Opportunities and Obligations to Take Turns in Collaborative Multi-Party Human-Robot Interaction

Martin Johansson and Gabriel Skantze
Department of Speech Music and Hearing, KTH
Stockholm, Sweden
{vhmj,skantze}@kth.se

Abstract

In this paper we present a data-driven model for detecting opportunities and obligations for a robot to take turns in multi-party discussions about objects. The data used for the model was collected in a public setting, where the robot head Furhat played a collaborative card sorting game together with two users. The model makes a combined detection of addressee and turn-yielding cues, using multi-modal data from voice activity, syntax, prosody, head pose, movement of cards, and dialogue context. The best result for a binary decision is achieved when several modalities are combined, giving a weighted F_1 score of 0.876 on data from a previously unseen interaction, using only automatically extractable features.

1 Introduction

Robots of the future are envisioned to help people perform tasks, not only as mere tools, but as autonomous agents interacting and solving problems together with humans. Such interaction will be characterised by two important features that need to be taken into account when modelling the spoken interaction. Firstly, the robot should be able to solve problems together with several humans (and possibly other robots) at the same time, which means that we need to model *multi-party* interaction. Secondly, joint problem solving is in many cases *situated*, which means that the spoken discourse will involve references to, and manipulation of, objects in the shared physical space. When speaking about objects, humans typically pay attention to these objects and gaze at them. Also, placing or moving an object can be regarded as a communicative act in itself (Clark, 2005). To solve the task efficiently, interlocutors need to coordinate their attention, resulting in so-called joint attention (Clark & Marshall, 1981).

These characteristics of human-robot interaction pose many challenges for spoken dialogue systems. In this paper, we address the problem of turn-taking, which is a central problem for all spoken dialogue systems, but which is especially challenging when several interlocutors are involved. In multi-party interaction, the system does not only have to determine when a speaker yields the turn, but also whether it is yielded to the system or to someone else. This becomes even more problematic when the discussion involves objects in a shared physical space. For example, an obvious signal that humans use for yielding the turn in a face-to-face setting is to gaze at the next speaker (Vertegaal et al., 2001). However, in situated interaction, where the gaze is also used to pay attention to the objects which are under discussion, it is not obvious how this shared resource is used. While modelling all these aspects of the interaction is indeed challenging, the multi-modal nature of human-robot interaction also has the promise of offering redundant information that the system can utilize, thereby possibly increasing the robustness of the system (Vinyals et al., 2012).

The aim of this study is to develop a data-driven model that can be used by the system to decide when to take the turn and not. While there are many previous studies that have built such models based on human-human (Koiso et al., 1998; Morency et al., 2008) or human-machine interaction (Raux & Eskenazi, 2008; Skantze & Schlangen, 2009; Bohus & Horvitz, 2011; Meena et al., 2014), we are not aware of any previous studies that investigate multi-party human-robot discussions about objects.

The system that we build the model for, and use data from, is a collaborative game that was exhibited at the Swedish National Museum of

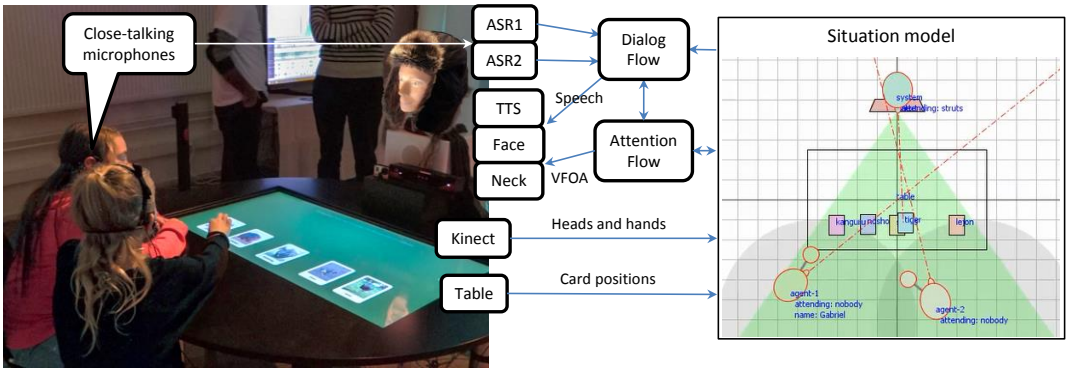

Figure 1: A schematic illustration of the dialogue system setting and architecture

Science and Technology in November 15-23, 2014. As can be seen in Figure 1, two visitors at a time could play a collaborative game together with the robot head Furhat (Al Moubayed et al., 2013). On the touch table between the players, a set of cards are shown. The two visitors and Furhat 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 visitors have to discuss the solution together with Furhat. As we have discussed in previous work (Johansson et al., 2013), we think that the symmetry of the interaction is especially interesting from a turn-taking perspective. The setting also provides a wide range of multi-modal features that can be exploited: voice activity, syntax, prosody, head pose, movement of cards, and dialogue context[1].

