How Interlocutors Coordinate with each other within Emotional Segments?

Firoj Alam, Shammur Absar Chowdhury, Morena Danieli, Giuseppe Riccardi
Department of Information Engineering and Computer Science,
University of Trento, Italy
{firoj.alam, shammur.chowdhury, morena.danieli, giuseppe.riccardi}@unitn.it

Abstract

In this paper, we aim to investigate the coordination of interlocutors behavior in different emotional segments. Conversational coordination between the interlocutors is the tendency of speakers to predict and adjust each other accordingly on an ongoing conversation. In order to find such a coordination, we investigated 1) lexical similarities between the speakers in each emotional segments, 2) correlation between the interlocutors using psycholinguistic features, such as linguistic styles, psychological process, personal concerns among others, and 3) relation of interlocutors turn-taking behaviors such as competitiveness. To study the degree of coordination in different emotional segments, we conducted our experiments using real dyadic conversations collected from call centers in which agent's emotional state include *empathy* and customer's emotional states include *anger* and *frustration*. Our findings suggest that the most coordination occurs between the interlocutors inside anger segments, where as, a little coordination was observed when the agent was empathic, even though an increase in the amount of non-competitive overlaps was observed. We found no significant difference between anger and frustration segment in terms of turn-taking behaviors. However, the length of pause significantly decreases in the preceding segment of anger where as it increases in the preceding segment of frustration.

1 Introduction

Behavioral and social signal processing are emerging interdisciplinary areas of research, which combine social science, psychology, and computer science. The aim of the research is to design computational models for processing human behavioral aspects, which can facilitate different domain experts while counseling, consulting and (or) providing services (Narayanan and Georgiou, 2013; Vinciarelli et al., 2009; Pantic et al., 2011; Vinciarelli et al., 2012; Stepanov et al., 2015). The idea is to analyze different overt and covert behavioral signals during social interactions and label them with some short and long term functional aspects (i.e., states and traits) in order to quantitatively measure them. The functional aspects include empathy, politeness, agreement, engagement, uncertainty, competitiveness and other typical, atypical, distressed and affective social behaviors. Using these short and long term states and traits, one can design an informative behavioral profile of an individual from the daily-life interactions. The measured behavioral profile can help to predict the next behavioral outcome/consequence and/or actions of an individual. This kind of behavioral profile can help domain experts in different application scenarios such as call center, health-care and teacher-student interactions.

In the field of social and psychological science, researchers have been trying to understand these functional aspects for a very long time, however, very recently there are attempts to design automatic computational models for real-world applications. Designing such automatic systems for measuring these behavioral and social functional aspects is still infancy due to many different challenges.

One of the important challenges is to understand how different behavioral cues are associated with one another and how we express them in different interaction scenarios. In this study, *we investigated, the coordination of interlocutors behavior in different emotional segments and how conversational turn-taking*

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 728–738, Osaka, Japan, December 11-17 2016.

dynamics are associated with emotional manifestations of the agent and customer. For the study, the conversational coordination between the interlocutors is defined as the tendency of speakers to predict and adjust each other accordingly on an ongoing conversation. We explored the coordination in terms of psycholinguistic features, lexical and turn-taking features using correlation analysis, cosine similarity, and regression analysis, respectively. For this study, we analyzed dyadic human-human spoken conversations, collected from the call centers in the domain of after-sale customer care, which has been annotated with turn-taking dynamics and emotional expressions. The turn-taking dynamics include competitiveness of overlaps, pauses, and lapses among others. Emotional expressions has been annotated for agent and customer separately with agent's emotional state include *empathy*, and customer's emotions include *anger* and *frustration*.

It has been a few decades to the study of automatically recognizing emotion in affective computing, which has been done in the lab as well as in real settings. The study includes classifying Ekman's six basic categorical emotions (Ekman, 1999) or dimensional levels of emotion such as valence and arousal (Russell, 1980). Still, there are challenges to make emotion recognition research in its practical use, which includes lack of publicly available realistic databases, issues of fusing multi-modal information, automatic segmentation, robustness in terms of generalizability across the domain, cross-corpus (Zeng et al., 2009; Schuller et al., 2011). A detailed overview of emotion recognition research in terms of theories, computation models, and relevant applications is provided in (Calvo and D'Mello, 2010).

The study of turn-taking dynamics such as speech overlap has also a long history. One of the first studies on speech overlap, as discussed in (Sacks et al., 1974), suggested that turn changes with overlap is a very rare case and occurs as a result of self-selection, which projects turn endings. Where as a recent study of (Heldner and Edlund, 2010) suggests that overlap is, in fact, a frequent phenomenon and is much more than just a turn-taking signal, which has also been discussed in (Chowdhury et al., 2015b).

