Modeling who speaks next for less structured multi-party interactions

Nick DePalma
Samsung Research Of America
Mountain View, CA

ABSTRACT

Interactive applications of intelligent systems have yet to reach appropriate fluency of turn-taking. While simple methods such as audio-visual activity detection can be used in dyadic interactions, multi-party interactions may require a better estimates of interpersonal interactions. This paper presents our model, MuPeT, of turn prediction which focuses primarily on leveraging a quantized measure of gaze angle for turn transition estimates as well as speaker activity. The primary objective of this model is to estimate who in the interaction will take the next turn. We will introduce the problem of turn prediction and describe how it applies to previous models of turn taking and could be an important factor in multiparty turn taking. We use speaker activity from a microphone array to dynamically estimate who has the floor, where they are looking, who is being addressed, and report on the model's overall turn prediction capabilities with respect to previous studies of a discriminative approach. Our model is formulated as a probabilistic program that estimates a number of latent variables that are useful for a deeper model of turn prediction that leverages social cue exchange. We compare results against a predefined ground truth as well as a previous model of turn-taking, and show that MuPeT significantly outperforms previous models on the problem of turn prediction.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; **Cooperation and coordination**; *Cognitive science*; *Intelligent agents.*

KEYWORDS

turn-taking, bayesian-modeling, conversational-modeling

1 INTRODUCTION

Intelligent assistants that interact with human partners may need to one day intelligently navigate the stream of social cues that humans use to communicate with each other. Recent studies have shown that humans interact with embodied agents using the same types of cues that they use to interact with humans, especially if the intelligent agent is embodied in similar ways with similar affordances [14, 16, 30]. Fluent turn-taking behavior, which we define as intelligent behavior where the agent can precisely measure when the conversational floor is open, when an partner has the floor, and when to seize the open floor[25], is still a significant challenge for intelligent agents in the real world. Past work in this field has focused primarily on dyadic interactions in which turns were being exchanged between one human and one intelligent agent [4, 8, 28]. These works are incredibly promising and bode well for improved fluency between human-agent teaming or even agent learning, both of which depend on fluent turn-taking for their interaction model. However, moving past dyadic interactions has proven non-trivial. This can be broken down into a number of factors, only one of which we address in this paper. One major issue is that detecting gesture, posture, and gaze as well as speaker activity necessitates a multi-modal approach to detect both the visual and auditory nature of how these cues are being exchanged. When two interlocutors are conversing between themselves regarding a topic that is not relevant to the machine, then it should be evident that the machine should not take a turn until it is being addressed. In dyadic interactions, it is implied that there is only the user and the machine in the interaction. This means that every utterance is addressed to either the machine or the user depending on which turn it is. Lastly, previous work has been focused on scheduling actions in multi-party turn taking [4] and is a major concern for temporally sensitive action scheduling. This concern for action timing is a delicate process that involves incorporating the time in which an action takes to initiate (so called preparatory action) before the agent has been considered to have taken the floor. Psychosocial literature has shown that the turn-transition conversational gaps are both culturally specific [27] as well as group dependent [2] and could also be individual specific. Our belief is that the ability to predict who will take the floor next could be a critical step in helping to schedule actions in multi-party turn taking.

This research provides the following contributions:

(1) a Bayesian formulation of speaker and addressee estimation;
(2) a computational model that models a turn's statistics to predict who will take the next turn;
(3) a study that quantifies the model against a modified common model for turn-prediction[10];
(4) An analysis on the performance of our addressee estimates measured per-turn.

Next we will review literature that could be considered relevant to the general problem of taking turns.

2 BACKGROUND

Research in turn-taking can be roughly organized in early psychological modeling in both dyadic and multi-party interactionsas well as human-machine studies. We have isolated the work most relevant to our own in Table 1. First we will cover the early psychological findings and hypotheses and follow up with recent computational modeling.

Sacks and Schegloff's early work in turn-taking [23] provide an early account of multi-party turn taking in conversation. They note that organizational structure arises in social systems in which resources are limited. The obvious example is conversationally in which the resource is limited to auditory floor. Transitions in this case are delicately executed and turns are variable in length. Turns

Similar Computational Models of Conversational Turn- Taking

	Model type	Objective	Input features	Analysis of
Ishii 2016 [10]	SVM	Multiparty turn-prediction	Gaze & Timing	**Next turn prediction**
Jovavich 2007 [12]	Dyn. Bayes Network	Addressee estimation	Utterance, gaze	Addressee correctness
op van Akker 2009 [1]	Rule based	Addressee estimation	Utterance, gaze	Addressee correctness
Sato et. al. 2014 [24]	Decision Trees	Multi-party turn-taking	Gaze	Timing
Bohus Horvitz 2010 [4]	Rule based	Multi-party turn-taking	Speech	Timing
This work (MuPeT)	**Bayesian Network**	**Turn prediction**	**Speech, gaze**	**Next turn prediction**

Table 1: Relevant and similar work to ours.

are exchanged in various manners both in a constructional (e.g. "Don't you think?") as well as explicitly through allocative measures (e.g. "Seen it?"). Unfortunately these systematics were too simplistic and did not consider deeply the empirical findings on non-verbal signalling in turn-taking [21]. Continuing on this work, Beattie [2] noted that simple social gaze cues were used to coordinate turns. More specifically, they noted that these gaze cues were context dependent on the specific task. Watanuki and Togawa [31] found that head nods and mutual gaze were inversely related in how the participants were acknowledging each other. More recently, the study of overlapped turns has become a central point of discussion. Schegloff's work in turn-taking through measurement of overlap [25] initiated some of the first insights into the dynamics that lead to appropriate and fluent turn-taking.

