ICASSP 2010, Dallas TX

PREDICTING INTERRUPTIONS IN DYADIC SPOKEN INTERACTIONS

Chi-Chun Lee and Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory (SAIL)
University of Southern California, Los Angeles, CA 90089, USA

ABSTRACT

Interruptions occur frequently in spontaneous conversations, and they are often associated with changes in the flow of conversation. Predicting interruption is essential in the design of natural human-machine spoken dialog interface. The modeling can bring insights into the dynamics of human-human conversation. This work utilizes Hidden Condition Random Field (HCRF) to predict occurrences of interruption in dyadic spoken interactions by modeling both speakers' behaviors before a turn change takes place. Our prediction model, using both the foreground speaker's acoustic cues and the listener's gestural cues, achieves an F-measure of 0.54, accuracy of 70.68%, and unweighted accuracy of 66.05% on a multimodal database of dyadic interactions. The experimental results also show that listener's behaviors provides an indication of his/her intention of interruption.

Index Terms— Interruption, Prediction, Hidden Conditional Field, Dyadic Interaction

1. INTRODUCTION

During dyadic spontaneous human conversation, interruptions occur frequently and often correspond to breaks in the information flow between conversation partners. Accurately predicting such dialog events not only provides insights into the modeling of human interactions and conversational turn-taking behaviors but can also be used as an essential module in the design of natural human-machine interface. Further, we can capture information such as the likely interruption conditions and interrupter's signallings by incorporating both conversation agents in the prediction model (we define in this paper the *interrupter* as the person who takes over the speaking turn and the *interruptee* as the person who yields the turn). This modeling is predicated on the knowledge that conversation flow is the result of the interplay between interlocutor behaviors. The proposed prediction incorporates cues from both speakers to obtain improved prediction accuracy.

Several previous works [1, 2, 3] have analyzed different aspects of interruption in human dialogs in terms of prosodic, gestural, and lexical cues exhibited under different conditions of interruption. This work is novel in the sense that it utilizes the information that happens before a turn change occurs to perform *prediction* of interruption rather than just *recognition*. Our hypothesis is motivated by a smiliar theory discussed in [4, 5] that intentions of speakers are transmitted multimodally. Hence during an interaction, the interrupter would exhibit different nonverbal behaviors while preparing to interrupt than when participating in coordinated smooth turn-taking conversation. This study relies on verbal behaviors of the interrupter and nonverbal behaviors of the interruptee. Many of our gestural cues are extracted with intuitively higher-level implication, such as mouth opening, raising of eyebrows, and rigid head motions, using direct motion capture data. They provide interpretable results and offer guidances for future efforts on automatic video feature extraction. Further, discriminant models have been shown to outperform generative models in several classification tasks, and the model assumption on the independence of observation across time is more relaxed. We utilize the Hidden Conditional Random Field (HCRF) [6], a dynamic discriminant model, for the interruption prediction task.

The IEMOCAP database [7] was used in the present study. It provides detailed information on different modalities (speech, gestures of face, head motions, and hand movements) expressed in natural human-human conversational settings. Furthermore, in order to cover more general cases of interruptions, interruptions were annotated based on human judgement instead of syntactic structure based solely on instances of overlapping speech [8]. The proposed prediction model achieves F-measure of 0.54, accuracy of 70.68%, and unweighted accuracy (average per class accuracy) of 66.05% by using acoustic cues from the interruptee and gestural cues from the interrupter for the duration of one second before turn change happens.

The paper is organized as follows; our research methodology is described in Section 2, experiment results and discussion are presented in Section 3, and conclusion and future work are given in Section 4.

2. RESEARCH METHODOLOGY

2.1. Database and Annotation

We used the IEMOCAP database for the present study [7]. It was collected for the purpose of studying different modalities in expressive spoken dialog interaction. The database was recorded in five dyadic sessions, and each session con-

The work was supported in part by funds from NSF and Army

sists of a different pair of male and female actors both acting out scripted plays and engaging in spontaneous dialogs in hypothetical real-life scenarios. In this paper, we are interested in the spontaneous portions of the database since they closely resemble real-life conversation. During each spontaneous dialog, 61 markers (two on the head, 53 on the face, and three on each hand) were attached to one of the interlocutors to record (x, y, z) positions of each marker. Figure 1 illustrates the placement of the markers. The markers were then placed onto the other actor and recorded again with the same set of scenarios to complete a session. The recorded speech data from both subjects were available for every dialog. The database was transcribed and segmented by humans, and time boudaries resulting from the automatic forced alignment are assumed to correspond to the actual speech portion of each subject.

