Sumit Basu, Tanzeem Choudhury, Brian Clarkson, and Alex Pentland. "Towards Measuring Human Interactions in Conversational Settings." Appears in Proceedings of the IEEE Int'l Workshop on Cues in Communication (CUES 2001) at CVPR 2001. Kauai, Hawaii, December 9, 2001.

Towards Measuring Human Interactions in Conversational Settings

Sumit Basu*, Tanzeem Choudhury*, Brian Clarkson*, and Alex Pentland
Media Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02139

Abstract

The Facilitator Room is a conference room that is outfitted with sensors and actuators in order to observe and influence human behavior in conversational settings. This work describes our efforts thus far in developing robust sensing mechanisms in the visual and auditory domains and designing statistical models to analyze and predict behavior. We review the "Influence Model," our primary analysis tool, which we developed for this purpose in [1]. The "Influence Model" models a group of interacting agents as a group of simple Markov chains that are each influencing each other's state transitions. We demonstrate the capabilities of this model on both synthetic data and real interaction data from the Facilitator Room. We describe our approaches for doing prediction with this model and close with a discussion of how we plan to influence interactions with the room's actuators.

1 Introduction

One of the most interesting aspects of human interactions is the "influence" that individuals and subgroups can have on each other. Certain people always seem to dominate the conversation, others seem particularly capable of getting people to agree with them, and some will rarely speak but cause a significant shift in focus when they do. Undoubtedly, some of this has to do with what words the participants say. However, it seems that how they say these words, in terms of speaking style and body language, can play a significant effect. Sociologists have long studied this manner of effect [2] and have shown some interesting properties. For instance, an interesting study in the Tipping Point [3] showed how a good salesman succeeded in his art by getting his audience to synchronize their body language with his – in particular, nodding their heads when he did.

We are interested in studying this type of effect in a quantitative way. In particular, we want to determine how much influence each participant has on the others. We cast this influence in terms of predictability – how useful is person B's state in determining person A's next state? We do this using the "Influence Model," described in Section 3. The advantages to being able to better predict individuals' states in such a way are many. The simplest application would be to use this information to determine who will speak next and steer a "smart camera" system for recording the meeting. Our long-term goal, though, is to use this information to affect the participants during their interaction. For example, we could try to use the actuators to make the quiet individual(s) speak up more and have the dominating speaker(s) quiet down, or vice versa. We can use the techniques for estimating influence between people to also estimate the influence between an actuator and people. With a reliable means of predicting the effects of the actuator, we can then design control loops for driving/limiting certain kinds of behavior.

In this paper, we first describe the Facilitator Room, an experimental setup with sensors and actuators we have constructed for studying these effects. We also describe our work in developing robust visual and auditory features to characterize the behaviors of the individuals in the room. In Section 3, we review the "Influence Model," a Dynamic Bayes Net (DBN) model we developed in [1] for analyzing human interactions. We show the results of applying this model on synthetic and real data to estimate the influences between individuals. We then describe our preliminary work on doing prediction with our model, along with our proposed methods to improve these early techniques. Finally, we close with a discussion of our findings thus far and our future plans.

2 The Facilitator Room

In the interests of studying the interactions between humans and the influences of various experimental variables, we have developed an experimental setup we call the "Facilitator Room." This room is a 15-foot by 15-

* The first three authors contributed equally to this paper and are listed alphabetically

foot space with three couches and a table. The room is instrumented with five pan-tilt-zoom cameras and an array of microphones:

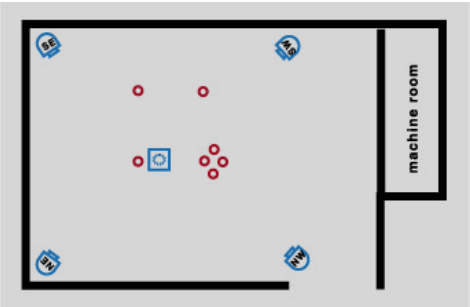

Figure 1 Sensor placement in the Facilitator Room. The red circles are microphones; the blue semicircles are cameras.