While computationally modeling the turn-taking phenomena is a daunting task, many researchers have focused on various components of the problem. We begin first by reviewing past work that modeled dyadic turn-taking. Following the work of Watanuki, Tagawa, and Beattie, Novick [20] provided an early account of modeling gaze between participants in a dyadic game centered task. The model relied on a rule based Prolog system. The authors highlight a notion of turn seizing that is preceded by a mutual-gaze moment where the participants look at each other. The turn is seized when mutual gaze is broken. This model of gaze may be too simple. In fact, significant work has been dedicated to understanding a deeper articulatory attention in dialogic multi-party turn-taking [29]. Thorisson [28] performed similar studies of how a virtual agent can seize the floor during an interaction with a human participant. His model focused primarily on a finite state machine model that passed state between "I-have-turn", "Floor-is-open", and "You-have-turn" states. The state transitions were mediated by a number of measurements from an exo-suit. We have extended this model, GANDALF, to compare our model against. Jonsdottir et. al. [11] extended this model to support using an adapted actor-critic Q-learning algorithm. Chao and Thomaz [8] evaluate a dyadic turn taking model within a human-robot interaction. Their findings suggest that overlapping turns can have a perceived effect on turn-taking in task-oriented interactions. We differentiate turn prediction as one component in a large pipeline from perception to action scheduling similar to Bohus and Horvitz[5]. But unlike Bohus and Horvitz, our focus is on the models ability to predict who will take the next turn.

Multi-party turn taking is a relatively new computational modeling problem that presents numerous new challenges in state estimation, turn modeling, timing, and scheduling. From the perspective of timing and scheduling, we believe that prediction may provide a leg up in the overall problem of turn scheduling. While we don't provide this evidence, we are optimistic with our preliminary results on prediction. Leveraging the notion of Vertegaal's articulatory attention[29] in multi-party interactions, we hypothesize that gaze plays a large role in turn prediction.

One key challenge in multi-party turn taking is addressee detection which has been studied in separate contexts. Jovanovic [12] built an addressee estimation system that evaluates three different dynamic Bayesian networks. This model uses constructional elements, specifically pronoun usages as well as gaze signals to estimate who is being addressed. Bayesian Augmented Naive bayes (BAN), Tree Augmented Naive Bayes (TAN), and GBN (General Bayesian Network) were measured separately. Addressee estimates were found on a per utterance basis. Addressee estimates in the TAN model that utilized utterances and gaze were measured at about 77% at best for the bigram variant. Op van Akker et. al. [1] evaluates Jovanovic's model along with their rule based system that leverages gaze cues toward addressee identification and finds the rule based "Traum's method" to outperform previous DBN models on annotated data. Bohus and Horvitz [4] were among the first to implement a turn-taking mechanism for multi-party interactions. Their system uses a rule based model for addressee estimation in a real time system for turn-taking with a virtual agent. Finally, some researchers have approached the problem in a unimodal way, focused primarily on the audio signal. Shriberg et. al. [26] build a system that leverages prosodic utterance, and direction of arrival data to understand who is being addressed. Finally Ishii et. al. [10] looked at next turn prediction using SVMs and hand engineered features. We report on their performance as well as our own on a common dataset.

Systemic issues in multi-party interactions still require significant effort to model. Leite et. al. [17, 18] perform a study between a virtual agent and multiple children, a novel population for computational modeling multi-party turn-taking. The turn-taking model was built with hand annotated features and turn state was determined using support vector machines. Overlapping turns was a significant focus of this work. Gorga and Otsuka [9] detail a hierarchical HMM model that attempts to model the conversational dynamics using a Bayesian network and learn predictive gestures that lead to gaps in the conversation. Similar to this model, we use gaze estimates for the "turn passing ceremony" but unlike their model, we are reporting results about who will take the next turn to determine how likely this ceremony leads to valid predictions.

Sato et. al. [24] explore turn-taking between three other interaction partners and a robot by playing a game. Gaze and prosody were used within a decision tree model to negotiate turns with human partners in a highly controlled environment. Finally Zhou and Wachs [33] build a spiking neural network model of turn taking using hand flexion features to determine whose turn will be taken next between a human and a robot. Our model is meant to be more robust to error (Section 6) working with non-annotated and cleaned data. In addition, it is meant to provide the ability to predict the next turn for later speech-act scheduling (Section 5).

In the following, we present our model of turn prediction that leverages addressee estimates in noisy environments to model a multi-party conversation.

3 REPRESENTATION

Our unsupervised model of turn-prediction uses variational inference (VI) to estimate the posterior parameters of our Bayesian model across a window of 10 time steps. The model results were tested on an open dataset for the sake of reproducability. We used the probabilistic programming paradigm to build our architecture. Similar to Wingate and Weber [32], we use a variation of probabilistic programming called Pyro built by Bingham et. al. [3]. The model itself is detailed below.