There has been a very few study, which explores finding how different turn-taking features are associated with emotional states. The association of turn-management labels, such as grab, accept, backchannel, and emotional states have been studied in (Koutsombogera et al., 2015). The importance of turn-taking information for predicting user-satisfaction in terms of user manifested emotion have been studied in (Chowdhury et al., 2016). They discussed that turn-taking cues significantly helps in the automatic prediction of user-satisfaction. To the best of our knowledge, a very little study have been conducted to examine what actually happens within an emotional segment in terms of turn-taking. In our study, we present a call center conversation corpus (in Section 2) in which we have the manual annotation of emotional states and overlap discourse. Using which we explored the coordination of interlocutors behaviors as our preliminary study, presented in Section 3 and 4, which can shade a light in future for designing automated computational model.

2 Corpus and Annotation

2.1 Corpus Description

The data used for our research is a collection of Italian human-human spoken conversations, sampled from a large set of call center conversations providing after-sales customer care support in the energy sector. We randomly selected these conversations over six months, which were recorded on two separate channels at a sample rate of $8kHz$, 16bits. The average duration of these conversations was 406 seconds. The corpus has been annotated with emotional states such as empathy, anger and frustration, and overlap-discourse such as competitive and non-competitive.

2.2 Annotation of Emotional States

As mentioned earlier, we annotated *empathy* on the agent channel, and *anger* and *frustration* on the customer channel. In the literature, there is a lack of operational definition of empathy. Therefore, we adopted the *modal model* of emotion by Gross (1998) in order to define empathy and design annotation guidelines for the annotators. Gross's modal model is based on appraisal theory, which has been studied by many psychologists for the investigation of emotional states. Appraisal models of emotion suggest that organisms appraise (i.e., evaluate, interpret, explain) events/situations based on the appraisal process in order to determine the nature of ensuing emotion as discussed by Scherer (2000).

According to the *modal model*, *"emotions involve person-situation transections that compel attention, have meaning to an individual in light of currently active goals, and give rise to coordinated yet flexible multisystem responses that modify the ongoing person-situation transection in crucial ways"* (Gross and Thompson, 2007; Gross, 2011). The key idea of the modal model is that emotional states unfold over time, and their response may change the environmental stimuli, and that may alter the subsequent instances of that and other emotional states. It is a useful framework for describing the dynamics of emotional states, which manifests over time, leads to the generation of an emotional sequence from the interlocutors' emotional manifestations. For example, the sequence of emotional states between an agent and a customer could be Frustration (C) → Empathy (A) → Satisfaction (C). A for the agent and C for the customer.

To design the annotation guideline, we have done an extensive analysis of one hundred conversations (more than 11 hours), and selected dialog turns where the speech signal showed the emergence of empathy, basic emotion, such as anger, and complex emotion such as frustration. In our qualitative analysis, we investigated the relevant emotional speech segments, which were often characterized by some perceivable variation in the speech signal. We observed that such variations could co-occur with emotionally connoted words, but also with functional parts of speech, such as adverbs and interjections, which could play the role of lexical supports for the variations in emotional states. We hypothesized that perceivable variations in the speech are a possible signal of an appraisal process. On the basis of those observations, we have designed annotation guidelines whose critical principle was to focus annotators' attention on their own perception of the variations in the speech signal as well as the variations in the linguistic content of the utterances. For example the annotation guidelines include the following recommendations for the annotators: 1) annotating the onset of the signal variations that supports the perception of the manifestation of emotions, 2) identifying the speech segments preceding and following the onset position, and 3) annotating the context (left of the onset) and target (right of the onset) segments with a label of an emotional state (e.g., frustration, empathy, etc.). In addition, the annotation guidelines include operational definitions of emotional states related to the given domain of application. For example, in this annotation task, the operational definition of empathy is defined as "an emotional state triggered by another's emotional state or situation, in which one feels what the other feels or would normally be expected to feel in his situation" (Hoffman, 2008).

The annotation task was performed by two expert annotators who worked on non-transcribed spoken conversations by following the annotation scheme reported above. In this task, the annotation unit is the speech segment. They annotated *Empathy* on the agent channel and *Frustration* and *Anger* on the customer channel. The annotators labeled *Neutral* on the segment that appeared before any emotional segment to define the context, as mentioned earlier. Finally, the annotated corpus includes 1894 customer-agent conversations (210 hours and 23 minutes in total). In order to evaluate the reliability of the annotation we measured inter-annotator agreement on the annotated segments, and obtained an average kappa 0.74. More details can be found in (Danieli et al., 2014; Danieli et al., 2015).