The center camera has a wide-angle lens and can see all the participants at once (see Figure 2). The corner cameras are each pointed at one of the couches (see Figure 3).

2.1 Features

From these sensors, we estimate a variety of features to characterize the behavior of the participants. While ideally we would have detailed tracking information and speech recognition, these are difficult to obtain in a robust way for data sessions that need to go for hours at a time without human intervention. Robustness is key here, sometimes at the cost of detail or precision.

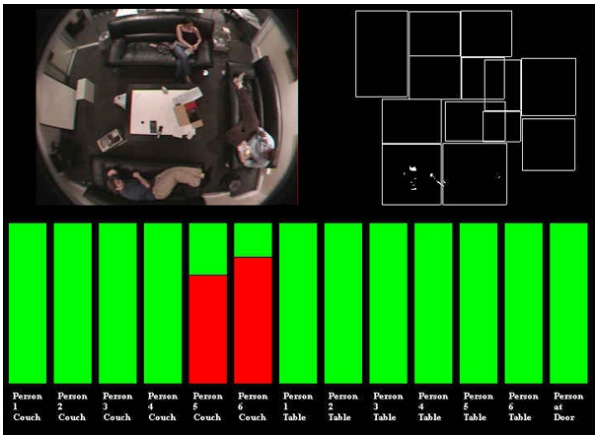

Figure 2: Per-region motion energy in the Facilitator Room

2.1.1 Visual Features

The primary visual feature we are using is per-region motion energy. The regions are marked out by hand from pilot data – each of the six seating positions are "instrumented" with active regions. There is no danger of losing correspondence since the participants stay in the same seats for the duration of the experiments. Though an extremely simple feature, motion energy gives us a concise summary of the level of body language in a given participant.

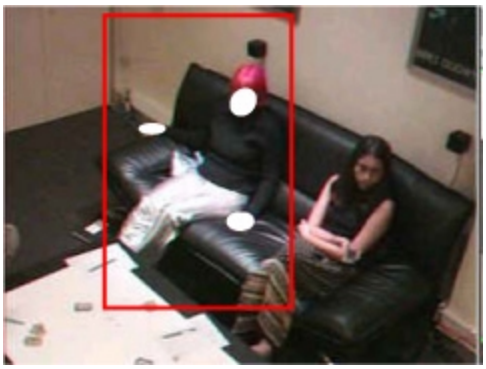

Figure 3: Per-region blob tracking

Another feature we are currently experimenting with is per-region blob tracking. This uses simple flesh tracking [4] on a region of the image marked out by each hand. The result is a much more detailed description of each person's body language. While there is bound to be some overlap in the regions, we can combine this information with the motion energy to see who is doing the moving. Our preliminary tests show that we can get fairly reliable tracking in most situations.

2.1.2 Auditory Features

Before we can make use of the auditory features, we need to know who is speaking – otherwise we do not know which participant to assign the features to. We are approaching this problem in several ways: phase-based source localization, energy from localized microphones near each speakers, and a method in which we treat the speaker location as a classification problem using microphone energies and phase estimates as features. Initial experiments with Hidden Markov Models resulted in ~80% of the speech frames being correctly assigned to one of 6 speakers.

Along with knowing the location of the sound, we wish to know when speech is actually present. For this we use the speech detection algorithm developed in [5]. This method gives us two levels of information – it marks out the voiced segments (i.e., vowels), as shown in Figure

4, and also the utterances, which are groupings of the vowels based on their separation.

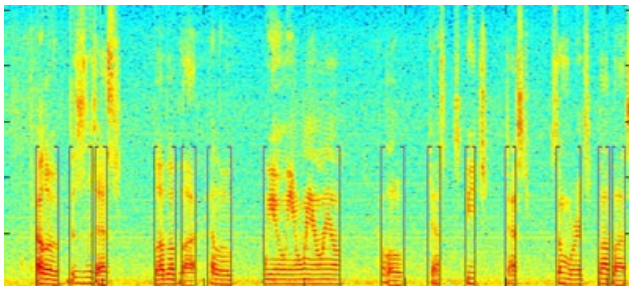

Figure 4: Results of the speech detection algorithm marking out voiced segments

We use the speech detection along with the speech location to decide how to assign auditory features to people – in this way we avoid assigning "false features" to the participants. The speech detection information is also used to compute an estimate of speaking rate. We do this by looking at the number of voiced segments per second. While this is a noisy estimate of rate, it does not require detailed phonemic information as most syllable-rate algorithms do.