The paper is organized as follows: In Section 2 we present and discuss related work, in Section 3 we describe the system and data annotation in more detail, in Section 4 we present the performance of the different machine learning algorithms and features sets, and in Section 5 we end with conclusions and a discussion of the results.

2 Background

2.1 Turn-taking in dialogue systems

Numerous studies have investigated how humans synchronize turn-taking in dialogue. In a seminal study, Duncan (1972) showed how speakers use prosody, syntax and gestures to signal whether the speaker wants to hold the turn or yield it to the interlocutor. For example, flat final pitch, syntactic incompleteness and filled pauses are strong cues to turn hold. In his analysis, Duncan found that as more turn yielding cues are presented together, the likelihood that the listener will try to take the turn increases. Later studies on human-human interaction have presented more thorough statistical analyses of turn-yielding and turn-holding cues (Koiso et al., 1998; Gravano & Hirschberg, 2011). Typically, for speech-only interaction, syntactic and semantic completeness is found to be the strongest cue, but prosody can also be informative, especially if other cues are not available. In face-to-face interaction, gaze has been found to be a strong turn-taking cue. Kendon (1967) found that the speaker gazes away from the listener during longer utterances, and then gazes at the listener as a turn-yielding cue near the end of the utterance.

Contrary to this sophisticated combination of cues for managing turn-taking, dialogue systems have traditionally only used a fixed silence threshold after which the system responds. While this model simplifies processing, it fails to account for many aspects of human-human interaction such as hesitations, turn-taking with very short gaps or brief overlaps and backchannels in the middle of utterances (Heldner & Edlund, 2010). More advanced models for turn-taking have been presented, where the system interprets syntactic and prosodic cues to make continuous decisions on when to take the turn or give feedback, resulting in both faster response time and less interruptions (Raux & Eskenazi, 2008; Skantze & Schlangen, 2009; Meena et al., 2014).

2.2 Turn-taking in multi-party interaction

Multi-party interaction differs from dyadic interaction in several ways (Traum & Rickel, 2001). First, in a dyadic interaction there are only two different roles that the speakers can have: speaker and listener. In multi-party interaction, humans may take on many different roles, such as side participant, overhearer and bystander (Mutlu et al., 2012). Second, in dyadic interaction, it is

[1] A video of the interaction can be seen at https://www.youtube.com/watch?v=5fhjuGu3d0I

always clear who is to speak next at turn shifts. In multi-party interaction, this has to be coordinated somehow. The most obvious signal is to use gaze to select the next speaker (Vertegaal et al., 2001). Thus, for multi-party interaction between a robot and several users, gaze is a valuable feature for detecting the addressee. Gaze tracking is however not trivial to utilize in many practical settings, since they typically have a limited in field-of-view, or (if head worn) are too invasive. In addition, they are not very robust to blinking or occlusion, and typically need calibration. Many systems therefore rely on head pose tracking, which is a simpler and more robust approach, but which cannot capture quick glances or track more precise gaze targets. However, previous studies have found head pose to be a fairly reliable indicator for gaze in multi-party interaction, given that the targets are clearly separated (Katzenmaier et al., 2004; Stiefelhagen & Zhu, 2002; Ba & Odobez, 2009). In addition to head pose, there are also studies which show that the addressee detection in human-machine interaction can be improved by also considering the speech signal, as humans typically talk differently to the machine compared to other humans (Shriberg et al., 2013). Vinyals et al. (2012) present an approach where the addressee detection is done using a large set of multi-modal features.

