violate' this schema. For example, speakers do not always take the turn even though they have been 'selected', and people may sometimes speak before the current speaker's speech segment is completed, resulting in overlapping speech. Indeed, the turn-taking model of Sacks et al. [24] has been challenged by other researchers, who argue that speakers do not always try to minimize gaps and overlaps, but that the criteria for successful interaction is highly dependent on the kind of interaction taking place [23]. In this view, overlaps can be a sign of engagement, and it is possible that conversational systems should not necessarily always avoid overlaps. Common signals, that often give rise to overlapping speech, are *backchannels* – short utterances (such as 'mhm' or 'aha'), which the listener provides to show continued attention [29].

2.2 Effects of Age and Gender on Turn-taking

Managing turn-taking is one of the first interactional skills humans have to learn. This learning begins by dyadic interaction with the

caregiver, who is responsible for regulating most of the turn-taking [9]. Later on, children also learn how to claim the floor in multi-party interaction, a skill that is mastered around the age of six. Even after that, children continue to learn how to take turns, and reduce gaps and overlaps. Whereas adults often take turns with very short gaps (around 250ms), children's gaps are typically longer – in some studies average gap length has been measured at 1.5-2 seconds [9]. This gap length shortens with age, as the child learns to better pick up turn-yielding cues, project the interlocutor's turn ending, and plan their own response [9]. Researchers have also compared child-child interaction with child-adult interaction [16]. Typically, children conversing with adults allow the adult to regulate the interaction, and then pick up these regulatory behaviours and employ them when talking to other children. Thus, we can expect interaction between children to be more balanced than interaction between adults and children.

The differences between male and female turn-taking behaviours has been the subject of several studies. In same-sex settings, men tend to follow the standard model by Sacks et al. (minimizing gaps and overlaps), to a larger extent, whereas a much larger amount of overlap has been found in conversations between women [6]. These overlaps should not be regarded as competitive, as most often they are the result of women helping to complete each other's utterances and giving backchannels. The conversational style between women has therefore been characterized 'cooperative', whereas conversation between men is more 'competitive' [7]. Despite being more talkative in general, in mixed-sex settings, women have been found to talk less than men [7], and men seem to interrupt women more than the other way around [1].

Examining the interaction between age and gender, some studies have investigated the mechanisms by which linguistic gender differences develop. Studies on children have shown that young girls seem to be more talkative and fluent than boys (both to their mothers and to other children), but that boys tend to dominate mixed conversations at a relatively early age [7]. In child-adult interactions, fathers seem to interrupt children more than mothers, and both parents interrupt girls more than boys [7].

2.3 Turn Regulation and Participation Equality in Human-robot Interaction

Several previous studies have found that conversational agents can shape the turn-taking and assignment of participant roles in multi-party interaction [5, 19, 21]. The most common signal is to use gaze to actively select the next speaker. For this to work, it must be clear who the agent is looking at. A problem with animated agents on 2D displays is that it is impossible for the user to see exactly where the agent is looking, a problem typically referred to as the 'Mona Lisa effect' [19]. This is because the agent and user don't share the same physical space. Thus, in a multi-party setting, this means that the agent cannot establish exclusive mutual gaze with one of the users, and in a situated interaction the object that is the target of the gaze cannot be inferred. However, studies have shown that if an animated face is back-projected on a 3D mask, as done with the Furhat robot head used here, the turn-yielding accuracy is improved [19], and humans can utilize the robot's gaze to disambiguate references to objects to achieve joint attention [20, 25]. In a study on turn-taking in multi-party child-robot interaction, [18] examined the likelihood of children answering quiz questions, depending on where and how the robot directed its gaze. However, we are not aware of any studies that systematically compare how children and adults respond to turn-taking signals from an artificial agent in the same interactional setting.

If indeed a robot can regulate the turn-taking in multi-party interaction, this opens up the possibility of allowing the robot to balance the level of participation between the participants. One example of such an attempt is [2], where a virtual agent was used to balance the contributions from a group of children playing a game. Such balancing also requires a measure of dominance or participation equality. In [28], a model for classifying children's social dominance in group interactions is presented. Based on manual annotation of social dominance, a model was trained. The main predictor that turned out to be useful was the children's gaze towards the robot. However, it should be noted that there was no two-way spoken interaction between the robot and the children, and thus is cannot be regarded as a conversation. Another example is [22], where a dominance estimation model (based on gaze and speech) was used to decide how the robot should regulate the interaction using gaze.

In this study we will also investigate how the robot can regulate the turn-taking and balance the participation. However, this study is different from most previous studies on this subject. Firstly, the robot has a similar role as the speakers, thus it does not have a clear 'function' (i.e., acting as a tutor or quiz host). As we have seen in our previous analyses, the addressee in this kind of interaction is often not so easy to determine [26]. Many utterances in more conversational settings are not targeted towards a specific person, but rather as open statements or questions. Thus, most turn-shifts will be done using self-selection. Secondly, the discussion in our task necessarily involves references to objects in the physical surroundings. In such settings, speakers also naturally look at these objects. This has been shown to clearly affect the extent to which humans otherwise gaze at each other to yield the turn [3, 13].

