interspeech 2008 incorporating SST 08

An Analysis of Multimodal Cues of Interruption in Dyadic Spoken Interactions

Chi-Chun Lee, Sungbok Lee, Shrikanth S. Narayanan

Speech Analysis and Interpretation Laboratory (SAIL)
Electrical Engineering Department
University of Southern California, Los Angeles, CA 90089, USA
{chiclee,sungbokl,shri}@usc.edu

Abstract

Interruptions are integral elements of natural spontaneous human interaction. Both competitive and cooperative interruption serve a distinct role in the flow of conversation. This paper analyzes their differences with features, *change* and *activeness,* employing audio, visual, and disfluency data. These features are able to capture differences between the two types of interruptions better than average feature values of any single modality. Also, discriminant analysis shows that the use of multimodal cues provides a 21% improvement in classification accuracy between the two types of interruptions relative to the baseline while any individual single modality cue does not provide significant improvement.

1. Introduction

Exploring the structure of human conversation has been at the center of studies on human interaction. The turn-taking model [1], which posits that at any point in time only one speaker has the floor to speak, is widely used to study various structures of human conversation, including overlapping talks, interruptions, and back channels. Recently, there has also been an increased interest within the engineering community to capture and model multi-person interactions afforded by advances in audio-visual processing technologies. For example, such data allow for quantitative meta analysis of the conversational dynamics in a meeting [2]. One key aspect of such analyses is the ability to detect and characterize interruptions. In this paper, we study multimodal cues of interruptions in a dyadic interaction setting using audio and visual information.

Interruption, which can be viewed as a deviation from the simple turn-taking model and one that occurs frequently in spontaneous speech, is thought to be an important element in identifying points of interest in human conversation and interaction. According to Goldberg [3], interruptions can be broadly classified into two categories—competitive and cooperative interruptions. Each type of interruption possess a similar form in its discourse characteristics locally, but serves different roles in aiding the information flow between speakers. Occurrences of competitive interruptions are usually disrupting to the flow of conversation between speakers while cooperative interruptions are more supportive to the flow. Therefore, understanding different cues and characteristics of each type of interruption is essential for performing automatic recognition and understanding of conversational interactions and for designing advanced virtual conversation agents. For example, Yang [4] analyzed maximum pitch at the utterance level across these two types of interruptions. While they show that maximum pitch value can be higher for competitive interruptions, their data do not provide statistical significance. Further, they also point out that the variability of pitch and energy value is also largely affected by the underlying emotion and intention of the speaker. Yang and Heeman [5] analyzed pitch and energy for the case of initiative conflict utterances, and they show with statistical significance that energy values are higher for the successful floor taking utterances but not pitch values. One limitation of this study is that it only considers a subset of interruptions.

Human communication often involves transmitting information multimodally, and it has been shown that both head and hand movements also carry significant meanings during human conversation [6,7]. Hence we consider an analysis of multimodal cues to interruptions in conversational interactions. The IEMOCAP database [8] was used because it provides information of different modalities in natural human-human conversational settings. We hypothesize that by analyzing different modalities, we can obtain better insights into natural human communication. Further, events such as disfluencies provide another layer of analysis because an occurrence of disfluency is often associated with increased cognitive processing. In fact, based on this analysis, the classification accuracy between these two types of interruptions substantially improves by combining such multimodal features.

The paper is organized as follows, research methodology is described in Section 2, experiment result and discussion is presented in Section 3, and conclusion and future work is in Section 4.

2. Research Methodology

Our focus is on differentiating between the two types of interruptions—competitive and cooperative. We adopt the definition of interruption given by Zimmerman and West [9] that an interruption is defined operationally as "incursions that are initiated more than two syllables away from the initial and terminal boundary of a unit type". This definition involves less subjective judgment, and is based solely on syntactic information. While it does not cover every possible case of interruptions, it does provide us an easier way for analysis and is also shown in [10] to be adequate for identifying what humans would commonly view as interruptions.

2.1. Database and Annotation

We use the IEMOCAP database for the present study. It was collected for the purpose of studying different modalities in expressive speech. The database was recorded in five dyadic sessions, and each session consists of a different pair of male-female actors acting out scripted plays and spontaneous dialogs in hypothetical scenarios. We are interested in the spontaneous portions of the database. There are eight hypothetical scenarios. During each spontaneous dialog, 61 markers (2 on the head, 53 on the face, and 3 on each hand) were attached to one subject to record (x, y, z) positions of each marker. 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. Emotional evaluation and disfluency described in Section 3.1.3 were also labeled by humans. In order to identify interruption utterances candidate for annotation, we use the automatic forced alignment results. The word boundary is assumed to match the actual speech portion of the subject. Overlapping utterances were extracted from the time stamps of the utterances. Then, the NIST syllabification tool [11] was used to map time stamps of phones into syllables based on word level alignment to find interruption utterances. In this paper,

we use a subset of the IEMOCAP database consisting of 1133 utterances from 16 dialogs with one pair of subject speakers.