In situated interaction, speakers also naturally look at the objects which are under discussion. The speaker's gaze can therefore be used by the listener as a cue to the speaker's current focus of attention. This has been shown to clearly affect the extent to which humans otherwise gaze at each other to yield the turn. Argyle & Graham (1976) studied dyadic interactions involving additional targets for visual attention. Objects relevant to the task at hand were found to attract visual attention at the expense of the other subject. In a study on modelling turn-taking in three-party poster conversations, Kawahara et al. (2012) found that the participants almost always looked at the shared poster. Also, in most studies on human-robot interaction, the robot has a clear "function", and it is therefore obvious that the user is either addressing the machine or another human. However, in a previous study on multi-party human-robot discussion about objects (Johansson et al., 2013), which had a task that is very similar to the one used here, we found that the addressee of utterances is not so easy to determine. Sometimes, a question might be posed directly to the robot, which then results in an *obligation* to take the turn. But many times, utterances in multi-party discussions are not targeted towards a specific person, but rather to both interlocutors, resulting in an *opportunity* to take the turn.

The approach taken in this study is therefore to combine the turn taking and addressee detection into one decision: *Should the system take the turn or not?*, and then allow a gradual answer from a clear "no" (0) to a clear "yes" (1). If the answer is 0, it could be because a speaker is holding the turn, or that a question was clearly posed to someone else. If the answer is 1, the system is obliged to respond, most likely because one of the users has asked a question directly to the robot. But in many cases, the answer could be somewhere in between, indicating an opportunity to respond. In future work, we plan to use such a score together with a utility function in a decision-theoretic framework (Bohus & Horvitz, 2011). Thus, if the system has something urgent to say, it could do so even in a non-optimal location, whereas if what it has to say is not so important, this would require an obligation in order to respond

3 Data collection and annotation

3.1 System description

As described in the introduction, we use data from a multi-party human-robot interaction game that was exhibited in a public setting. The system was implemented using the open source dialogue system framework IrisTK (Skantze & Al Moubayed, 2012) and is schematically illustrated in Figure 1. The visitors are interacting with the Furhat robot head (Al Moubayed et al., 2013), which has an animated face back-projected on a translucent mask, as well as a mechanical neck, which allows Furhat to signal his focus of attention using a combination of head pose and eye-gaze. A Kinect camera (V2) is used to track the location and rotation of the two users' heads, as well as their hands. This data, together with the position of the five cards on the touch table are sent to a Situation model, which maintains a 3D representation of the situation. Two behaviour controllers based on the Harel statechart mechanism offered by IrisTK run in parallel: The Dialog Flow and the Attention Flow. The Attention Flow keeps Furhat's attention to a specified target (a user or a card), even when the target is moving, by consulting the Situation model. The 3D position of the target is then transformed into neck and gaze movement of Furhat (again taking Furhat's position in the 3D space into account).

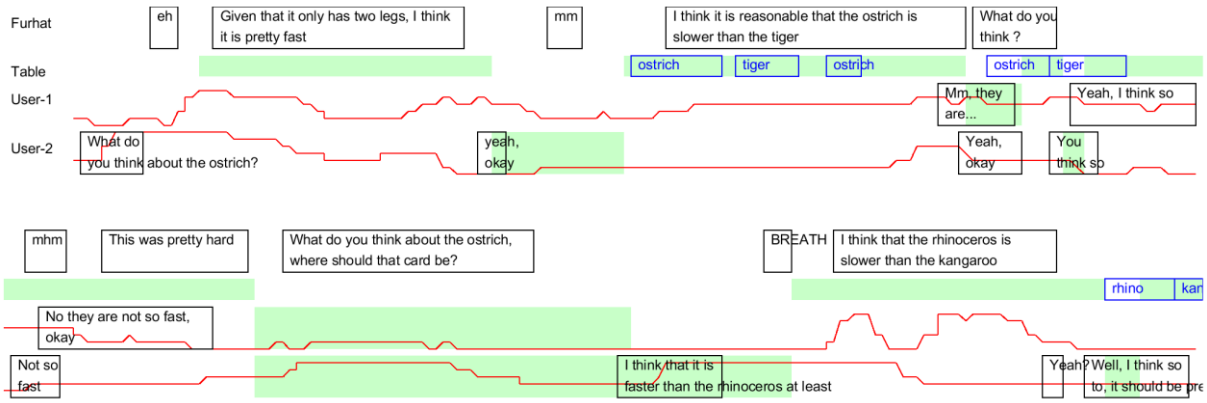

Figure 2: Dialogue fragment from an interaction (translated from Swedish). The shaded (green) track shows where Furhat's attention is directed. Card movements are illustrated in blue. Users' head poses are illustrated with red plots, where a high y-value means the angular distance towards Furhat is small.