Finally, we are also estimating the pitch and pitch variance for each second of speech. We have not used this in our experiments to date, but are planning to in the very near future.

2.2 Actuators

The room is also outfitted with a number of actuators meant to influence the behavior of the participants. Currently we have speakers mounted behind each seat intended to mask sounds with white noise, whisper items to individuals, and so on. We have also installed lights

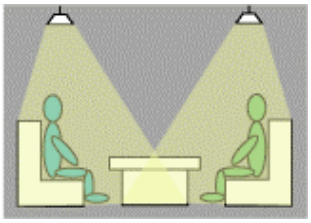

Figure 5: Computer-controlled lighting

focused on each seat whose colors and intensities are under computer control. These are meant to change overall room lighting conditions and also to spotlight individuals to affect others' response to them. (see Figure 5).

There are five projectors in the room: three on the walls, one going to a main screen, and one on the table. These are intended to show relevant information at appropriate times in the hopes of changing the conversation pattern.

In the experiments in this paper, we have only begun to use the potential of this room – at this point we are not using the actuators. However, we cannot study the effects of actuators until we have modeled the baseline interactions among individuals, and thus in this study we focus on the latter.

3 The Influence Model

In his PhD dissertation, Asavathiratham [6] introduced the "Influence Model," a generative model for describing the connections between many Markov chains with a simple parameterization in terms of the "influence" each chain has on other chains. His work showed how complex phenomena involving interactions between large numbers of agents could be simulated through this simplified model. This is very relevant to our scenario; thus in our previous work [1] we extended his model by adding observations and developing a algorithm for learning the parameters from data. We briefly review the highlights of that development here.

The graphical model for the influence model is identical to that of the generalized *N*-chain coupled HMM (see Figure 6), but there is one very important simplification. Instead of keeping the entire $P(S_t^i \mid S_{t-1}^1,...,S_{t-1}^N)$ table, we only keep $P(S_t^i \mid S_{t-1}^j)$ and approximate the full conditional distribution with:

$$P(S_t^i \mid S_{t-1}^1,...,S_{t-1}^N) = \sum_j \alpha_{ij} P(S_t^i \mid S_{t-1}^j)$$

In other words, we form the distribution for a given chain's next state by taking a convex combination of the pairwise conditional probabilities. As a result, we only have N QxQ tables and N α parameters per chain, resulting in a total of $NQ^2 + N^2$ transition parameters. These are far fewer parameters than any of the above models. The real question, of course, is whether we have retained enough modeling power to determine the interactions between the participants.

Asavathiratham refers to the α's as "influences," because they are constant factors that tell us how much the state transitions of a given chain depend on a given neighbor. It is important to realize the ramifications of these factors being constant: intuitively, it means that *how much* we are influenced by a neighbor is constant, but *how* we are influenced by it depends on its state. Another way to look at this is that we are only modeling the first-order effects of our neighbors' influences on us: if Joe yelling causes us to be quiet with certainty and Mark's yelling causes us to yell back with certainty and our α's for both are equal, the combination of both yelling will result in a distribution of our next action that has its probability mass equally distributed over yelling and not yelling. This is what we are giving up in terms of

modeling power while the fully-connected coupled HMM would allow us to explicitly model the effect of the joint event of Joe and Mark yelling together, the influence model does not (note, however, that the set of pairwise coupled HMMs would also not be able to model this joint effect).

This simplification seems reasonable for the domain of human interactions and potentially for many other domains. Furthermore, it gives us a small set of interpretable parameters, the α values, which summarize the interactions between the chains. By estimating these parameters, we can gain an understanding of how much the chains influence each other.