In summary, several studies have examined both turn-taking regulation and participation equality in human-robot conversations, as well as gender and age effects on turn-taking. However, we are not aware of any previous studies that have systematically investigated these issues in combination.

3. HUMAN-ROBOT INTERACTION SETTING

The interactional setting that is under investigation here, and was briefly described in the Introduction, is illustrated in Figure 1. Two users are seated at a large table with a multi-touch screen, opposite the Furhat robot head (see Figure 2), which has an animated face back-projected on a translucent mask, as well as a mechanical pan-tilt neck [20]. This allows Furhat to direct the gaze using a combination of head and eye movements. The animated face allows for very accurate and expressive lip movements, facial gestures and gaze, which have been shown to be easy for users to read [20]. The synthetic voice (unit selection) is also complemented by non-verbal expressions, such as sighs, breathing, filled pauses and different types of backchannels. In our experience, Furhat does not typically give rise to an uncanny valley effect, despite the human-like appearance, possibly because of its slightly cartoonish design.

Both users were wearing unidirectional headset microphones, which allowed for the recording of two separate good quality audio streams (given the noisy setting in the museum). A Kinect camera (v2) was used to track the location and rotation of the users' heads, and a GoPro camera was used for logging.

The interaction starts when two users are seated and press a 'Start' button on the touch screen. Furhat initiates the interaction by asking them for their names. Then five cards are shown on the table and Furhat describes the sorting criterion, after which the discussion starts. When the task has been discussed for some time, a button is shown on the table that can be pressed to reveal the solution. Furhat

then comments on the solution, comparing it with his own belief (admitting mistakes or pointing out that they should have listened to him). After that, the players can play another round if they wish. An example interaction is shown in Table 1[1]. After having played two rounds, the users are encouraged by Furhat to allow others to play. The system only manipulates the cards when switching between tasks – only the players can move the cards during the discussion.

Table 1: Example interaction (translated from Swedish)

U-1	I wonder which one is the fastest [looking at cards]
U-2	I think this one is fastest [touching a lion card], what do you think? [looking at robot]
R	I'm not sure about this, but I think the lion is the fastest animal [looking at cards]
U-1	Okay [moving the lion]
R	Now it looks better
U-2	Yeah… How about the zebra? [looking at robot]
R	I think the zebra is slower than the horse. What do you think? [looking at U-1]
U-1	I agree
U-2	I'm not sure, the zebra has to be fast to escape the lion…
R	mhm

The system was implemented using the open source framework IrisTK[2] [23], which provides an event-driven modularized architecture for real-time multi-modal dialog processing. The speech recognition was done with two parallel cloud-based large vocabulary speech recognition modules, which allowed Furhat to understand the users even when they were talking simultaneously. Since the speech is of a very conversational nature, the word-error-rate is typically very high (around 60%). However, our previous analysis [26] has showed that this is mainly due to errors in function words, and less due to errors in content words. In any case, given this high rate of error, modelling such dialog on a deeper semantic and pragmatic level would not be feasible. However, our main goal here is to create a believable dialog agent that can be used as a test-bed for studying turn-taking. Our approach has therefore been to exploit the redundancy offered by the multi-modal nature of the interaction, where for example the movement of the cards can be used together with the noisy speech recognition to infer which objects are being discussed. For a more detailed technical description of how this was accomplished and how the system was implemented, please refer to [26].

4. DATA PROCESSING AND ANALYSIS

Interactions were recorded during the 9 days the system was exhibited at the Swedish National Museum of Science and Technology. For this analysis, we have filtered out complete interactions where the same pairs played at least one round together, and where the logs were complete. Even though the interaction continued if one of the players was lost in the Kinect face tracking, we have only analysed instances where the Kinect tracked both players throughout the whole interaction. This selection was done partly because we base some of our analyses on the Kinect data, but also because failure in Kinect face tracking may indicate that the users were not properly seated at the table and taking part in the interaction. Since we are doing automatic analysis, we want to make sure that their speech patterns can be interpreted as reactions to the ongoing conversation and to Furhat's behaviour. This selection amounts to a total of 292 interactions (with 584 users). The average length of these interactions was 4.5 minutes. Since people freely walked up to the system and interacted with it, we were not able to collect data on actual age and gender. Therefore, we have relied on manual annotation of age and gender based on the video recordings, by an annotator experienced with children. We are of course aware of the uncertainty in this annotation, but we should stress that we have only based our analysis on the rough distinction between children and adults, and not the exact age. The result of the annotation in terms of distribution is shown in Figure 4.

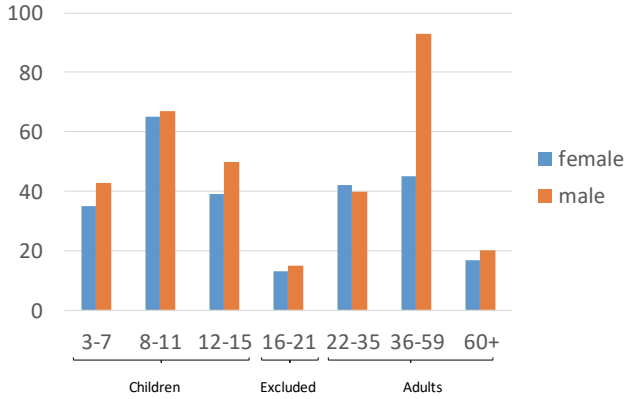

Figure 4: Number of users in different age groups.