This, together with the 3D design of Furhat, makes it possible to maintain exclusive mutual gaze with the users, and to let them infer the target of Furhat's gaze when directed towards the cards, in order to maintain joint attention (Skantze et al., 2014). Although the system can be configured to use the array microphone in the Kinect camera, we used close talking microphones in the museum. The main motivation for this is that the Kinect array microphone cannot separate the sound sources from the two users and we wanted to be able to run parallel speech recognizers for both users in order to capture overlapping speech (for both online and offline analysis). The speech recognition is done with two parallel cloud-based large vocabulary speech recognizers, Nuance NDEV mobile[2], which allows Furhat to understand the users even when they are talking simultaneously.

The Dialogue Flow module orchestrates the spoken interaction, based on input from the speech recognizers, together with events from the Situation model (such as cards being moved, or someone leaving or entering the interaction). The head pose of the users is used to make a simple decision of whether Furhat is being addressed. The game is collaborative, which means that the visitors have to discuss the solution together with Furhat. However, Furhat does not have perfect knowledge about the solution. Instead, Furhat's behaviour is motivated by a randomized belief model. This means that visitors have to determine whether they should trust Furhat's belief or not, just like they have to do with each other. Thus, Furhat's role in the interaction is similar to that of the visitors, as opposed to for example a tutor role which is often given to robots in similar settings. An excerpt from an interaction is shown in Figure 2, illustrating both clear turn changes and turns with overlapping speech.

3.2 Collected Data

The dialog system was exhibited at the Swedish National Museum of Science and Technology, in November 15-23, 2014. During the 9 days the system was exhibited, we recorded data from 373 interactions with the system, with an average length of 4.5 minutes. The dataset contains mixed ages: both adults playing with each other (40%), children playing with adults (27%), and children playing with each other (33%). For the present study, 9 dialogues were selected for training and tuning the turn-taking model, and one dialogue was selected for final evaluation and for verification of the annotation scheme.

3.3 Data Annotation

In order to build a supervised machine learning model for detecting turn-taking cues, we need some kind of ground truth. There have been different approaches to deriving the ground truth in previous studies. In studies of human-human interaction, the behaviour of the other interlocutor is typically used as a ground truth (Koiso et al., 1998; Morency et al., 2008). The problem with this approach is that much turn-taking behaviour is optional, and these studies typically report a relatively poor accuracy (albeit better than baseline). Also, it is not clear to what extent they can be applied to human-machine interaction.

In this paper we follow the approach taken in Meena et al. (2014) – to manually annotate appropriate places to take the turn. Although this is quite labour intensive, we think that this is the best method to obtain a consistent ground truth

[2] http://dragonmobile.nuancemobiledeveloper.com/

about potential turn-taking locations. To this end we used turn-taking decisions from one annotator (one of the authors), thus building models of one specific human's behaviour rather than an average of multiple humans' behaviour. However, as described further down, we have also evaluated the amount of agreement between this annotator with another annotator on the evaluation set.

Similarly to most previous studies on turn-taking reported above, we treat the end of Inter-Pausal Units (IPUs) as potential turn-taking locations. Each channel of the recorded audio was first echo-cancelled and then automatically segmented into IPUs, using an energy-based Voice Activity Detector (VAD), with a maximum of 200ms internal silence. The logged utterances from the dialogue system were then added as a third track of IPUs. A decision point was defined after every segmented user IPU where the system had not been speaking in the last three seconds. Figure 3 presents an example of sequences of subject IPUs with the location of decision points overlaid. Note that we also include locations where the other speaker is still speaking (1 in the figure), since the other speaker might for example be talking to herself while the first speaker asks Furhat something.

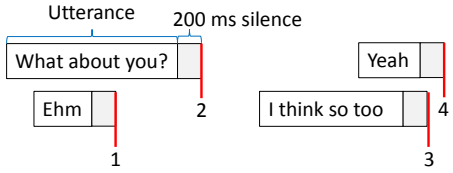

Figure 3: Four numbered decision points

A set of 688 decision points from the 9 selected dialogues were annotated for turn-taking decisions. The annotator was presented with five seconds of audio and video taken from the robot's point of view. A turn-taking decision was then annotated on a continuous scale ranging from "Absolutely don't take the turn" to "Must take the turn". The scale was visually divided into four equally wide classes to guide the annotator. The first section "**Don't**" (35% of annotated instances) represents instances where it would be inappropriate to take the turn, for example because the other interlocutor was either the addressee or currently speaking. The next section, "**If needed**" (19%), covers cases where it is not really appropriate, but possible if the system has a clear reason for saying something, while "**Good**" (21%) covers instances where it would not be inappropriate to take the turn. The final section, "**Obliged**" (25%), represents instances where it would be inappropriate not to take the turn, for example when the system clearly was the sole addressee.