For the following, we will denote an index of a particular person using the variable k, the number of people interacting with each other as K, and $k \in \{0::K\}$. Time is denoted as an index t in a set of turns T, or in other words $t \in \{0::T\}$. Time in this instance is being measured in turns, not seconds. Finally, a turn is made up of samples. To index these samples, we will use the variable i in some set of size N, or $i \in \{0::N\}$. Positions in the model are modeled as angles in radians, denoted as some θ. For instance, $\theta_{i,k}$ is the angle of a particular person, k, at sample i. Gaze is denoted with the variable ∂ and current speaker is measured with the variable S_i at some time step i. Turn in this model does not allow for overlap. Finally f denotes a deterministic function in the model that uses a set of latent parameters that are found via VI and p denotes a probability distribution that came from a specified probability function. We will use the vocabulary "global" to refer to variables that are not a part of any plate. For instance, the sensor error, σ, is estimated across all time steps and across all position estimates and hence could be called "global" from the scope of the plates.

We formulate the model in a piece-wise fashion in which each set of posterior parameters of the model represent a piece of the interaction dynamics that are needed for the next step or next turn estimates. The model can be broken down into the following modules: a) The current position estimate of all of the interaction partners and who of them is speaking at any one time, b) an estimate of who the speaker is addressing at each time step, and finally c) a per-turn prediction of who will take the floor next. We estimate relative location to the machine using observations from both visual and audio channels. In addition, we use approximate gaze angle to generate a distribution over those being addressed and use that information for turn prediction.

In the following sections, we will split each module in Figure 1 into its component parts for the purpose of detailing the design of the model itself. We validate the model on a publicly available dataset, the AMI Corpus[6, 7, 13] in Section 5.

3.1 Location estimates

Our objective is to model the posterior distribution as the set of mean angles of each of the interaction partners with respect to the machines origin. In the case of the AMI corpus, it is sometimes at the center of the circle. Figure 1a and 2 notes that for K people in the interaction, each modeled with a normal distribution ($\mathcal{N}$) on the metric of radians with global sensor error, σ. In Figure 1a, $\theta_{A,i} \sim \mathcal{N}(\mu_{S_{i,k}}, \sigma_{S_{i,k}})$ where the K μs are being modeled independently in radians. Each one of those parameters is initialized with another normal prior (i.e. $\mu_k \sim \mathcal{N}(\mu_m, \sigma_m)$) while the sensor error measured across all K participants uses a uniform prior: $\sigma \sim \mathcal{U}(0, \pi)$.

To estimate whose turn it is, we use the microphone array's direction of arrival at sample i. Direction of arrival in a highly insulated room can provide a clear signal of who is speaking. We will estimate this as a ordinal index, $S_i \in \{0::K\}$. Note a particular subtlety. A turn t is indexed by a run of samples S_j in which all S are equal. S_t is defined as following: $S_i = arg\,max_k P(\theta_{A,i} \mid \mathcal{N}(\mu_k, \sigma))$. This way, when we are interested in the position of the speaker in radians, we can query the model for $\mu_{k=S_t}$ and use it for the geometry of gaze projection to another conversational partner. We have detailed the results of this estimate in Section 5.2.

Finally, we are interested in modeling the prior probability, p_d to query for a particular interlocutors natural inclination toward taking a turn. To do this, we are interested in modeling a multinomial distribution with parameters p_d. We use a conjugate prior of a Dirichlet distribution to find these parameters: $S_t \sim \mathcal{M}(p_d)$ and $p_d \sim Dir(\alpha)$. This set of parameters may be interesting as it provides a global estimate about who may be speaking next that does not rely specifically on any social signals.

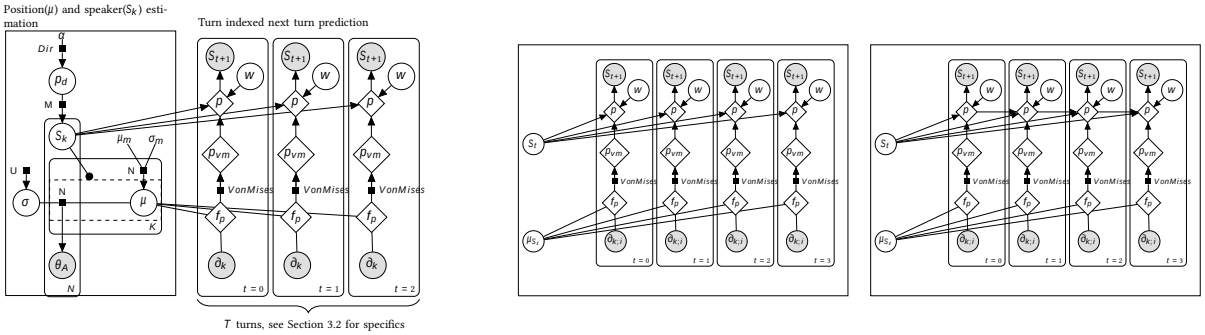
Next we use these estimates to predict who will take the next turn.