As is evident, the distribution of gender is fairly even across age groups (overall 43.5% female), with the exception of middle aged visitors, for whom males are more common. To make the distinction between children and adults clearer, we exclude from our analysis all interactions where one of the participants was judged to be between 16 and 21 years old. The relatively few users in this range could perhaps be explained by the technical museum setting, which seems to attract adults in company with children, but perhaps fewer people in their late adolescence or early adulthood. For our analysis, we treat users below this span as children, and users above it as adults. This leaves us with 270 interactions and 540 individual users. To verify the age and gender coding, 20% of the interactions were coded by a second annotator. The agreement with the first annotator (in terms of child/adult and male/female) was 100%. The number of combinations of demographic categories in user pairs is shown in the rightmost column in Figure 5. As is evident, even though all combinations are represented, they are not evenly distributed; there are, for example, 33 pairs with two adult males, but only eight pairs with two adult females.

In this study, we have chosen to base all further analysis on completely automatic processing of the interaction logs, partly due to the large amount of annotation that would otherwise be required, but also to avoid subjective indicators and increase the reproducibility of the analysis. We mainly analyse speech activity, which is extracted by applying voice activity detection on the audio streams (with a 200 ms end-of-speech threshold). This way, we can measure which speaker responds when Furhat says something, the total speaking time and overlaps between the speakers. We measure participation equality in terms of speaking time, which we think is a

[1] A video of a complete interaction with visualization can be seen at https://www.youtube.com/watch?v=5fhjuGu3d0I and some footage of different users at https://www.youtube.com/watch?v=PtMOI0WHeE4

[2] http://www.iristk.net

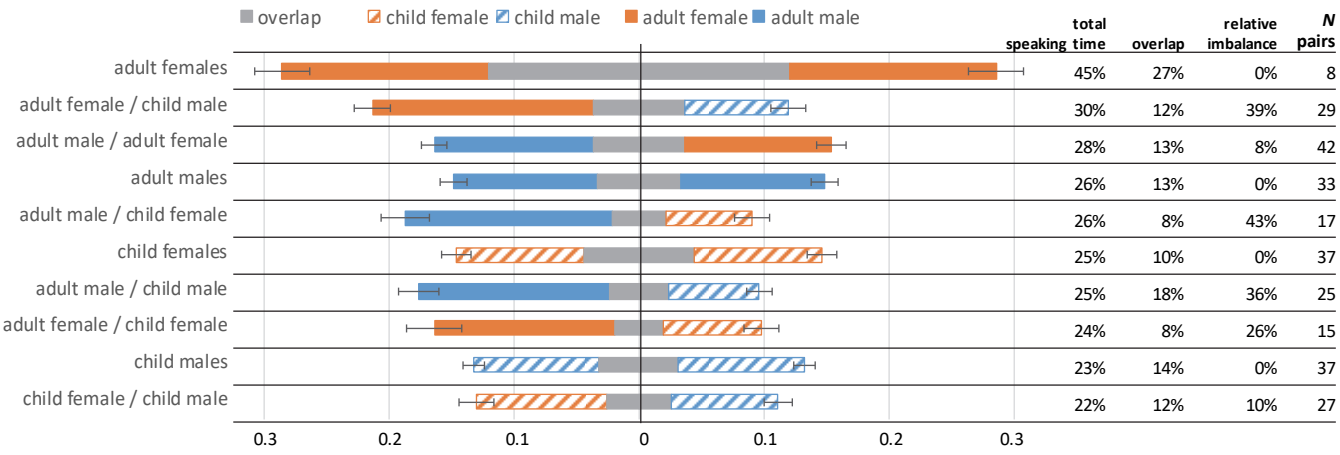

Figure 5: Average speaking time (with SE) and overlaps in different user type pairs. Total speaking time shows a proportion of the whole interaction where at least one of the users is speaking. Overlap is relative to total speaking time.

more objective measure than, for example, 'social dominance' used in similar studies [22, 28], but the reader should be aware that these are not exactly the same thing. We also analyse the users' head pose movements, which are registered by the Kinect, as a proxy for gaze. This way, we can measure whether users are looking at Furhat, given the same threshold that was used in the online system. Although this cannot be used to capture quick glances or track more precise gaze targets, previous studies have found head pose to be a fairly reliable indicator of visual focus of attention in multi-party interaction, given that the targets are clearly separated [4]. We do not analyse interruptions, even though this has been the subject of many studies (especially on gender effects), since we have not found a method for extracting these automatically and reliably. It is tempting to use overlaps as an indicator for this, but overlaps do not have to indicate interruptions (as discussed in the Background), and interruptions can occur without overlap [10]. Furthermore, when analysing the logs, it was clear that overlaps were mostly of a collaborative nature, mainly due to the nature of the task.

5. RESULTS

We report here the results from the data analysis. An alpha level of .05 was used for all statistical tests.