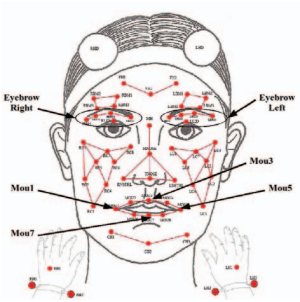

Fig. 1. Markers Placement.

We used the Anvil software [9] as our annotation tool as it provides a multimodal annotation interface. Our interruption annotation scheme is based on subjective judgement rather than the syntactic structure. Interruption was labeled if the utterance made by the interrupter was to intentionally stop the interruptee's flow of speech. Annotator was instructed to be aware that an interruption can happen without occurrence of overlapping speech, and an overlapping speech instance that is cooperative in nature should be noted as smooth transition. In total, we annotated 1763 turn transitions in which 1558 were smooth transitions and 215 were interruptions. Since the distribution of these two types of turn transitions is highly unequal, we downsampled the data by including only three sessions (six subjects) of the IEMOCAP database with three dialogs chosen for each recording session. Subjects and dialogs were selected to include a majority of the annotated interruptions. In total, there are 382 turn transitions annotated with 130 interruptions and 252 smooth transitions used as our dataset in this paper.

2.2. Feature Extraction

For every given turn transition, we extracted two sets of features; one corresponds to the interrupter's body gestural cues, and another one corresponds to the interruptee's acoustic cues. The features were calculated for a total of one second in duration before the interrupter starts speaking at 60 frames per second. We assume that the duration captures relevant behaviors associated with turn taking. Only acoustic cues were extracted from the interruptee because there are no markers placed on the interruptee, and only gestural cues were extracted from the interrupter because the interrupter has not started speaking during the time of interest.

2.2.1. Interrupter Gestural Features

The following features were extracted for the interrupter.

- Mouth opening distances denoted as (M_z, M_x)
- First-order polynomial parametrization for right and left eyebrows denoted (A_r, B_r, A_l, B_l)
- Six degress of rigid head motion - pitch, roll, yaw, translation in x, translation in y, translation in z denoted as (P, R, Y, T_x, T_y, T_z)

M_z was calculated as absolute distance between markers Mou3 and Mou7 as shown in Figure 1, and M_x was calculated as distance between markers Mou1 and Mou5. The eyebrow's shape was parametrized by a linear equation for each frame ($Z = A^* \times X + B^*$). In our preliminary experiment, second-order polynomial parametrization resulted in a negligible coefficient for the X^2 term. We only considered the (x, z) direction. People rarely have eyebrow movement in the y direction that is the forward and backward direction in our database after normalization of head movements. A_r and A_l are the slopes of the polynomial calculated from the right and left eyebrow marker positions respectively; B_r and B_l are the intercepts. The slope and intercept can be easily associated with tilting and raising of eyebrows. T_x, T_y and T_z were derived from the nose marker, and P, R, Y were computed from all the markers using a technique based on Singular Value Decomposition (SVD) [7].

2.2.2. Interruptee Acoustic Features

The interruptee's energy and pitch values (denoted as E, F) were calculated using the Praat toolbox [10] at 60 frames per second during the same time windows described previously.

Concatenation of the two sets of features along with deltas computed from interrupter's eyebrow parametrization, mouth opening distance, and interruptee's acoustic cues resulted in a 22-dimensional feature vector to serve as the observation inputs for our prediction model.

2.3. Review of Hidden Conditional Random Field

Details of HCRF are described in [11]. An HCRF models the conditional probability of a class label y given a set of observation vectors $\mathbf{x}$ in terms of the Equation 1,

$$P(y|x,\theta) = \sum_s P(y, \mathbf{s}, |\mathbf{x}, \theta) = \frac{\sum_s e^{\Psi(y,s,x;\theta)}}{\sum_{y' \in Y, \mathbf{s} \in S^m} e^{\Psi(y', \mathbf{s}, \mathbf{x}; \theta)}} \quad (1)$$

where $\mathbf{s}$ corresponds to hidden states in the model which captures the underlying structure of each class label, and the potential function $\Psi(y, \mathbf{s}, \mathbf{x}; \theta)$ parameterized by θ is a measure

Table 1. *Summary of Experiment I*

Model	**F-Measure**	Accuracy	Unweighted	Precision	Recall
Chance	N/A	65.96%	50.00%	N/A	N/A
Logistic Regression	0.39	68.06%	58.85%	0.56	0.30
HCRF w/o Feature Selection	0.48	64.66%	60.37%	0.48	0.47
HCRF w/ Feature Selection	**0.54**	**70.68%**	**66.05%**	**0.57**	**0.51**

of compatibility between a label y, a set of observations $\mathbf{x}$ and a configuration of hidden states $\mathbf{s}$.