All utterances that fit the definition of interruption were marked, and two human evaluators were asked to label each interruption into three categories: competitive interruption, cooperative interruption, and back channels (i.e, "yeah") using the Anvil [12]. The rough guidance of labeling is based on the categorization described in [13]. An example of each type of interruption is given below,

Competitive Interruption
M: simply step to line 2A ma'a|m, I'm sorry.
F: |would you be able to...
Cooperative Interruption
M: downtown, was it beautiful, |of course it was beautiful.
F: |so beautiful, full moon...

Any discrepancies were resolved after discussion between the labelers. Back channellings are excluded from the analysis. Table 1 is a summary of each type of interruption in the data considered. The cell represents the number of interrupting utterances of each actor and the number in the parenthesis corresponds to the number of the utterances where the marker information is available.

Table 1: *Summary of Interruptions*

	Competitive	Cooperative	Total
Male	39 (16)	17 (7)	56 (23)
Female	66 (32)	31 (18)	97 (50)
Total	105 (48)	48 (25)	153 (73)

2.2. Feature Extraction

Three main categories of features are extracted for analysis: acoustic features: intensity, gestural features: hand motions, and disfluency features from the transcriptions. Based on the definition of interruptions, we assumed that the most prominent cues will be in the overlapping portion for each interruption utterance. We also considered pitch value as a feature. However, we found that pitch estimation algorithm fails to capture foreground speaker's pitch accurately in this region because of the cross talk. Therefore, pitch is excluded from this preliminary analysis.

Aside from the feature value statistics, two novel simple measures, *change* and *activeness,* of the dynamic behavior of the feature within the specified time interval were also calculated. *Change* roughly corresponds to the difference between the final state of the feature at the end of the time interval with respect to the initial state. *Activeness* roughly corresponds to the amount of fluctuations of the feature within the the time intervals.

2.2.1. Speech-Intensity

Raw intensity values are obtained every 10 ms with window length 30 ms using Praat tool [14]. The list of intensity features obtained are,

- Maximum intensity at first jump-in word
- Maximum intensity within overlapping region
- Mean intensity at first jump-in word
- Mean intensity within overlapping region
- Intensity *change* at first jump-in word
- Intensity *activeness* within overlapping region

Intensity *change* at the jump-in word is calculated as,

$$I_w = I_{fe_w} - I_{fs_w} \tag{1}$$

where fe_w , fs_w correspond to ending and starting frame of the word. Intensity *activeness* is calculated at the overlapping region using (2),

$$I_{ov} = \frac{\sum_{i=fs_{ov}+1}^{fe_{ov}} |I_i - I_{i-1}|}{T_{e_{ov}} - T_{s_{ov}}} \tag{2}$$

where fe_{ov} , fs_{ov} correspond to the ending and starting frame of the overlapping region, and $T_{e_{ov}}$, $T_{s_{ov}}$ correspond to ending time and starting time of the overlapping region.

2.2.2. Hand Motions

Hand motions features are obtained from the maker information provided in IEMOCAP database. The list of hand motions features obtained are,

- Right hand *activeness* at first jump-in word
- Right hand *change* at first jump-in word
- Left hand *activeness* at first jump-in word
- Left hand *change* at first jump-in word

since there are 3 markers on each hand, the representative marker of each hand is obtained by averaging positions of the markers. The raw hand *activeness* is calculated as,

$$V_k = \frac{\sum_{i=fs_w+1}^{fe_w} \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + (z_i - z_{i-1})^2}}{T_e - T_s} \tag{3}$$

where x_i , y_i , z_i are the coordinate of the representative marker of each hand, $T_e - T_s$ is the duration of a word. This value can be viewed both as the average speed of the hand at the word level or as a fluctuation measure of the positions of each hand during the time interval.