3.2 Addressee and turn prediction estimates

One of the key assumptions in our model is that those who are being addressed will most likely respond next in the conversation. There have been plenty of researchers who have focused on addressee detection. Jovanovic [12] built an addressee estimation system that used a Bayesian network in a similar way for estimation but did not use social signals like gaze estimates. Like the work of Gorga et. al. [9], we quantize the gaze observations for the sake of stabilizing ∂ temporo-spatially in an unstructured environment but instead we are investigating the phenomena of predicting who will take the next turn which will have significant importance to turn-taking behavior of virtual and embodied agents going forward. Figure 1b details the turn taking model in two separate experiments that are reported on in Section 5.3 and Section 5.4.

In this section, we will detail how each of the two models (MuPeT-A and MuPeT-B) work (see Figure 1b). Temporally the two models are similar. There are T turns in N observations where $T << N$. For each turn, a variable number of observations map to the particular turn. For instance, if turn t spans observations from $i = \{j_0:::j_n\}$ then the window of gaze estimates through that turn could create a distribution over objects and people. For the sake of the experiment,

θ_A: Mic. direction of arrival, θ_A 2 » $\pi;\pi^1\!/_4$ S_k: Speaker index estimate, S_k M $^1p_d{}^o$ f_p: Radial target of gaze	σ: Sensor error estimate, σ U ^{1}0; π^o p_d: Probability of taking a turn, p_d $Dir^1\alpha^o$ p_{vm}: Addressee dist., p_{vm} = V M ^1f_p; 1:0^o	μ_k: Pos. of interlocutor k, μ_k N $^1\mu_m;\sigma_m{}^o$ $\partial_{k;i}$: Gaze dir. of speaker, ∂ 2 f $\frac{\pi}{4}$; 0; $\frac{\pi}{4}$ g p: Time t estimate to predict S_{t+1}



(a) Isolating multiple party members and overview of MuPeT model

(b) The MuPeT experiment evaluates two different models that tests conditional dependencies on previous turns. More details in Section 3.3.

Figure 1: The Multi Person Turn prediction or MuPeT model. This model attempts to predict who will take the next turn in a multiparty conversation.

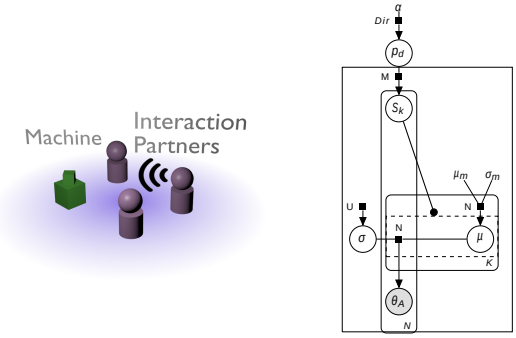


Figure 2: Localizing multiple people surrounding the machine.

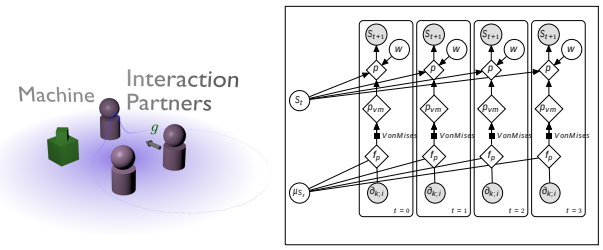


Figure 3: Computing the addressee and next turn prediction, p. Each turn uses an aggregate frequency of gaze to estimate the addressee of the turn's speech act. In some cases, as we will discuss, a decay model is used from turn-to-turn.

we ignore the gaze targets on objects and only focus on gaze targets of people within the conversation. At the highest sense, we are creating a probability distribution over K participants and choosing target $S_t = ar\partial max_k{}^1p_k{}^o$ where p is prediction distribution for time $t+1$. The exact function for p depends on which model we are discussing. In the case of MuPeT-A, the function only spans estimates across the j_0 to j_n indices. In the case of MuPeT-B, the distribution initializes uniformly and is preserved from the last time the speaker spoke. This function corresponds to the idea that conversational partners may typically pass turns to the same interlocutor like you would in a game of friendly catch. For instance, let's say that there are three conversational partners: A, B, and C. Let's say that the speakers (S_t) are estimated across some set of 7 turns as B,A,C,A,B,A,B and the addressee distribution for interlocutor A was -, [0,0.25,0.75], -, [0,0.75,0.25], -, [0,0.65,0.35],- for each turn where the dash (-) represents someone else's turn and the array represents the probability of transitioning to the next person. MuPeT-A will only use the addressee estimate during that turn. MuPeT-B will initialize with the previous time it was speaker A's turn (a non-dash) with some amount of exponential decay before incorporating the new signals.