5.1 Speaking Time and Overlaps

We start by making a basic analysis of the total speaking time (as a proportion of the total time of the interaction), to see if there are any effects of age or gender. The average speaking time across all interactions for different user types is shown in Figure 6. As is evident, adult females speak most (M = 19.3%, SD = 9.2), followed by adult males (M = 16.2%, SD = 8.1), child females (M = 13.0%, SD = 8.8) and child males (M = 12.0%, SD = 6.9). A 2x2 ANOVA with age and gender as between-subjects factors revealed a main effect of age, $F(1, 536) = 52.85$, $p < .001$, $\eta_p^2 = 0.09$ and gender, $F(1, 536) = 8.28$, $p = .004$, $\eta_p^2 = 0.015$, but no interaction effect. Thus, on average, females speak more than males, and adults speak more than children. These results are in line with what was found in the literature (2.2).

However, this analysis only indicates the overall average. As the literature suggests, speaking time is likely to depend on the constellation of user types. The average amount of speaking time, and overlaps, for different pairs are shown in Figure 5. As is evident, the speaking time for the same user type varies a lot, depending on the pairing. Overall, children have much less speaking time when

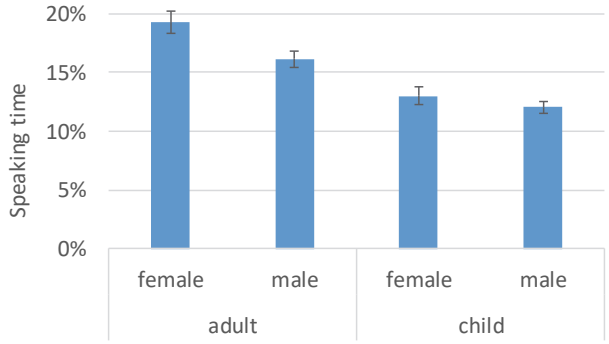

Figure 6: Average speaking time (with SE), as a proportion of the whole interaction, for different user types.

paired with an adult than when paired with another child, which is in line with what has been found in related work. Correspondingly, adults have much more speaking time when paired with a child, than when paired with another adult. From our impressions, adults in these settings often take on a protective role with the child, trying to 'explain' what is happening, regulating the interaction, and to some extent ignoring the robot. At the same time, this leads (most likely without intention) to less participation from the child.

Although adult females have the overall highest speaking time in general (Figure 6), the combination of two adult females clearly reinforces this effect, and leads to a strikingly high total proportion of 45% speaking time (compared to, for example, two adult males with 26% total speaking time). They also have a very large amount of overlap (27%), which is twice as much as adult males (13%). Interestingly, this is also in line with what has been found in the literature [6].

We also investigated the amount of gazing towards Furhat, as measured by the head pose data from the Kinect. We measured both the proportion of gazing towards Furhat when he was speaking (which indicates to what extent the users attended to what Furhat said), and the amount of gazing towards Furhat when the user was speaking (which indicates to what extent the users addressed Furhat). We did not find any gender effects, but age had an effect, which is shown in Figure 7. A two-way ANOVA was conducted, with age and peer age as variables, which showed a statistically significant interaction effect on proportion of gaze towards Furhat, both while Furhat was

speaking ($F(1, 536) = 8.65$, $p < .01$) and while the user was speaking ($F(1, 536) = 8.3$, $p < .01$). Thus, it seems like children attend less to the robot in general, and that adults in adult-child pairs attend more to the child. This strengthens the interpretation that adults in such settings take on a regulatory role and spend attentional resources on monitoring their children. Furthermore, since the children seem to be attending more to the cards or the human peer, they may be less receptive to turn-taking cues from the robot.

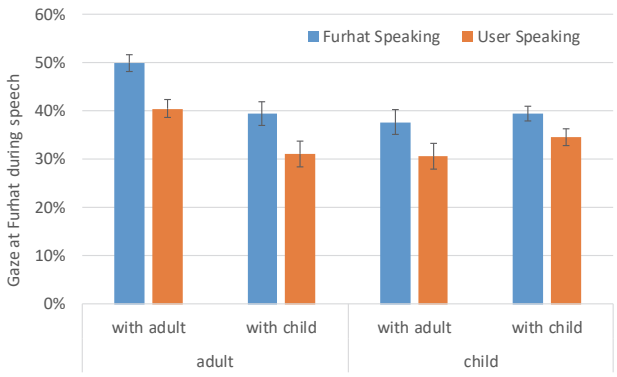

Figure 7: Average gazing towards Furhat (with SE), while user is speaking, and while Furhat is speaking.

5.2 Imbalance

It is important to note that that Figure 5 only indicates the overall imbalance between user types, which makes the homogenous pairs look perfectly balanced. In reality, these interactions are of course often imbalanced. To make a more systematic comparison of participation equality in different combinations of user types, we have defined an *Imbalance score* that was calculated for each interaction according to the following equation, where ST_{max} denotes the total speaking time for the speaker that speaks the most, and ST_{min} denotes the total speaking time for the other speaker.

$$Imbalance\ (absolute) = 1 - \frac{ST_{min}}{ST_{max}}$$

This function gives an absolute score between 0 and 1, indicating the amount of imbalance between the speakers (where 0 means perfect balance, and 0.5 that one speaker speaks twice as much as the other). The average of these scores for different pairs of user types are shown in Figure 8 as an *absolute* imbalance score. For the demographically heterogeneous pairs, we also calculated a *relative* imbalance score, which is the same as the absolute score, but with a sign (negative or positive) that indicates the direction of the imbalance between the two types of users. A one sample t-test was conducted to see which of the relative scores deviated from 0. These results are also shown in Figure 8 (where * denotes significant deviation using Bonferroni correction, $\alpha = .05 / 6 = .0083$). Note that a high absolute imbalance still allows for a relative imbalance closer to zero, if the imbalance goes in different directions between pairs and thereby cancel each other out when averaging.