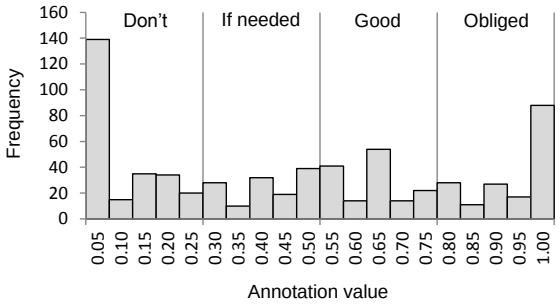

Figure 4: Histogram of annotated decisions on a scale from 0 (must not take turn) to 1 (must take turn)

The distribution of the decisions, illustrated in Figure 4, indicates a fairly even distribution across the x-axis, but with higher frequencies of annotations at the extremes of the scale.

For verification of the annotation scheme and final evaluation, we annotated a second set of 43 decision points from a tenth dialogue using both the original annotator and a second annotator. The inter-annotator agreement for the four classes was good, K_w=0.772 (Cohen's Kappa, equal weights), and neither annotator classified any decision point as "Don't" when the other had classified it as "Obliged".

4 Results

For this analysis we will first focus on the classes "Don't" and "Obliged" to make a binary turn-taking decision in section 4.1. We will then switch focus to the full range of annotations and predict turn-taking decisions numerically on a scale in section 4.2. Finally we evaluate the resulting models in 4.3 using annotations from a second annotator.

4.1 Binary Decision – Don't vs. Obliged

For every turn-taking decision the outcome will eventually be either to take the turn or to not. For the annotated classes "Don't" and "Obliged", there is a one-to-one mapping between the class and the correct turn-taking decisions. The classes "If needed" and "Good" on the other hand encode optional behaviour; both the decision to take the turn and to not take the turn can be considered correct at the same time, an opportunity to take the turn and not an obligation.

In this section we therefore build a model to distinguish between "Don't" and "Obliged". For this we explore the RIPPER (JRIP), Support Vector Machine (SVM) with linear kernel function and Multilayer Perceptron (MLP) classifiers

in the WEKA toolkit (Hall et al., 2009), using the default parameters. All results in this section are based on 10-fold cross-validation. For statistical analysis, we have used two-tailed tests and chosen an alpha level of 0.05.

Features	JRIP	SVM	MLP
VAD *	0.727	0.734	0.723
Head pose *	0.690	0.724	0.709
Cards *	0.717	0.526	0.671
Prosody *	0.648	0.574	0.649
POS *	0.602	0.630	0.634
System DA	0.506	0.506	0.500

Table 1: Weighted F_1 score of the feature categories used in isolation. Results significantly better than baseline are marked with *.

Baseline

The majority-class baseline, always providing the classification "Don't", yields a weighted F_1 score of 0.432.

Voice Activity Features

A very basic feature to consult before taking the turn is to listen if **anyone is speaking**. Using only this feature the weighted F_1 score reaches 0.734, significantly better than the baseline. In addition, we also use features to add context: The amount of time each of the system and the other interlocutor has been **quiet**, and the **length of the last turn,** defined as a sequence of IPUs without IPUs from other speakers in-between, as well as **length of the last IPU** for the system and each of the two interlocutors. Thus, the total of VAD features is 9. The "anyone speaking" feature is the single feature yielding the highest weighted F_1 score, performing on par with the combination of all VAD features (Table 1).

Prosodic Features

As prosodic features, we used final pitch and energy. A pitch tracker based on the Yin algorithm (de Cheveigné & Kawahara, 2002) was used to estimate the F_0 at a rate of 100 frames per second. The F_0 values were then transformed to log scale and z-normalized for each user. For each IPU, the last voiced frame was identified and then regions of **200ms** and **500ms** ending in this frame were selected. For these different regions, we calculated the **mean**, **maximum**, **standard deviation** and **slope** of the normalized F_0 values. To calculate the slope, we took the average pitch of the second half of the region minus the average of the first half. Additionally, we calculated the maximum and standard deviation of the normalized F_0 values over the full IPU. We also Z-normalized the energy of the voiced frames and then calculated the **maximum energy** for the 200ms and 500ms regions and the full IPU. Thus, we used 13 prosodic features in total. Using MLP on the combination of all features yielded the highest weighted F_1 score (0.649, see Table 1). The features based on pitch were more useful than the ones based on energy.