Computing the likelihood works in the following way. First, let's use Figure 3, left as an example. The green machine is computing that the speaker at angle $\mu_{S_t} = 0rad$ is speaking. That particular interlocutor is looking to the machine's left so $\partial_{S_t} = \frac{\pi}{4}rad$. The gaze observation is computed relative to the speaker's position but we need a global estimate. We're interested in computing $p = VonMises^1f_p; 1:^o$ where f_p is the projected global gaze target. We project the gaze target back onto the circle and use it as the gaze referent:

$$\theta_G = \mu_{S_t} + \pi + \partial_{S_t} \quad (1)$$

$$v_1 = \langle cos(\theta_G), sin(\partial_{S_t}) \rangle \quad (2)$$

$$v_2 = \langle cos(\mu_{S_t}), sin(\mu_{S_t}) \rangle \quad (3)$$

$$G = arctan(v_1 \quad 2v_1 \quad v_2 v_2) \quad (4)$$

$$f_p(\partial_{S_t}, \mu_{S_t}, \mu_k) = \mu_k \quad G, 8k \quad (5)$$

This estimate can be computed for every time-aligned (to index i) gaze and position estimate from the dataset. This resulting f_p is binned in various ways as we will discuss in Section 3.3.

3.3 Per Turn Binning

As discussed previously, we bin the model based on a few hypotheses to get the final p distribution at the end of the speaker's turn. The first hypothesis is that the prediction is based on just the observations of where the speaker gazed during that turn. The other hypothesis is that the prediction is based on not just this turn but previous turns by the current speaker. We will cover the both functions that we use to evaluate these hypotheses in this section.

First, for MuPeT-A, we only consider the observations during each turn $t \in \{0::T\}$, let's say from indices j_0 to j_n and that there are n samples in a particular turn. In this case, $p = \frac{1}{j_0 \; j_n} \sum_{i=j_0}^{j_n} p_{vm;i}$. This corresponds to the idea that predicting the next turn is completely depends on the gaze observations during the speaker's turn. In other words, we use a simple average over all estimates of the next speaker during that turn based on our addressee estimate.

For the MuPeT-B model, we model the distribution to be dependent on the previous turn of the speaker. In Figure 1b we model those dependencies as a separate function. In this function, it could depend on a previous time step. The update step is defined as following:

$$p_{t;k} = \begin{cases} \frac{1}{j_n \; j_0+1} e^{\; \lambda \delta t} p_{vm;t \quad 1;k} + \sum_{i=j_0}^{j_n} p_{vm;i;k}; & \text{if } S_t == k: \\ p_{vm;t \quad 1;k}; & \text{otherwise:} \end{cases} \quad (6)$$

For the MuPeT-B case, we consider not just the observations during the turn t, but during the previous turn in which S_t last spoke. The difference is in time, δt is defined in sample iterations. In this hypothesis, we pose the question of whether we can borrow the overall distribution of who a particular speaker is talking to across turns.

In the next section, we will discuss how inference of the posterior parameters was determined.

4 LEARNING AND INFERENCE

We use data from the AMI corpus to evaluate our model. The AMI corpus includes 32 scenario meetings at three different locations. Each scenario involves $K = 4$ participants. For each participant, gaze, microphone array activity data, speech annotations, and gesture are available. The data was preprocessed to run in Torch. Each observation in all N samples was either a change in gaze or microphone array corresponding to the ∂_i and $\theta_{A;i}$ respectively.

The posterior parameters of this probabilistic programming model were found via variational inference. We used the autodelta algorithm [22] with a trace graph ELBO [19] for the guide generation and lower bound function generation, respectively. Constraints on the parameters are as follows: $\{\sigma_m \in (0; \pi)\}, \mu_m \in (\quad \pi; \pi)\}$, $w_0 + w_1 = 1.0$, and $w \in \langle 0; 1 \rangle$. The rest of the parameters are found in the following section. The model is trained via a 2 step inference process in which an estimate of the first set of global parameters is found followed by a turn-indexed set of parameters is found for turn prediction. Within this model, each turn prediction estimate at turn t is compared to ground truth from the AMI dataset at as the speaker at turn $t + 1$. Optimization is performed with the Adam algorithm [15] at a learning rate of 0.5 for 200 samples on 3 chains.

Figure 4 details inference during model training stage.

5 EVALUATION

We evaluated our model using the AMI corpus. The following is a detailed account of the parameters and findings.

5.1 Parameter optimization

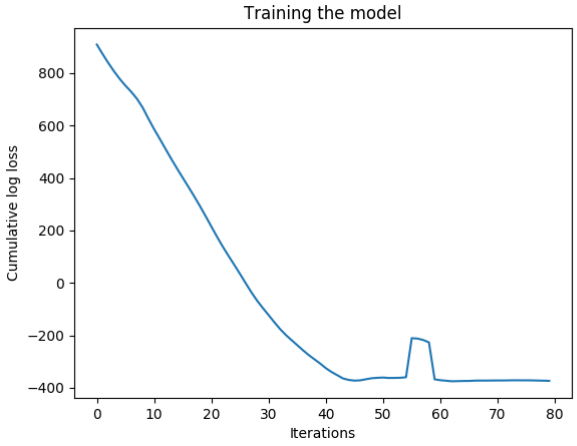


Figure 4: Log probability during model inference with the AMI corpus. Iterations on x-axis and log-probability on the y-axis.