As is evident from the figure, the imbalance between different user types varies a lot. Interestingly, the top and the bottom of the figure follows the patterns found in the literature review in 2.2 on turn-taking and gender. The most imbalanced pairs are child females with adult males. It is also interesting to note that while both adult males and females with child males have a high relative imbalance, there is no significant relative imbalance between adult females and child females. The most balanced pairs are adult females, which also had the most speaking time and amount of overlap (Figure 5). This might perhaps strengthen the characterization of female-female talk as 'cooperative', as argued by [7]. However, contrary to what was reported in [7], there is no significant imbalance in total speaking time between males and females when two adults or two children interact.

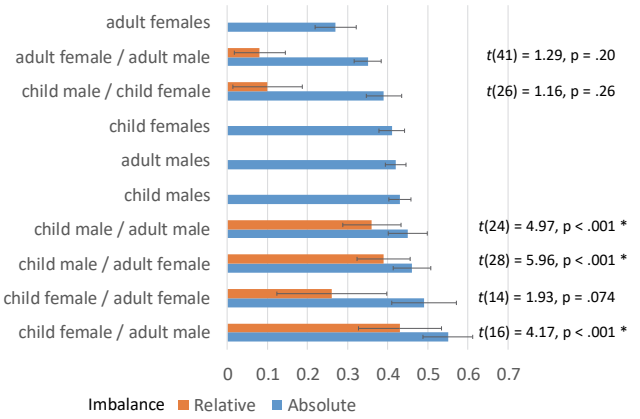

Figure 8: Average imbalance of speaking time (with SE) in different pairs of users (dominant speaker to the right). Test-statistics for relative imbalance > 0.

5.3 Regulating Turn-taking

It is important to note that there might be different causes behind the imbalances in speaking time. For example, it could be because one of the users speak in longer utterances, or it could be because one of the users is more prone to grab the turn. To address this question, we investigated the imbalance in turn-taking after the turn was yielded by the robot. This was calculated using the same formula that was used for calculating the imbalance of speaking time above, but instead we counted the number of times each speaker responded within three seconds (note that this makes it possible for none of the users, or both, to respond). Overall, this yielded a turn-taking imbalance of 0.367, which is fairly close to the overall speaking time imbalance of 0.422. Thus, the imbalance can in large part be explained by the turn-taking. To better understand the mechanisms behind the imbalance, and how it can be mitigated, we now turn to a more detailed analysis of the turn-taking behaviours of the users.

At the museum, Furhat's turn-yielding behaviour was randomly selected for each turn, both in terms of addressee and speech act (question or statement). This gives us an opportunity to study the causal effect of these behaviours on the relative imbalance in which user speaks next. For each interaction, we defined the user that ended up speaking the most as the 'dominant' user, and the other the 'non-dominant' user. In Table 2, we show the relative imbalance in four different conditions:

1. The robot asking a question to both participants, rapidly switching the gaze between them (both users may self-select).
2. The robot making a statement and looking at the cards (both users may self-select).
3. The robot looking at the dominant user, asking a question (robot selects next speaker).
4. Same as previous, but for the non-dominant user.

Table 2: Effects of different turn-regulating strategies. A positive relative imbalance indicates that the dominant user responded.

Strategy	N	Dominant responds	Non-dominant responds	Relative imbalance
1. Question, attend both	788	67.4%	51.0%	0.243
2. Statement, attend table	1092	55.8%	27.9%	0.499
3. Question, attend dominant user	313	81.5%	13.7%	0.831
4. Question, attend non-dominant user	339	28.0%	67.8%	-0.587

As is evident from Table 2, when both users can self-select, the dominant user is in general more likely to grab the turn. It is interesting to compare condition 1 and 2, since condition 1 creates a larger expectation that at least one of the users should respond, and also provides a clearer way of responding, whereas condition 2 leaves the continuation more open. As is evident from the table, the imbalance is larger for condition 2. Thus, non-dominant users seem to be less prone to self-select in these cases, and such turn-shifts are therefore more likely to increase the imbalance. Looking at condition 3 and 4, it is clear that a direct question from Furhat to one of the users will make that user much more likely to respond, even when the non-dominant user is addressed. Thus, it should be possible for the robot to decrease the imbalance by addressing this user more often.

By analysing direct questions (condition 3 and 4) in more detail, we can also gain a better understanding of how this 'next-speaker-selection accuracy' (i.e., the likelihood that the targeted user responds within 3 seconds) can be improved. First, we investigated the effect of including the addressee's name in the question (which was done randomly). For questions of the type 'What do you think (Peter)?', the accuracy increased from 66.9% to 83.9% if the addressee's name was included ($\chi^2(1) = 5.77$, $p = 0.016$). Thus, getting the users' name is valuable for improving the balance.