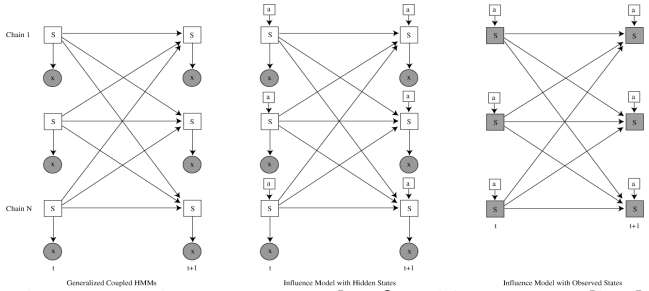

Figure 6: Figure 6: Graphs for (a) a generalized coupled HMM, (b) an Influence Model with hidden states, (c) an Influence Model with observed states.

3.1 Learning for the Influence Model

Despite the parameterization, performing inference on the Influence Model is still intractable using exact methods. We are currently investigating a number of approximate inference mechanisms to combat this problem. For our current work, we have decided to simplify the estimation problem by allowing the states S_t^i to be observed for each chain. We obtained our state sequences by fitting an HMM to each chain's observations and performing a Viterbi decoding (we will discuss the potential issues with this approach later). The chain transition tables were then easily estimated (by frequency counts) directly from these state sequences. Since our goal is to estimate the inter-chain influences (via the α_{ij}'s) this "clamping" of the observation and chain transition parameters helps combat the over fitting problems of the full model.

We can now easily optimize for the α values using gradient descent. Let us first examine how the likelihood function simplifies for the observed Influence Model:

$$P(S \mid \{\alpha_{ij}\}) = \left(\prod_i P(S_0^i) \right) \prod_i \prod_t \sum_j \alpha_{ij} P(S_t^i \mid S_{t-1}^j)$$

Converting this expression to log likelihood and keeping only the terms relevant to chain *i*, we have:

$$\alpha_{ij}^* = \arg\max_{\alpha_{ij}} \left[\sum_t \log \sum_j \alpha_{ij} P(S_t^i \mid S_{t-1}^j) \right]$$

This *per chain* likelihood is concave in α_{ij} (see [1] for details. Now taking the derivative w.r.t. α_{ij}:

$$\frac{\partial}{\partial \alpha_{ij}}(.) = \sum_t \frac{P(S_t^i \mid S_{t-1}^j)}{\sum_k \alpha_{ik} P(S_t^i \mid S_{t-1}^k)}$$

Notice that the gradient and the *per chain* likelihood expression above are inexpensive to compute, $O(TN)$. This along with the facts that the *per chain* likelihood is concave and the space of feasible α_{ij}'s is convex means that this optimization problem is a textbook case for constrained gradient ascent with full 1-D search (see p.29 of [7]). Furthermore, in all examples in this paper, 20 iterations were sufficient to ensure convergence, which amounted to less than 10 seconds of CPU time.

4 Evaluating the Influence Model

In this section, we describe the performance of our model on a synthetic dataset and also on data collected from real human interactions in a conversational setting. We believe that the Influence Model can outperform Generalized Coupled HMMs and be a useful tool for modeling interactions for the following reasons: first, the influence parameter specifically models the strength of the influence that one chain's dynamics has on the other. As mentioned earlier, this influence strength could be a very informative measure for human interactions. Next, the Influence Model has exponentially fewer parameters than the Generalized Coupled HMM, and therefore the model can be learned with far less training data.