During parameter optimization, we found that the learning procedure accurately optimizes the parameters of the posterior distribution. Figure 4 illustrates the overall log-probability less than 100 time steps. We measure the cumulative loss of the log likelihood function in the model's posterior parameters as VI is performed. For our dataset, this is enough to show convergence. As an example, for the IS2001 scenario, we find that the global parameters are $\sigma = 0.05 rad$ and $\mu = \langle -\frac{3}{4}\pi; -\frac{1}{4}\pi; \frac{1}{4}\pi; \frac{3}{4}\pi \rangle$ which makes sense since the microphone array is located in the center of the interaction with each interlocutor surrounding the microphone array. We visualize the positions of the interlocutors in Figure 5.2. p was found to be $\langle 0.55; 0.19; 0.13; 0.13 \rangle$.

5.2 Recovered interlocutor locations and speaker activity

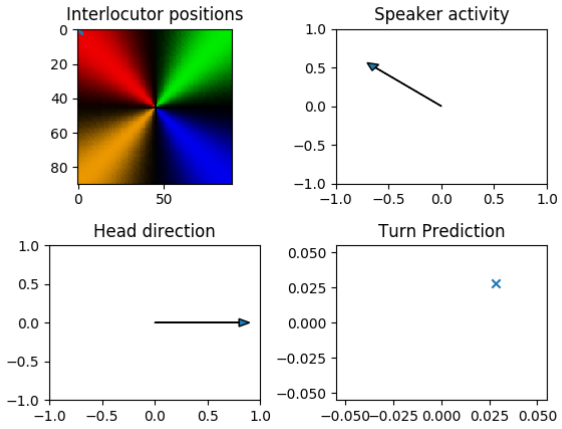


Figure 5: Estimated positions of the participants across the dataset for the IS 2001 scenario. The visualization is rendered as if it is taken top down. So in this example, μ_0 is speaking (red), looking to their left, and it is predicted that μ_1 will speak next. The k indices are specified in the following order: [red, green, blue, orange] for $k = 0; 1; 2; 3$ respectively.

Following Section 5.1, we'd like to analyze the latent parameters that were inferred from the microphone array and gaze tracking. For scenario IS2001, we snapshot one instance in time in Figure 5, top left, in which we see that the positions of the interlocutors are properly estimated from the IS2001 dataset. These μ_k parameters represent the positions of each of the interlocutors to be used in subsequent calculations. When performing geometric transformations later in the inference pipeline, these estimates provide a stable signal from which to use.

Speaker activity is another signal that needs to be stabilized to compute addressee and turn prediction measures. Figure 5, top right, helps us to visualize the need for speaker activity in computing the global geometric addressee. Figure 6 shows the cleaned up signal as measured by μ_{S_i}.

To model the accuracy of the speaker location, we plot the the activity of the microphone array against the estimated speaker indexed mean position, μ_{S_i} in blue along with the actual speaker signal from the microphone direction of arrival in red. Figure 6 shows the estimated S_i speaker index on the estimated position μ_{S_i} over 700 sample steps. The y-axis is in radians. As you can see, the model appropriately cleans the signal while estimating the current speaker in a conversation.

5.3 Modeling turn predictions with exponential decay

As discussed in Section 3.3, we are interested in modeling temporal dynamics of predicting the next interlocutor to take a turn. To

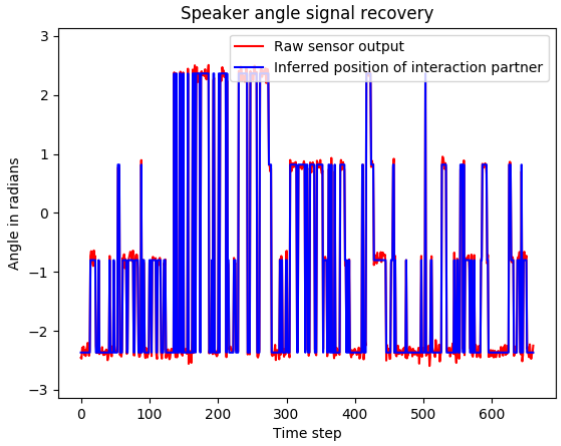


Figure 6: Speaker activity signal from the AMI dataset. X-axis is 0 to about 700 sample steps from the microphone array. Y-axis is direction of arrival activity in radians. We demonstrate that the new estimates are significantly cleaned up from the signal, providing a better estimate to use during the later inference steps.

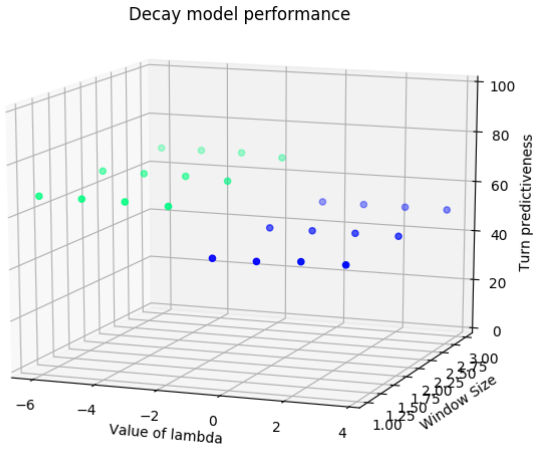


Figure 7: Performance of the recursive decay functions effect on turn prediction. As a reminder, we define the recursive decayed update function component as $F^1\lambda; W = kj_n$ $j_0 k^0 = \frac{1}{j_n \; j_0+1} e^{\;\lambda\delta t} p_{vm;t \;\; 1;k}$.

do this, we are interested in the various parameters of our model. Very generally, the decay model manifests itself as an exponential coefficient that depends on the different in time since last turn, δt. It also depends on the window size (how many turns to consider). Figure 5.3 visualizes values of λ in relation to the window size w against it's predictive power. We found that there is a significant change in predictive power mainly on the sign of λ. In addition, we found that for the model MuPeT-B, that there is a trend between increasing the window size and it's predictive value. This is to be

expected as considering more turns may provide a better estimate of who a particular interlocutor passes turns to next.