In previous studies we have found that when the robot keeps addressing the user who was speaking most recently, that user is less likely to respond again, compared to when the robot switches to the other user [26, 27]. Thus, by switching addressee, the next-speaker selection accuracy may increase. Since children are most often the non-dominant speaker, we also wanted to investigate whether this effect is different for children and adults. In the game, a typical example of keeping the addressee is when the robot follows up a comment by one of the users with a question like 'why do you think so?'. An alternative is to switch addressee by asking the other user 'and what do you think?'. The effect of these two strategies, split between children and adults, is shown in Figure 9. A chi-square test revealed a significant difference between keeping and switching addressee for children ($\chi^2(1) = 8.48$, $p = .0036$), but not for adults ($\chi^2(1) = 0.11$, $p = .74$). Thus, repeated questions towards children are less likely to reduce the imbalance. When a non-dominant child has not spoken recently, the robot has a better chance of reducing the imbalance by addressing that child with a question.

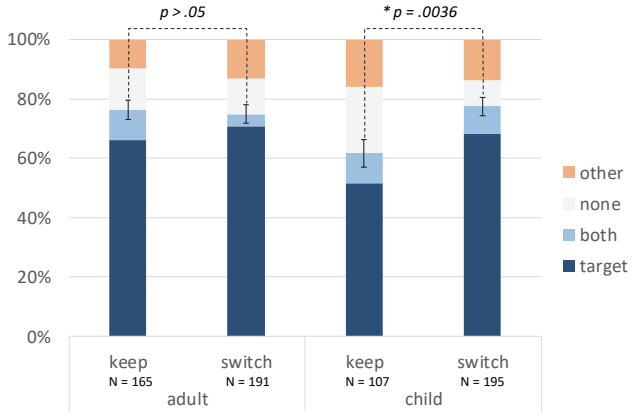

Figure 9: Next-speaker-selection accuracy (with SE) for cases where the robot targeted a particular user with a question, depending on age and whether the previous speaker is addressed.

Next, we examined the effect of mutual gaze, since we have also found this to be an important factor in previous studies [26]. Figure 10 shows the next-speaker-selection accuracy in targeted questions, depending on whether the addressee is looking back at the robot at the time the turn is yielded (as registered by the Kinect). Again, a chi-square test revealed a significant difference for children ($\chi^2(1) = 8.62$, $p = .0033$), but not for adults ($\chi^2(1) = 2.71$, $p = .10$). Thus, seeking mutual gaze is especially important when yielding the turn to children.

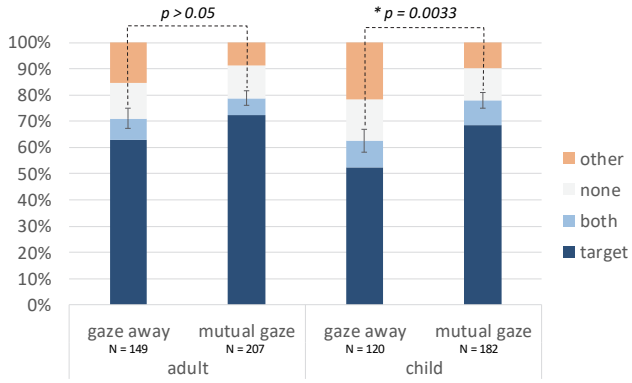

Figure 10: Next-speaker-selection accuracy (with SE) for cases where the robot targeted a particular user with a question, depending on age and mutual gaze

5.4 Predicting Imbalance

Given the possibility for the robot to regulate the turn-taking and thereby shift the balance in the interactions, an interesting question is how early on in the interaction the system could predict the participation equality. The earlier this could be done, the more proactive the robot could be. To explore this, we evaluated a multiple linear regression model for predicting the final relative imbalance. To explore how soon into the interaction the imbalance can be predicted, we evaluated the model at different time steps *T*, using the available features up to that point. We included the following sets of features that we deemed to be indicative:

- IMBALANCE: The current relative imbalance up to point *T*, as defined in 5.2.
- DEMOGRAPHICS: Information about the user types. The combination of age and gender of the two users were encoded into six dummy variables.
- GAZE: The head pose movements of the users (registered by the Kinect) as a proxy for their gaze behaviour. As discussed in Section 2.3, gaze has been shown in related studies to be highly indicative of social dominance. We measured the proportion of gaze by both users towards Furhat, (a) in total, (b) while Furhat was speaking, and (c) while the user was speaking, up to point *T*.

A multiple linear regression model was built at 15 seconds intervals (up to 90 seconds into the dialogue), and the correlation coefficient (R) was used for evaluation of the predictive power. The features from IMBALANCE and DEMOGRAPHICS were both useful in the prediction, as shown in Figure 11, whereas GAZE only gave a negligible improvement at 15 seconds (an increase of the R-value of 0.01, not shown in the figure). This is contrary to what has been found in related attempts at modelling social dominance [22, 28], where gaze has been an important factor.