Syntactic Features

Syntax has been shown to be a strong turn-yielding cue in previous studies (Koiso et al., 1998; Meena et al., 2014). For example, hesitations can occur in the middle of syntactic constructions, whereas turn ends are typically syntactically complete. In previous studies, the **part-of-speech** (POS) of the last two words has been shown to be a useful feature. Thus, we use the POS of the last two words in an IPU as a bigram. The POS tags were automatically extracted using Stagger (Östling, 2013) based on results from cloud-based large vocabulary speech recognizers, Nuance NDEV mobile ASR, as an automated system would need to rely on ASR. Despite a word error rate (WER) of 63.1% (*SD*=39.0) for the recognized IPUs, the generated POS feature performed significantly better than the baseline (Table 1). However, the increase is not very high compared to previous studies. This could both be due to the relatively high WER, but also due to the fact that syntax in itself does not indicate the addressee of the utterance.

Head Pose Features

Unlike the other feature categories, head pose can be used to both yield the turn and to select the next speaker, and is therefore expected to be a strong feature for the current task. We represent the interlocutors' head poses in terms of **angular distance** between the direction of the interlocutor's head and the robot's head. The angular distance is made available as **absolute angular distance** as well as signed **vertical and horizontal** angular distance separately. The sign of the horizontal distance is adjusted to account for the mirrored position of the two interlocutors. This representation allows the system to infer if someone is looking at the system (low absolute distance), towards the table (negative vertical distance) or towards the other interlocutor (high horizontal distance).

The head pose features are generated separately for the speaker ending the IPU and the other interlocutor as well as in two composite versions

representing the joint (maximum) and disjoint (minimum) distance. The features are generated both at the end of the speech in the IPU and at the time of the decision point. Thus, there are a total of 24 features available for estimating visual focus of attention. Sorting the individual features from highest weighted F_1 score to lowest, we get the following top four groups in order: Last speaker (end of speech), last speaker (decision), disjoint (decision) and then joint (end of speech). As expected, the use of head pose gives a significantly better result than the baseline (Table 1).

Card Movement

The activity of the game table is represented in terms of card movement activity via 3 feature types. Note that we only know if a card is being moved, but not by whom. The first feature type is the **duration** of ongoing card movement. If no card is being moved at the moment, the value is set to 0. The second feature type is the duration of the most recently completed card movement. The final feature type is the **time passed** since the last movement of any card. These features are generated for two points in time; the end of the IPU relating to the decision point and the time when the decision is to be made. Thus, there are 6 card movement features in total. As can be seen in Table 1, this feature category alone performs significantly better than baseline, which is a bit surprising, given that the card movements are not necessarily linked to speech production and turn-taking.

The System's Previous Dialogue Act

To represent the dialogue context, we used the last system dialogue act as a feature. Whereas this feature gave a significant improvement in the data-driven models for dyadic turn-taking presented in Meena et al. (2014), it is the only feature category here that does not perform significantly better than the baseline (Table 1). The overall low performance of this feature could be due to the nature of multi-party dialogue, where the system doesn't necessarily have every second turn.

Combined Feature Categories

Until now we have only explored features where every category comprised one single modality. All feature categories, summarized in Table 1, have performed significantly better than the baseline with the exception of the system's last dialogue act.