To evaluate the effectiveness of our learning algorithm we first show results on synthetic data. The data was generated by an Influence Model with three chains in lock step: a leader chain, which was evolving randomly (i.e., flat transition tables), and two followers who meticulously followed the leader (i.e., an influence of 1 by chain 2 and a self-influence of 0). This can be thought of as an idealized version of the situation where audience members are "synchronized" with the speaker, nodding their heads as he does, etc. We sampled this model to obtain a training sequence which was then used to train another randomly initialized Influence Model. As described above, the $P(S_t^i \mid S_{t-1}^j)$ distributions were estimated by counting and the α_{ij}'s by gradient ascent. Note how this model learns the "following" behavior –

chains 1 and 3 follow chain 2 perfectly. The alpha matrix captures the strength of chain 2's dynamics on chains 1 and 3 very well. The learned alpha matrix is

$$\begin{bmatrix} 0.0020 & 0.9964 & 0.0017 \\ 0.2329 & 0.4529 & 0.3143 \\ 0.0020 & 0.9969 & 0.0011 \end{bmatrix}$$

where the rows represent who is influenced and the columns represent the influencer. We also trained a Generalized Coupled HMM on this data with the EM algorithm, using the Junction Tree Algorithm for inference. Again we sampled from the lock step model and trained a randomly initialized model. In this case, the learned model performed reasonably well, but was unable to learn the "following" behavior perfectly due to the larger number of parameters it had to estimate ($P(S_t^i \mid S_{t-1}^1,...,S_{t-1}^N)$ vs. $P(S_t^i \mid S_{t-1}^j)$).

In order to measure the effect of training data size on the quality of the model, we test how well the learned model predicts the next state of the follower chains given the current state of all the chains for different size training set. Table 1 shows the prediction results for both the Influence Model and the Generalized Coupled HMM (GCHMM). Clearly, the Influence Model requires a lot less training data to model the dynamics accurately.

Training Data Size	Influence Model Chain #		GCHMM Chain #	
	1	3	1	3
10	100%	99.5%	67%	50.5%
20	99.5%	100%	66%	90.5%
50	100%	100%	100%	100%
100	100%	100%	100%	100%

Table 1 Prediction Results for the follower chain in the synthetic dataset

After verifying the performance of our algorithm on synthetic data, we tested our models on data of natural human interactions in the facilitator room. We recorded two hours of data of five participants playing an interactive debating game. The game, Opinions, comes with stack of cards that has different controversial debate topics. We recorded ten games (debate sessions) for our experiment. In order to ensure that we saw a debate session between all possible pairs of players, we listed all pairs and chose pairs from the list without replacement. The first participant in the list entry rolled a die to pick a side (proponent or opponent). Each debater spoke for one minute after which the stage was open for discussion between all participants. No restrictions were imposed on the participants' interaction style during the game. The features calculated automatically from the data were per person motion energy (30 Hz), speech energy (30 Hz), and voicing state (60 Hz). Also, the speaker turns (i.e., who was speaking when) were hand labeled for all the games.

In the first experiment, we used the hand-labeled speaker turns only. Each player had two states – speaking and silent. When multiple players were speaking at the same time, all of them were considered to be in the speaking state. The full set of features for the game was the binary state vector for all of the players, which was

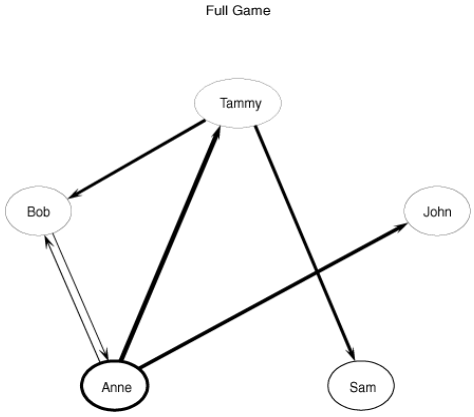

Figure 7: Influence Graph for full game showing strong links for the dominating speakers Tammy and Anne.

afterwards non-uniformly resampled in order to remove consecutively repeating states. Therefore, if all the players were in the same state for *t* timesteps, those *t* identical observations were effectively replaced with one time step. This effectively broke up the data such that there would be one feature vector per conversational turn. If the features were not resampled in this way, the self-transitions would overwhelm the effects of any inter-person influences.

We estimated the influence matrix α for the entire dataset (all ten games) and also for each game separately. Tammy and Anne were observed to be the dominating speakers in all of the datasets, and this appears in the learned graphs as the strong connections Tammy and Anne have to the other participants. This is true both in games in which they were debaters (Figure 8) and also the overall graph for the entire data session (Figure 7).