Not surprisingly, IMBALANCE is an increasingly useful feature (since the final imbalance is the target function). However, the correlation between the current and final imbalance is not very strong early on in the dialogue. Interestingly, we can see that information about the mix of age and gender clearly improves the prediction, especially at an early stage.

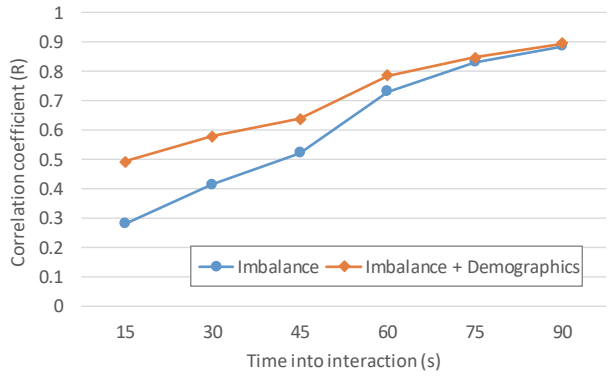

Figure 11: Prediction of imbalance at different time steps, with and without demographic information.

6. CONCLUSIONS AND DISCUSSION

Taken together, the results indicate that it is possible to predict participation imbalance at a fairly early stage, and that it is possible for the robot to reduce the imbalance, by asking directed questions to the non-dominant user. Overall, demographically heterogeneous pairs are more imbalanced in terms of participation, but it should be stressed that homogenous pairs also have a high level of imbalance and might need turn-regulation. The highest levels of imbalance are found in pairs of adults and children, where children are less prone to take the turn. By avoiding yielding the turn with open statements too often, and posing more directed questions towards the child, the balance can be improved. This effect can be reinforced further if the robot also switches addressee often, seeks mutual gaze with the addressee and, possibly, uses the name of the addressee.

We have not found any previous studies on participation equality in multi-party interaction, where all combinations of age (adult/child) and gender (male/female) have been systematically studied in the same conversational setting. Thus, we think the present study is unique in this regard. Since we have employed completely automated measures, it should be possible to replicate and verify the findings in other settings. Furthermore, using a robot in this setting adds a level of control: we can make sure the same question is executed in exactly the same way across interactions, and that the selection of turn-taking signals is chosen randomly.

Much of the findings in terms of participation equality are in line with the patterns reported in the literature: Women speak more than men [7], and adults speak more than children [9]. Children speak less when paired with an adult than with another child [16]. The most imbalanced pairs are those with adults males and child females [7]. Women have much more total speaking time, proportion of overlaps and equality in speaking time when paired with another woman, compared to when paired with a man [6, 7]. However, we have not found any significant relative imbalance when a man and a woman, or a boy and a girl, are paired. This is contrary to what is reported in the literature, in which males have been found to talk more than females in mixed-gender settings [7]. One should of course be aware that our study was performed in a Swedish context, and that the results may not necessarily be generalizable to other cultures. Furthermore, the setting is of course very specific: a technical museum where visitors play a game together with a robot.

The generally high levels of overlaps (especially in conversations involving two women) pose a challenge for current conversational systems, which most often rely on a simplified no-gap-no-overlap model of turn-taking. Furthermore, depending on the microphone setting, overlapping speech may be hard to pick up. Before these issues are solved, it would be interesting to explore how the robot could also regulate turn-taking in order to avoid overlaps.

The study we have presented here was done 'in the wild', where visitors of the museum voluntarily walked up to the robot and played the game. Although this has given us the opportunity to collect and analyse data from a diverse population, one should be aware that we did not randomly assign subjects in pairs. As such, the study constitutes a natural experiment, and we had no control over the number of subjects in each demographic group. It should also be pointed out that although the subjects to a large extent conformed to the system's turn-regulating signals on the turn-level, and each such turn should in itself reduce the imbalance, we do not know whether there are any long-lasting effects on the imbalance, i.e., whether these interventions will encourage non-dominant subjects to start self-selecting more. Therefore, we encourage future studies involving more controlled experiments that can address these questions. Finally, one should also be aware that adjusting for participation equality might not always be desirable. For example, in adult-child interactions, the adult may take on a tutoring role, which should not necessarily call for a correction of speaking time imbalance. Thus, there are other factors that should be taken into account for making this decision.

7. ACKNOWLEDGEMENTS

This work is supported by the Swedish research council (VR) project *Coordination of Attention and Turn-taking in Situated Interaction* and the EU Horizon 2020 project *BabyRobot*. Thanks to Martin Johansson and Jonas Beskow for their contributions to setting up the system, and thanks to everyone helping out with the exhibition: Saeed Dabbaghchian, Björn Granström, Joakim Gustafson, Raveesh Meena, Kalin Stefanov and Preben Wik. The author would also like to thank the anonymous reviewers for their very helpful remarks.

8. REFERENCES

[1] Anderson, K.J. and Leaper, C. 1998. Meta-analyses of gender effects on conversational interruption: Who, what, when, where, and how. *Sex Roles*. 39, 3/4 (1998), 225–252.

[2] Andrist, S., Leite, I. and Lehman, J. 2013. Fun and fair: influencing turn-taking in a multi-party game with a virtual agent. *Proceedings of the 12th International Conference on Interaction Design and Children - IDC '13*. (2013), 352–355.