5.4 Overall turns predicted correctly

Model type	Prediction performance
Ground Truth	0.62
Initial probability only (p_d)	0.46
MP-GANDALF	0.21
Ishii 2016	0.43
MuPeT-A	0.587
MuPeT-B	**0.606**

Figure 8: Predictive value of model given the estimate of the current speaker and the distribution inferred from the social gaze cues across all timesteps using the MuPeT model.

Finally, we are interested in comparing our model, MuPeT-B against a few other models. In particular, a previous model of turn prediction, Ishii 2016 [10] as well as a naive estimate based on hand annotated gaze estimates per turn. In other words, we compute the addressee per turn and use it as a prediction based on high quality similar to MuPeT-A, hand annotated data whereas our model uses noisy signals from the sensors. We will call this ground truth as we consider it an upper bound using our method. For our evaluation of the MuPeT-B model, we compute the turn prediction estimate using the default values of $\lambda = 1$ and $W = 5$. Finally we measure the predictiveness of the Ishii algorithm [10] that hand craft the gaze transitions for 2 time gaze steps and classify next speaker prediction with support vector machines.

Using the ground truth of the dataset (S_{t+1}), we measure the predictiveness of each model. Figure 8 demonstrates that our model of turn prediction using exponential decay performed better than previous models on the open AMI corpus dataset.

5.4.1 Constraints. During the learning process, we constrained our model in a few minor ways. While some of these were designed to keep clusters from merging, others were designed only to make inference faster. First, individuals cannot occupy the same position in space: $k\mu_i \quad \mu_j k < 0{:}5rad; 8i; j$ and $i \neq j$. In other words, when estimating the positions of every individual in conversation, no two interlocutors can overlap in position. Second, the position of the participant must be in relation to the central location of the machine: $\pi <= \mu_i <= \pi; 8i$.

We also constrain the sensor error to be strictly positive since it represents the standard deviation: $\sigma > 0$. Finally, blending weights were constrained to be strictly between zero and one: $0 <= w_0; w_1 <= 1{:}0$. In addition, we set a maximum of $K = 4$ participants as a hyperparameter.

6 CONCLUSIONS AND OPPORTUNITIES

Borrowing Vertegaal's [29] vocabulary for articulatory attention, our goal was to operationalize this cue and determine whether or not gaze cues prime individuals into taking turns. We have found that this phenomena is considerably deeper than this one cue alone. Multi-sample aggregation of gaze cues does not alone signal turns in conversation but does play a role. Our understanding of the results in Section 5.4 is that they demonstrate how important these gaze cues are for turn prediction alone. We are optimistic about the future of this problem and believe that the form of this problem can only be fully solved through multi-modal interactions.

There has been much work in understanding the temporal issues in turn taking. Table 1 details just how much of the work has been dedicated to these issues. For instance, speaker prosody and linguistic information has been shown to detail the upcoming nature of an open floor [26]. This work has focused primarily on speaker cues from a microphone and is not multimodal. Other work has focused on gesture [28] that could only be sensed from an exo-suit or from a visual field. Some work focused on the mixed initiative nature of the problem to detect open floor states and to know when the agent has the opportunity to jump in [8]. But regardless of the approach, it is becoming increasingly clear that this problem is largely one that requires a coordinated effort across many different nonverbal and verbal cues.

In closing, one important distinction that we want to emphasize is that the addressee may be different from the the person who may take the next turn. While we are currently modeling them as equivalent, we don't now believe that this could be the case. Turn prediction is more than just addressee understanding. In fact, it may incorporate the initial probabilities per group as p_d in Figure 1 with the intuition being that some participants take a very dominant role in the conversation. For instance, we found p_d to over-represent certain participants as turn dominant. We believe the future of this problem will include many more modalities than just audio and visual and will leverage signals that are prosodic in nature or to detect body language in deeper ways. Furthermore, we believe our model of turn-prediction could lead to better turn-taking in a mixed initiative, unstructured interaction and hope to show this in future work.

Acknowledgements

Thank you to the staff of Samsung Research America for helping fund this work. Particularly the advice and mentorship of Gene Becker, Akash Sahoo, and Eric Lipschutz for their support. Also thank you to Min Jung Lee for administrative help.

REFERENCES

[1] Rieks op den Akker and David Traum. 2009. A comparison of addressee detection methods for multiparty conversations. In *Workshop on the Semantics and Pragmatics of Dialogue.*

[2] Geoffrey Beattie. 1983. *Talk: An analysis of speech and non-verbal behaviour in conversation.* Open University Press.