However, since this value varies too much across utterances, in order to obtain an utterance-independent measure, we take the ratio between each V_k and the mean of V_k at the utterance level. This measure can tell us whether each hand is more or less active for this word during the utterance. *Activeness,* r_w ,value is calculated as,

$$r_w = \frac{V_1}{\langle V \rangle} \quad , \quad \text{where,} \quad \langle V \rangle = \frac{\sum_{k=1}^{tot_w} V_k}{tot_w} \tag{4}$$

V_1 corresponds to the the first word of the interrupting utterance calculated using (3) and tot_w is the total number of words in the utterance.

Change of hand movement measures whether the hand has moved significantly from one region to another within a word. To quantify this, we assume that all positions of each hand in a dialog can be separated into four distinct regions. This assumption is based on the observation from the video data. Hence, we use k-means clustering with k = 4 and euclidean distance as distance measure to cluster (x, y, z) of each representative hand marker of all speech segments for each dialog, see Figure 1 .

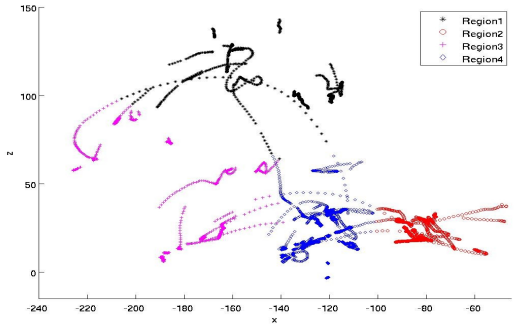

Figure 1: An e*xample of clustering of right hand motions into 4 regions*

Figure 1 is plotted by projecting (x, y, z) onto the (x, z) axis for better view. In this case, right hand's positions are clustered into four regions. Three of them are at

Table 2*: Summary Results of Interrupting Utterances*

	Speech Intensity Value						Hand Motions				Disfluency
	Maximum		Mean		Activeness	Change	Activeness		Change		Number
	Word	*Overlap*	*Word*	*Overlap*			*Left*	*Right*	*Left*	*Right*	
Competitive	67.75	70.8	60.48	61.49	91.92	11.57	2.86	2.55	15	15	34
Cooperative	65.42	68.33	59.38	60.21	71.39	8.00	1.65	2.15	1	3	7
p-value	**0.05**	**0.013**	0.18	0.09	**0.007**	**0.02**	**0.02**	0.25	**0.006**	0.07	**0.016**

approximately the same height but at either left, middle, or right side of the body, and the other one is much higher in height. Since, most of the data that we considered has shown this type of clustering, it shows that our assumption on the finite number of regions that each hand motions resides in is a fair assumption. The *change* of hand movement is then a binary feature: 1 meaning that the hand has moved across clusters within the word, and 0 meaning otherwise. This measure can be seen as the difference between the final state of hand positions and the initial state given the time interval.

2.2.3. Disfluencies

We asked the evaluators to assign a binary feature, 1 meaning an occurrence of disfluency and 0 otherwise, for each given interrupting utterance. We only count a disfluency occurrence when it happens close to the turn transition of the utterance to be consistent with our assumption of the locations of prominent cues. The disfluencies in speech that we adopted are from the categorization in [15] ,

- *false start:* the speaker abandons and utterance or constituent and then start over
- *repetition:* the speaker repeats some part of utterance
- *filled pause:* "er", "um", "uh", "ah"

2.2.4. Feature Normalization

All numerical features are normalized with respect to the neutral utterances of every subject using the following scheme to obtain speaker-independent measures. For a given feature, F , of a speaker, we calculate F_{ref} as the reference value of the feature,

$$F_{ref} = \frac{\sum_{sbj}\sum_{neu} F}{num_{sbj} * num_{neu}} \quad (5)$$

where sbj , neu correspond to subject speakers and neutral utterances. And num_{sbj} , num_{neu} , correspond to number of subject, and number of neutral utterances of each subject. Then the normalized feature F_{norm} is calculated as,

$$F_{norm} = \frac{F}{C_{sbj}} \quad \text{, where} \quad C_{sbj} = \frac{\sum_{neu} F_{sbj}}{num_{neu}} * \frac{1}{F_{ref}} \quad (6)$$

where C_{sbj} is normalizing constant for each subject. This normalization retains emotional information in the feature, while performing normalization at the same time.

3. Results and Discussion

The experiments are set up trying to answer two main questions,

- *Does each feature listed in Section 2 behave differently for the two types of interruptions?*
- *Can we obtain a better discriminating power by incorporating multimodal cues?*

Hypothesis testing (*two sample t-test, two proportions test, fisher's exact test*-for small sample size), and discriminant analysis are performed to address these questions.