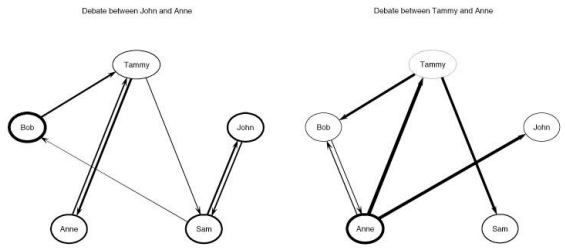

Figure 8: Influence graphs for two debates: (a) John and Anne (b) Tammy and Anne

The one-step prediction accuracy for the natural interaction data is shown in Table 2. Although this result looks promising it is somewhat misleading – since the self-transition probabilities for both speaking and non-speaking states were much higher than the probability of transition out of the states, the model always predicted that it should stay in the same state.

Chain	Prediction Accuracy
Tammy	74%
Bob	80%
Anne	83%
Sam	76%
John	89%

Table 2 One-step prediction accuracy

For the second experiment using the natural interaction data, we learned the Influence Model using the motion energy, speech energy and voicing feature for each person. Viterbi decoding of the individual HMMs obtained the "observed" state labels for the model that was used to learn the α_i values. The α matrix obtained this way was very diagonal, i.e. had very strong self-influence (~1) and very weak influence from other chains (~0).

We believe this is more than anything a result of how we chose our state space. By individually clustering the states of each chain using HMMs, we were ignoring the couplings between the chains, and thus ended up with states that were not helpful in predicting the other chains. We can test this hypothesis by looking at mutual information between a given chain's state and the previous state of all the other chains. This quantity measures the predictability of a given variable from another. For our experiment, we expect this would be quite low. We plan to counteract this problem by incorporating this metric into our clustering, using the methods described in [8].

5 Conclusion

While the Facilitator Room is still a long way from facilitating human interactions, we are already beginning to see some interesting results on the path to behavior analysis. We feel the influence model is interesting in itself – it is a powerful parameterization that can quickly learn meaningful parameters. Once we can learn a good state space, it should be a powerful tool for analyzing human interactions.

The next steps open to us now are many – our first goal is to continue working with the Influence Model on the feature data, attempting the new state-clustering mechanisms we described above. In addition, we want to try to predict certain kinds of events – speaker changes, gaps in the conversation, interruptions, etc. Finally, we want to start using our actuation mechanisms, estimating their influences, and putting together predictive controllers that can drive/avoid certain kinds of behavior.

References

1. Basu, S., Choudhury, T., Clarkson, B., and Pentland, A., *Learning Human Interactions with the Influence Model*. 2001, MIT Media Lab: Cambridge, MA.
2. Solomon, H., *Mathematical Thinking in the Measurement of Behavior: Small Groups, Utility, Factor Analysis. A Study of the Behavior Models Project*. 1960, Glencoe, IL: Free Press.
3. Gladwell, M., *The Tipping Point: How little things make can make a big difference*. 2000, New York: Little Brown.
4. Wren, C., Azarbayejani, A., Darrell, T., and Pentland, A., *Pfinder: Real Time Tracking of the Human Body*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1997. **19**(7): p. pp 780-785.
5. Basu, S., Clarkson, B., and Pentland, A. *Smart Headphones: Enhancing Auditory Awareness through Robust Speech Detection and Source Localization*. in *ICASSP*. 2001. Salk Lake City, UT.
6. Asavathiratham, C., *The Influence Model: A Tractable Representation for the Dynamics of Networked Markov Chains*, in *Dept. of EECS*. 2000, MIT: Cambridge. p. 188.
7. Bertsekas, D.P., *Nonlinear Programming*. 1995, Belmont: Athena Scientific.
8. Tishby, N., F. Pereira, and W. Bialek, *The Information Bottleneck Method*, The 37th annual Allerton Conference on Communication, Control, and Computing, pp. 10, 1999.