The following objective function is used in [6] to train the parameters of the model using a hill-climbing optimization technique called the Broyden Fletcher Goldfarb Shanno (BFGS) method,

$$L(\theta) = \sum_{i=1}^{n} \log P(y_i|\mathbf{x}_i, \theta) - \frac{1}{2\sigma^2}||\theta||^2 \quad (2)$$

where n is the number of training sequences. The first term is the log-likelihood of data, and the second term assumes a Gaussian priors with regularization factor, σ^2, on parameters, θ. The optimal parameter is obtained as, $\theta^* = \text{argmax}_\theta\, L(\theta)$.

At testing stage, for a new sequence $\mathbf{x}$ given the optimal parameters obtained from the training data θ^*, we can assign the label of the sequence using Equation 3 through standard belief propagation techniques,

$$\underset{y \in Y}{\text{argmax}}\, P(y|\mathbf{x}, \theta^*) \quad (3)$$

3. EXPERIMENT RESULTS AND DISCUSSION

Two experiments were set up to evalute the performance of the interruption prediction model with the following goals.

- **Experiment I**: Compare dynamic modeling of HCRF with the static model using a Logistic Regression Model. Further optimization of prediction performance through feature selection
- **Experiment II**: Compare interrupter-only model, interruptee-only model, and the optimized model

For both experiments, we performed *z-normalization* with respect to speaker identity, and this normalization makes our feature vectors across speakers comparable. We also performed a six-fold (leave-one-subject-out) cross validation to evaluate the performance. The label that annotates whether the turn-transition utterance is an interruption or smooth transition served as ground truth for computing different prediction metrics. Since the database is skewed toward smooth transitions, several different metrics other than accuracy percentage, such as unweighted accuracy, F-measure, precision and recall, are reported below. F-measure is our primary measure to assess the performances of our prediction model. Training and testing were both done using HCRF library [12].

3.1. Experiment I

In Experiment I, three prediction models were trained. We first trained a baseline model using logistic regression because it can be seen as a static version of the discriminant model. The baseline model was trained with the full 22-dimensional feature vector on every frame of the training sequences given the class label. At testing, the decision was made with majority vote over the frames. The second model was obtained by training an HCRF model with the full 22-dimensional feature vector, and the number of hidden states and the regularization factor were set to be 4 and 1 empirically. Lastly, forward feature selection was performed through an inner five-fold cross validation for each of the six fold validations. We selected features that optimize the accuracy percentage on the inner five-fold cross validation for every given fold. The third model was trained using the final feature set, which was the union of the features selected in each of the six folds, with the number of hidden states set to 4, and the regularization factor set to 1. Results are shown in Table 1.

The best performing model is HCRF with feature selection, which obtains an F-measure of 0.54 with 70.68% accuracy and 66.05% unweighted accuracy. The results indicate that dynamic modeling improves prediction accuracy. Specifically, HCRF without the Feature Selection model obtains a 23.1% relative improvement in F-measure over the Logistic Regression model.

3.2. Experiment II

Experiment II was performed by training interrupter-only and interruptee-only HCRF models to compare with the best performing model - a combination of features from both speakers after feature selection. The interrupter-only model used an 18-dimensional feature vector corresponding to the interrupter's gestural cues, and the interruptee-only model used a 4-dimensional feature vector corresponding to the interruptee's acoustic cues. The number of hidden states was set to 4 with the regularization factor being 0.1 in the interrupter-only model and 1 in the interruptee-only model. Table 2 shows a summary of results from Experiment II.

As results indicate in Table 2, the best performing model in terms of F-measure is the one that models both speakers' behaviors. In particular, the combination of models improves 20% and 31% relatively in F-measure compared with interruptee-only and interrupter-only models, respectively. The combination model with the full 22-dimensional

feature vector listed in Table 1 also has a relative 6.7% and 17.1% higher F-measure compared with interruptee-only and interrupter-only models, respectively.

Table 2. *Summary of Experiment II*

Model	**F-Measure**	Accuracy	Unweighted
Chance	N/A	65.96%	50.00%
Interrupter-only	0.41	64.66%	57.57%
Interruptee-only	0.45	68.59%	61.11%
Optimized	**0.54**	**70.68%**	**66.05%**

3.3. Discussion

We can gain some insights by examining the features selected along with the performance summary. Table 3 shows the feature selected for each fold and the feature set used to generate the final prediction model.