3.1. Hypothesis Testing

Our null hypothesis is that the means or proportions of the features are equal for both interruptions, while the alternative states that it is higher for competitive interruptions. Competitive interruptions are more likely to occur when the interrupter is in a higher-emotional state and is likely to be more intrusive than cooperative interruptions. We expect that the features to reflect similar characteristics. Table 2 is a summary of our results based on the features described in Section 2.

3.1.1. Speech-Intensity

The first thing to point out is that neither the difference in the mean intensity at the jump-in word nor at the overlapping region between two types of interruption is statistically significant. We speculate that the competitiveness does not manifest in the average characteristics of intensity. Instead, it seems to be reflected more in the values of maximum intensity. This shows that the prominent cues are more localized than distributed. The intensity *change* and *activeness* features both show statistically significant differences across two types. This results shows that not only the values of intensity, but also the behavior of the intensity values are also more active in competitive than in cooperative interruptions.

3.1.2. Hand Motions

The left column in Table 2 corresponds to the left hand, and the right column corresponds to the right hand. The fluctuation of hand positions and the range of movement of the left hand are significantly higher for competitive interruptions. Note that *change* is tested using *fisher's exact test*, since we do not have enough sample data for *two sample proportions test*. We only have two subjects, so the speaker dependency, such as right/left handedness, of these measures are unknown. But, we can see that hand motion features do show a difference between the two types of interruptions.

3.1.3. Disfluencies

The *p-value* is calculated using *two proportions test*. Table 2 also shows that the number of occurrences of disfluencies is significantly higher in the case of competitive interruptions. We hypothesize that this outcome may due to the role of each type of interruption plays in the flow of conversation, since competitive interruption seems likely to place a heavier load of cognitive processing on the interrupter especially when the original speaker does not yield the turn quickly.

3.2. Discriminant Analysis

Discriminant analysis was performed on three sets of features using the SPSS software,

- *Intensity-only Features*
- *Hand Motions-only Features*
- *Combination of Both Modalities*

Table 3 is a summary of prediction accuracy percentage for each type of interruptions using different set of features. The features repeated are the ones that have shown statistically significance from Section 3.1. From Table 1 there is less available interruption utterances for hand motions, we take

the intensity values of those interrupting utterances where hand motions are available. Also, the prior probability of each type of interruption estimated from the considered dataset are incorporated into discriminant analysis for classification purpose. The Box's M values are all less than 0.05, meaning that the homoscedasticity assumption does not hold. This may be due to the limitation of the number of our data samples, so the covariance is pooled within each group.

Table 3: Summary of Classification Result

	Competitive	Cooperative	Overall
Chance	100.0%	0.0%	65.7%
Intensity-only	89.6%	28.0%	68.5%
Hand Motions-only	54.2%	88.0%	65.8%
Combination	89.6%	60.0%	**79.5%**

3.2.1. Intensity-only

The features used in this analysis are maximum intensity value, intensity *change*, and *activeness*. The Wilk's Lambda value is 0.904 and the discriminant function's significance level is 0.070. The discriminant function is not significant as a whole, but from Table 3, 89.6% of all competitive interruptions are classified accurately, but only 28% of cooperative interruptions are classified correctly. We speculate that both types still possess a certain degree of disruptive characteristic that may still be reflected in the intensity, therefore intensity features alone are not sufficient for recognizing cooperative interruptions.

3.2.2. Hand Motions-only

The features used in this are left hand's *activeness*, *change*, and right hand's *change*. Although right hand's *change p-value* is slightly > 0.05, we still include in this analysis since it is not that far off. The Wilk's Lambda value is 0.859, and the discriminant function's significance is 0.014. From Table 3, hand motions features provide different information than intensity. In this case, an absence of hand motions can almost signal an occurrence of cooperative interruptions, but the confusion still exists since this can also happen in the case of competitive interruptions.

3.2.3. Combination

The Wilk's Lambda value is 0.782 and the significance is 0.10. From Table 3, intensity values and hand motions alone does not provide much gain in overall classification accuracy. However, since each set of features provide a different information, by combining both features, the classification accuracy improves an absolute 14% (21% relative) above baseline to 79.5%. This proves our hypothesis that each modality does carry important but complementary information that humans use to realize their communication goal. Also a step-wise analysis was performed, and the most prominent features in this case are: maximum intensity value and left hand's *change*. By just using these two features, we can obtain an overall classification accuracy of 71.2%, which is higher than using any of the single modality features.