Features	JRIP	SVM	MLP
Head pose (HP)	0.690	0.724	0.709
HP+VAD	0.742	0.786	0.764
HP+Cards (C)	0.780	0.753	0.772
HP+Prosody (P)	0.700	0.698	0.789
HP+POS	0.754	0.731	0.772
HP+System DA (SDA)	0.725	0.739	0.728
Best combination			
HP+POS+C+P+SDA	0.745	0.796	**0.851**

Table 2: Weighted F_1 score for different feature set combinations using RIPPER (JRIP), Support Vector Machine (SVM) and Multilayer Perceptron (MLP) classifiers

Features	GP	LR
System DA	0.090	0.129
Prosody	0.146	0.135
POS	0.193	0.188
Cards	0.351	0.226
VAD	0.416	0.368
Head Pose (HP)	0.447	0.376
HP+System DA	0.482	0.373
HP+Prosody	0.500	0.377
HP+POS	0.471	0.393
HP+Cards	0.572	0.431
HP+VAD	0.611	0.523
Best combination		
HP+VAD+Cards	**0.677**	0.580

Table 3: Correlation coefficient for different feature set combinations using Gaussian Processes (GP) and Linear Regression (LR) classifiers

In this section we explore the combinations of features from different modalities, summarized in Table 2. Combinations including head pose typically performed best. The maximum performance using automatically generated features is 0.851 using 5 feature categories: head pose, POS, card movements, prosody and the system's dialog act.

4.2 Regression Model

While the end result of a turn-taking decision has a binary outcome, the distribution of annotations on a scale (Figure 4) suggests that there are stronger and weaker decisions, reflecting opportunities and obligations to take turns. As discussed above, such a score could be used together with a utility to take turns in a decision-theoretic framework. Thus, we also want to see whether it is possible to reproduce decisions on the scale. For this we explore the Gaussian Processes (GP) and Linear Regression (LR) classifiers in the WEKA toolkit. All results in this section are based on 10-fold cross-validation.

The individual feature categories have positive but low correlation coefficients (Table 3). Combining the feature categories with highest corre-

lation coefficients improve performance. The head pose in combination with VAD and card movements, using Gaussian Processes yields the highest correlation coefficient, 0.677.

4.3 Evaluation

We finally evaluated the best performing models built from the initial 9 dialogues on a separate test set of 43 decision points from a tenth dialogue, annotated both by the original annotator and a second annotator.

For the binary decision, we selected the MLP classifier with features from head pose, POS, card movements, prosody and the system's dialogue act. When evaluated on the test set annotated by the original annotator and the new annotator, the weighted F_1 score was 0.876 and 0.814 for 29 and 32 instances respectively. These are promising results, given the classifier's performance of 0.851 in the training set cross-validation (Table 2) and that the test set was from a previously unseen interaction.

The regression model was evaluated using the Gaussian Processes classifier with features from head pose, VAD and card movement. The correlation coefficients for the original annotator and the new annotator were 0.5959 and 0.5647 over 43 instances each, compared to 0.677 in the training set cross-validation (Table 3). The lower values could be due to a different distribution of annotations in the test set and the relatively small data set.

5 Discussion and Conclusions

In this study we have developed data-driven models that can be used by a robot to decide when to take the turn and not in multi-party situated interaction. In the case of a simple binary decision on whether to take the turn or not, the weighted F_1 score of 0.876 on data from previously unseen interactions, using several modalities in combination, is indeed promising, given a relatively small training material of 9 interactions and 688 instances. The decision process for the annotator is also simplified by not making separate decisions for turn ending and addressee detection. It should also be pointed out that we have only relied on automatically extractable features that can be derived in an online system. We have also achieved promising results for a regression model that could be used to identify both opportunities and obligations to take turns.

We have observed that combining features from different modalities yield performance improvements, and different combinations of features from diverse modalities can provide similar performance. This suggests that the multimodal redundancy indeed can be used to improve the robustness of the dialogue system. This is very relevant to the specific dialogue system in this study as head pose data sometimes is unavailable. Two possible remedies would be to only use classifiers that are robust against missing features, or to use multiple classifiers to step in when features are unavailable.

The results support that head pose, despite sometimes missing, is very useful for turn-taking decisions. This was expected, as head pose is the only of our available features that can be used to both select addressee and act as a turn-yielding cue. The results also indicate that POS provide useful information, even when based on ASR results with high WER. Provided that higher ASR performance becomes available, we could also benefit from other more sophisticated features, such as semantic completion (Gravano & Hirschberg, 2011), to predict turn-transition relevant places.

It is also interesting to see that the card movement is an important feature, as it suggests that moving of objects can be a dialogue act in itself, as discussed in Clark (2005). This makes situated dialogue systems – where the discussion involves actions and manipulation of objects – different from traditional dialogue systems, and should be taken into account when timing responses in such systems. This also suggests that it might be necessary to not just make turn-taking decisions at the end of IPUs, but rather continuous decisions. It is not obvious, however, how this would be annotated.