[3] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. 2018. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research* (2018).

[4] Dan Bohus and Eric Horvitz. 2010. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction.* ACM, 5.

[5] Dan Bohus and Eric Horvitz. 2011. Decisions about turns in multiparty conversation: from perception to action. In *Proceedings of the 13th international conference on multimodal interfaces.* ACM, 153–160.

[6] Jean Carletta. 2007. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation* 41, 2 (2007), 181–190.

[7] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried

Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI Meeting Corpus: A Pre-announcement. In *Proceedings of the Second International Conference on Machine Learning for Multimodal Interaction (MLMI'05)*. Springer-Verlag, Berlin, Heidelberg, 28–39. https://doi.org/10.1007/11677482_3

[8] Crystal Chao and Andrea L Thomaz. 2012. Timing in multimodal turn-taking interactions: Control and analysis using timed petri nets. *Journal of Human-Robot Interaction* 1, 1 (2012), 4–25.

[9] Sebastian Gorga and Kazuhiro Otsuka. 2010. Conversation scene analysis based on dynamic Bayesian network and image-based gaze detection. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*. ACM, 54.

[10] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2016. Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 6, 1 (2016), 4.

[11] Gudny Ragna Jonsdottir, Kristinn R Thorisson, and Eric Nivel. 2008. Learning smooth, human-like turntaking in realtime dialogue. In *International Workshop on Intelligent Virtual Agents*. Springer, 162–175.

[12] Natasa Jovanovic. 2007. To Whom It May Concern-Addressee Identification in Face-to-Face Meetings. (2007).

[13] Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt. 2006. A corpus for studying addressing behaviour in multi-party dialogues. *Language Resources and Evaluation* 40, 1 (2006), 5–23.

[14] Sara Kiesler, Aaron Powers, Susan R Fussell, and Cristen Torrey. 2008. Anthropomorphic interactions with a robot and robot–like agent. *Social Cognition* 26, 2 (2008), 169–181.

[15] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[16] Hatice Kose-Bagci, Ester Ferrari, Kerstin Dautenhahn, Dag Sverre Syrdal, and Chrystopher L Nehaniv. 2009. Effects of embodiment and gestures on social interaction in drumming games with a humanoid robot. *Advanced Robotics* 23, 14 (2009), 1951–1996.

[17] Iolanda Leite, Hannaneh Hajishirzi, Sean Andrist, and Jill Lehman. 2013. Managing chaos: models of turn-taking in character-multichild interactions. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 43–50.

[18] Iolanda Leite, Hannaneh Hajishirzi, Sean Andrist, and Jill F Lehman. 2013. Take or wait? learning turn-taking from multiparty data. In *Workshops at the Twenty-Seventh AAAI Conference on Artificial Intelligence*.

[19] Andriy Mnih and Karol Gregor. 2014. Neural variational inference and learning in belief networks. *arXiv preprint arXiv:1402.0030* (2014).

[20] David G Novick, Brian Hansen, and Karen Ward. 1996. Coordinating turn-taking with gaze. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, Vol. 3. IEEE, 1888–1891.

[21] Richard JD Power and Maria Felicita Dal Martello. 1986. Some criticisms of Sacks, Schegloff, and Jefferson on turn taking. *Semiotica* 58, 1-2 (1986), 29–40.

[22] Daniel Ritchie, Paul Horsfall, and Noah D Goodman. 2016. Deep amortized inference for probabilistic programs. *arXiv preprint arXiv:1610.05735* (2016).

[23] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In *Studies in the organization of conversational interaction*. Elsevier, 7–55.

[24] Ryo Sato and Yugo Takeuchi. 2014. Coordinating turn-taking and talking in multi-party conversations by controlling Robot's eye-gaze. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 280–285.

[25] Emanuel A Schegloff. 2000. Overlapping talk and the organization of turn-taking for conversation. *Language in society* 29, 1 (2000), 1–63.

[26] Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tür, and Larry Heck. 2012. Learning when to listen: Detecting system-addressed speech in human-human-computer dialog. In *Thirteenth Annual Conference of the International Speech Communication Association*.

[27] Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences* 106, 26 (2009), 10587–10592.

[28] Kristinn R Thórisson. 2002. Natural turn-taking needs no manual: Computational theory and model, from perception to action. In *Multimodality in language and speech systems*. Springer, 173–207.

[29] Roel Vertegaal. 1998. Look who's talking to whom. *Mediating Joint Attention in multiparty* (1998).

[30] Joshua Wainer, David J Feil-Seifer, Dylan A Shell, and Maja J Mataric. 2006. The role of physical embodiment in human-robot interaction. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 117–122.

[31] Keiko Watanuki and Fumio Togawa. 1995. Some signals of emotional arousal: Analysis of conversations using a multimodal interaction database. In *Fourth European Conference on Speech Communication and Technology*.

[32] David Wingate and Theophane Weber. 2013. Automated variational inference in probabilistic programming. *arXiv preprint arXiv:1301.1299* (2013).

[33] Tian Zhou and Juan P Wachs. 2018. Early Turn-taking Prediction with Spiking Neural Networks for Human Robot Collaboration. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1–9.