[3] Argyle, M. and Graham, J.A. 1976. The central Europe experiment: Looking at persons and looking at objects. *Environmental Psychology and Nonverbal Behavior*. 1, 1 (1976), 6–16.

[4] Ba, S.O. and Odobez, J.-M. 2009. Recognizing visual focus of attention from head pose in natural meetings. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*. 39, 1 (2009), 16–33.

[5] Bohus, D. and Horvitz, E. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. *Proceedings of ICMI* (Beijing, China, 2010).

[6] Coates, J. 1994. No Gap, Lots of Overlap; Turn-taking Patterns in the Talk of Women Friends. *Researching language and literacy in social context: A reader*. Multilingual Matters.

[7] Coates, J. 2004. *Women, Men and Language: A Sociolinguistic Account of Gender Differences in Language, 3rd edition*. Routledge.

[8] Duncan, S. 1972. Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology*. 23, 2 (1972), 283–292.

[9] Ervin-Tripp, S.M. 1979. Children's verbal turn-taking. *Developmental pragmatics*. E. Ochs and B. Schieffelin, eds. Academic Press. 391–414.

[10] Gravano, A. and Hirschberg, J. 2012. A Corpus-Based Study of Interruptions in Spoken Dialogue. *Interspeech-2012*. (2012).

[11] Gravano, A. and Hirschberg, J. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*. 25, 3 (2011), 601–634.

[12] Johansson, M. and Skantze, G. 2015. Opportunities and Obligations to Take Turns in Collaborative Multi-Party Human-Robot Interaction. *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)* (2015), 402.

[13] Kawahara, T., Iwatate, T. and Takanashi, K. 2012. Prediction of Turn-Taking by Combining Prosodic and Eye-Gaze Information in Poster Conversations. *Interspeech 2012* (2012).

[14] Kendon, A. 1967. Some functions of gaze direction in social interaction. *Acta Psychologica*. 26, (1967), 22–63.

[15] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A. and Den, Y. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese Map Task dialogs. *Language and Speech*. 41, (1998), 295–321.

[16] Martinez, M.A. 1987. Dialogues among Children and between Children and Their Mothers. *Child Development*. 58, 4 (1987), 1035–1043.

[17] Meena, R., Skantze, G. and Gustafson, J. 2014. Data-driven models for timing feedback responses in a Map Task dialogue system. *Computer Speech and Language*. 28, 4 (2014), 903–922.

[18] Al Moubayed, S. and Lehman, J. 2015. Regulating Turn-Taking in Multi-child Spoken Interaction. *Intelligent Virtual Agents* (2015).

[19] Al Moubayed, S. and Skantze, G. 2011. Turn-taking Control Using Gaze in Multiparty Human-Computer Dialogue: Effects of 2D and 3D Displays. *Proceedings of the International Conference on Audio-Visual Speech Processing 2011*. (2011), 99–102.

[20] Al Moubayed, S., Skantze, G. and Beskow, J. 2013. The furhat back-projected humanoid head-lip reading, gaze and multi-party interaction. *International Journal of Humanoid Robotics*. 10, 1 (2013).

[21] Mutlu, B., Kanda, T., Forlizzi, J., Hodgins, J. and Ishiguro, H. 2012. Conversational Gaze Mechanisms for Humanlike Robots. *ACM Trans. Interact. Intell. Syst.* 1, 2 (2012), 12:1-12:33.

[22] Nakano, Y.I., Yoshino, T., Yatsushiro, M. and Takase, Y. 2015. Generating Robot Gaze on the Basis of Participation Roles and Dominance Estimation in Multiparty Interaction. *ACM Transactions on Interactive Intelligent Systems*. 5, 4 (2015), 1–23.

[23] O'Connell, D., Kowal, S. and Kaltenbacher, E. 1990. Turn-taking: A critical analysis of the research tradition. *Journal of psycholinguistic*. (1990).

[24] Sacks, H., Schegloff, E. and Jefferson, G. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*. 50, (Dec. 1974), 696–735.

[25] Skantze, G., Hjalmarsson, A. and Oertel, C. 2014. Turn-taking, feedback and joint attention in situated human-robot interaction. *Speech Communication*. 65, (2014), 50–66.

[26] Skantze, G., Johansson, M. and Beskow, J. 2015. Exploring turn-taking cues in multi-party human-robot discussions about objects. *ICMI 2015 - Proceedings of the 2015 ACM International Conference on Multimodal Interaction* (2015).

[27] Skantze, G., Moubayed, S. Al, Gustafson, J., Beskow, J. and Granström, B. 2012. Furhat at Robotville: A Robot Head Harvesting the Thoughts of the Public through Multi-party Dialogue. *Proceedings of IVA-RCVA* (Santa Cruz, CA, USA, 2012).

[28] Strohkorb, S., Leite, I., Warren, N. and Scassellati, B. 2015. Classification of Children's Social Dominance in Group Interactions with Robots. *ICMI* (2015), 227–234.

[29] Yngve, V.H. 1970. On getting a word in edgewise. *Papers from the sixth regional meeting of the Chicago Linguistic Society* (Chicago, Apr. 1970), 567–578.