The first thing to notice in Table 3 is that an energy-related feature from the interruptee is always selected as one of the features. This is not surprising because the abrupt jump-in during the interruptee's speech correlates highly with what people perceive as an interruption, while a smooth transition often accompanies a pause between speaker turns. Indeed, if we look at Table 2, using interruptee-only acoustic features alone shows improvement in unweighted accuracy compared to chance.

The more interesting phenomenon is that the feature selection process also selected some of the intuitive interrupter's gestural features, such as mouth-opening and head rigid movement. In fact, Table 2 shows that by using interrupter-only cues, we still obtain an improvement in unweighted accuracy compared to chance. This implies that the background listener's behaviors provide information on his/her own intention of interrupting.

Table 3. *Features Selected*

Fold	Interruptee	Interrupter
One	Energy, ΔEnergy	Slope_Right_Eyebrow
Two	Energy	Roll
Three	ΔEnergy	Slope_Right_Eyebrow
Four	ΔEnergy	ΔMouth_Open_z
Five	ΔEnergy	Yaw
Six	Energy	Mouth_Open_z, Translation_x
Final	E, ΔE	$M_z, \Delta M_z, A_r, R, Y, T_x$

In summary, the best prediction model is obtained through a combination of interrupter and interruptee features with F-measure of 0.54, 70.68% accuracy, and 66.05% unweighted accuracy. The result shows that interruption usually happens when the interrupter jumps in during the interruptee's speaking turn. It also shows that the interrupter's gestural behaviors provide information on the intention of his/her interruption. While the prediction work is limited because of the assumption of time boundary availability, the experimental results still show encouraging results in predicting interruptions by monitoring the speaker's interaction in a dialog.

4. CONCLUSION AND FUTURE WORK

Interruptions in dialogs often provide essential information on changes in the conversation flow. Prediction of such an event before it happens can be of great use in human-machine dialog interface. This work investigated the usage of HCRF as the prediction model and obtained promising prediction accuracy by monitoring both interlocutors' behaviors before a turn change occurs. The results reinforce our hypothesis that speakers' multimodal behaviors can be a good predicting indicator of the upcoming speech intention; in particular, the listener's behaviors before turn taking is shown to indicate his/her intention of interruption.

Future work will extend the prediction modeling to predict occurrences of interruption without knowledge of exact turn change boundaries with different fusion to model interlocutors' behaviors. Further inclusion of other features, such as lexical content and dialog acts, should be investigated as they can also provide information on the intention of speakers. Accurate modeling of interruption in a spoken interaction can bring insights into the design of a natural dialog system in terms of differences in behaviors under various turn-taking structures. This could also provide improved insights into the study of human-human conversations.

5. REFERENCES

[1] C.-C. Lee, S. Lee, and S. Narayanan, "An analysis of multimodal cues of interruption in dyadic spoken interactions," in *Interspeech*, Brisbane, Australia, 2008.

[2] L.-C. Liang, "Visualizing spoken discourse: prosodic form and discourse function of interruptions," in *Second SIGdial Workship on Discourse and Dialog*, vol. 16, 2001, pp. 1–10.

[3] F. Yang and P. Heeman, "Avoiding and resolving initiative conflicts in dialog," in *NAACL HLT*, Rochester, NY, Apirl 2007.

[4] D. MacNeill, *Hand and Minds: What Gestures Reveal about Thoughts*. Chicago, IL: University of Chicago Press, 1992.

[5] D. Haylan, "Challenges ahead. head movements and other social acts in conversation," *AISB*, 2005.

[6] S. B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *IEEE Computer Society Conference and Computer Vision and Pattern Recognition*, vol. 2, 2006, pp. 1521–1527.

[7] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, pp. 335–359, 2008.

[8] D. G. Okamoto, L. Rashotte, and L. Smith-Lovin, "Measuring interruptions: Syntactic and contextual method of coding conversation," *Social Psychology Quarterly*, vol. 65, no. 1, pp. 38–55, 2002.

[9] M. Kipp, "Anvil - a generic annotation tool for multimodal dialogue," in *Eurospeech*, 2001, pp. 1367–1370.

[10] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.03) [Computer program]," March 2009. [Online]. Available: http://www.praat.org/

[11] A. Quattoni, M. Collins, and T. Darrell, "Conditional random field for object recognition," *NIPS*, no. 17, 2004.

[12] L.-P. M. and, "Hidden-state conditional random field library." [Online]. Available: http://sourceforge.net/projects/hcrf/