4. Conclusions and Future Work

Occurrence of interruptions are often driven by various factors including emotional state, topic interest, and intentions, etc. This information is carefully encoded with multimodal cues, so a human can understand whether an occurrence of interruption falls into categories of competitive or cooperative interruption and respond appropriately during the conversation. While most of the literature has been focused on single modality, i.e. speech, we have shown in this paper that other information channels, e.g, hand motions and disfluencies, can also be indicative about the type of interruption. This does not only provide us with an improved understanding about natural human interaction, but also helps in monitoring group interaction by incorporating other cues when certain features, such as pitch, can not be accurately estimated for the region of interest. Our analysis presented in Section 3 also shows that the cues are often encoded differently using different modalities. Therefore, by analyzing the combination of multimodal cues, we are able to obtain a better classification result of interruption type.

In this initial study, we have worked only with a subset of the IEMOCAP database due to the amount of human annotation needed to perform detailed analysis. To show with more statistical evidence, we will incorporate more sessions from the database in future work. In addition, we have considered hand motions as the only gestural information in this paper. However, other cues, such as rigid head motions [16] and posture shifts, have also been shown to play important roles in emotional expressions and conversation flow management. Further, in order to better model the interaction of conversation between speakers, speech acts is a good candidate for labeling what the speaker's underlying intentions. There is still an abundance of information embedded in an occurrence of interruption that we have considered. If we can unfold more of this information, we can gather improved insights into designing a natural virtual human or in automatically identifying region of interest in understanding the dynamics of human interactions. These are the goals of our future work.

5. Acknowledgment

This paper was supported in part by funds from NSF and Department of Army.

6. References

[1] S. Duncan, “Some Signals and Rules for Taking Speaking Turns in Conversations”, *Journal of Personality and Social Psychology,* 23:283-92, 1972

[2] C. Busso, P.G. Georgiou, and S. Narayanan, "Real-time monitoring of participants interaction in a meeting using audio-visual sensors," in *ICASSP*, Honolulu, HI, USA, April 2007, vol. 2, pp.685-688.

[3] J. Goldberg. “Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power- and rapport-oriented acts”, *Journal of Pragmatics*, 14, 883-903, 1990

[4] L.c Yang, “Visualizing Spoken Discourse: prosodic form and discourse function of interruptions”, in *Proc. Second SIGdial Workshop on Discourse and Dialog*, v.16, pp 1-10, 2001

[5] F. Yang. and P.A Heeman A., “Avoiding and Resolving Initiative Conflicts in Dialog”, in *Proc. NAACL HLT 2007*, Rochester NY, April 2007, pp 17-24

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

[7] D. Haylen, “Challenges Ahead. Head Movements and other social acts in conversation”, *AISB*, Hertfordshire, UK., 2005.

[8] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database,"Submitted to Journal of Language Resources and Evaluation, 2007.

[9] D.H Zimmerman and C. West, “Sex Roles, Interruptions and Silences in Conversations”, *Language and Sex: Difference and Dominance,* edited by Barrie Thorne, and Nancy Hneley, Rowely, MA: Newbury House, 1975

[10] DG Okamoto, LS Rashotte, L Smith-Lovin, “Measuring Interruption: Syntactic and Contextual Methods of Coding Conversation”, *Social Psychology Quarterly,* vol. 65, no.1, pp 38-55

[11] W.M Fisher, *Syllabification Software,* The Spoken Natural Language Processing Group, National Institute of Standards and Technology, June 1997. http://www.nist.gov/speech/tools/

[12] M. Kipp, “Anvil–a generic annotation tool for multimodal dialogue”, *Proc. of the Eurpspeech,* pp. 1367 – 1370, 2001

[13] H.Z Li., Y-o Yum, R. Yates, L. Aguilera, Y. Mao, and Y. Zheng, “Interruption and Involvement in Discourse: Can Intercultural Interlocutors be Trained?”, *Journal of Intercultural Communication Research,* vol.34, no.4, pp 233-254, Dec. 2005

[14] P. Boersma and D. Weenick, “Praat, a system for doing phonetics by computer,” Institute of Phonetic Sciences of the University of Amsterdam, Netherlands, Technical Report 132, 1996, http://www.praat.org

[15] P. A. Heeman and J. F. Allen.“Speech repair, intonational phrases, and discourse markers: Modeling speakers’ utterances in spoken dialogue”, *Computational Linguistics*,25(4):525–571, 1999

[16] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 3, pp. 1075-1086, March 2007.