With the promising results of this study, we plan to expand on this work and integrate the turn-taking models into the live dialogue system, and see to what extent this improves the actual interaction. Of particular interest for future work is the regression model that could predict turn-taking on a continuous scale, which could be integrated into a decision-theoretic framework, so that the system could also take into account to what extent it has something important to say.

Acknowledgements

This work is supported by the Swedish research council (VR) project *Coordination of Attention and Turn-taking in Situated Interaction* (2013-1403, PI: Gabriel Skantze).

References

Al Moubayed, S., Skantze, G., & Beskow, J. (2013). The Furhat Back-Projected Humanoid Head - Lip reading, Gaze and Multiparty Interaction. *International Journal of Humanoid Robotics, 10*(1).

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

Ba, S. O., & 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), 16-33.

Bohus, D., & Horvitz, E. (2011). Decisions about turns in multiparty conversation: from perception to action. In *ICMI '11 Proceedings of the 13th international conference on multimodal interfaces* (pp. 153-160).

Clark, H. H., & Marshall, C. R. (1981). Definite reference and mutual knowledge. In Joshi, A. K., Webber, B. L., & Sag, I. A. (Eds.), *Elements of discourse understanding* (pp. 10–63). Cambridge, England: Cambridge University Press.

Clark, H. H. (2005). Coordinating with each other in a material world. *Discourse studies, 7*(4-5), 507-525.

Duncan, S. (1972). Some Signals and Rules for Taking Speaking Turns in Conversations. *J. of Personality and Social Psychology, 23*(2), 283-292.

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

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations, 11*(1).

Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics, 38*, 555-568.

Johansson, M., Skantze, G., & Gustafson, J. (2013). Head Pose Patterns in Multiparty Human-Robot Team-Building Interactions. In *International Conference on Social Robotics - ICSR 2013*. Bristol, UK.

Katzenmaier, M., Stiefelhagen, R., Schultz, T., Rogina, I., & Waibel, A. (2004). Identifying the Addressee in Human-Human-Robot Interactions based on Head Pose and Speech. In *Proceedings of International Conference on Multimodal Interfaces ICMI 2004*. PA, USA: State College.

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

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

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & 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*, 295-321.

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

Morency, L. P., de Kok, I., & Gratch, J. (2008). Predicting listener backchannels: A probabilistic multimodal approach. In *Proceedings of IVA* (pp. 176-190). Tokyo, Japan.

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

Raux, A., & Eskenazi, M. (2008). Optimizing endpointing thresholds using dialogue features in a spoken dialogue system. In *Proceedings of SIGdial 2008*. Columbus, OH, USA.

Shriberg, E., Stolcke, A., & Ravuri, S. (2013). Addressee detection for dialog systems using temporal and spectral dimensions of speaking style. In *Interspeech 2013* (pp. 2559-2563).

Skantze, G., & Al Moubayed, S. (2012). IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of ICMI*. Santa Monica, CA.

Skantze, G., & Schlangen, D. (2009). Incremental dialogue processing in a micro-domain. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*. Athens, Greece.

Skantze, G., Hjalmarsson, A., & Oertel, C. (2014). Turn-taking, Feedback and Joint Attention in Situated Human-Robot Interaction. *Speech Communication, 65*, 50-66.

Stiefelhagen, R., & Zhu, J. (2002). Head orientation and gaze direction in meetings. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems* (pp. 858-859).

Traum, D., & Rickel, J. (2001). Embodied Agents for Multi-party Dialogue in Immersive Virtual Worlds. In *Proc. of IJCAI Workshop on Knowledge and Reasoning in Practical Dialogue Systems* (pp. 766-773). Seattle, WA, US.

Vertegaal, R., Slagter, R., van der Veer, G., & Nijholt, A. (2001). Eye gaze patterns in conversations: there is more to conversational agents than meets the eyes. In *Proceedings of ACM Conf. on Human Factors in Computing Systems*.

Vinyals, O., Bohus, D., & Caruana, R. (2012). Learning speaker, addressee and overlap detection models from multimodal streams. In *Proceedings of the 14th ACM international conference on Multimodal interaction* (pp. 417-424).

de Cheveigné, A., & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America, 111*(4), 1917-1930.

Östling, R. (2013). Stagger: An open-source part of speech tagger for Swedish. *Northern European Journal of Language Technology (NEJLT), 3*, 1-18.

Appendix A. Gameplay Interaction – One Complete Round