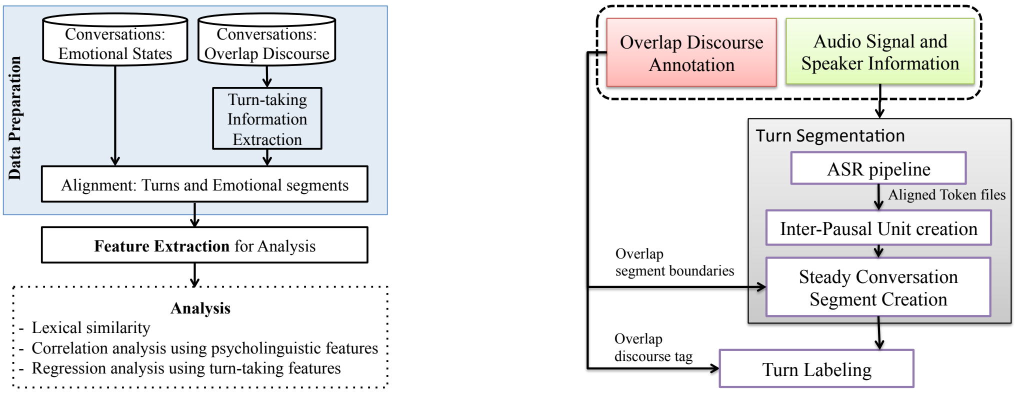
2.3 Annotation of Overlap Discourse

For the annotation of overlap discourse, we selected a subset of 565 conversations with approximately 62 hours of spoken content. Annotators manually segmented overlapping speech, then, categorized and labeled them with the competitive and non-competitive acts. The annotations were performed by two Italian native expert annotators by following the guideline described in (Chowdhury et al., 2015a). The guideline includes Competitive (Cmp) scenarios, in which the intervening speaker (overlapper) starts prior to the completion of the current speaker (overlappee), and both the speakers display interest in the turn for themselves, and also the speakers perceive the overlap as problematic. As for Non-Competitive (Ncm) scenarios, the overlapper starts in the middle of an ongoing turn. No evidence is shown by both the speakers to grab the turn for themselves. The overlapper used the overlap to signal the support for the current speaker's continuation of speech. Both of the speakers perceive the overlap as non-problematic.

The inter-annotator agreement of the annotations is 0.70, which was measured using kappa statistics.

3 Methodology

In Figure 1, we present the experimental system of our study. In the data preparation phase, we selected a subset of conversations in which we have annotations of emotional states and overlap discourse. The turn-taking information extraction system utilized an Automatic Speech Recognition (ASR) system (Chowdhury et al., 2014) to create turn segments and extract turn information (see Section 3.1.1). Later, this information was aligned with the annotations of emotional segments to find the turn-taking information (more details can be found in Section 3.1.2). Using the aligned turn-taking information for an emotional segment, we extracted turn-taking features. We also used turn information to obtain lexical and psycholinguistic features per speaker from the segment. In the analysis phase of our experiment, we investigated lexical similarities and correlation of psycholinguistic features between speakers for different emotional segments. We also used multilevel logistic regression method to understand the association between turn-taking features and emotional segments, and how the association differs from one emotion to another.


Data Preparation
Conversations: Emotional States
Conversations: Overlap Discourse
Turn-taking Information Extraction
Alignment: Turns and Emotional segments
Feature Extraction for Analysis
Analysis
- Lexical similarity
- Correlation analysis using psycholinguistic features
- Regression analysis using turn-taking features
Overlap Discourse Annotation
Audio Signal and Speaker Information
Turn Segmentation
ASR pipeline
Aligned Token files
Inter-Pausal Unit creation
Overlap segment boundaries
Steady Conversation Segment Creation
Overlap discourse tag
Turn Labeling

(a) Experimental pipeline of this study. (b) Turn-taking information extraction system.

Figure 1: System diagrams.

3.1 Data Preparation

For the analytical study, we selected a set of 523 conversations with the manual annotation of emotional states and overlap discourse. This set includes 310 conversations with emotional segments. Among the emotional segments around 11.28% of emotion are annotated as anger, 26.11% as frustration and 62.61% of annotated emotion in agent channel has empathy. From the rest of the 213 conversations, containing no emotional annotations, we selected segments and labeled them with no-emotion (NoEmo).

During the data preparation, we faced two important problems in order to define and align the emotional segment in association with turn-taking discourse: 1) emotional segment are very short in length, which made the task very difficult to get sufficient turn information, 2) an speaker respond to other speaker's emotion with a latency. To overcome these problems, we re-defined the following boundary of manual emotion segment with an impact window of length $2 * d$, where d is the length of the manual annotation of the emotional segment. Hence, the length of our emotional segment is $d + 2 * d = 3 * d$. We also investigated preceding context of each customer's emotional segment and defined it as $Pre.Emo$ with a window of length $3 * d$. The $NoEmo$ segments have been selected from conversations where no emotion in both agent and customer side has been annotated. From the middle of each conversation, we selected and extracted two $NoEmo$ segments with a length of the average emotional segment, (≈ 42 sec). We extracted the $NoEmo$ segments from both agent and customer channels. As mentioned earlier, empathy, Emp, has been annotated in the agent channel only. Thus the preceding context of agent's emotional segment is defined as $Pre.Emp$. Hence, the investigated emotional and non-emotional segments include Pre.Emp, Emp, Ang, Fru, Pre.Ang, Pre.Fru and NoEmo.

3.1.1 Turn-taking Information Extraction

The Turn-Taking Information Extraction System, described in Figure 1 (b), consists of a *turn segmentation and labeling system*. The system uses lexical and manual overlap discourse annotation information to segment and labels the turn types. The pipeline uses the time aligned ASR output as tokens to create Inter-Pausal Units (IPUs) for each input channel. IPUs are defined as the consecutive tokens with no less that 50 ms gaps in between. Using the start and end time information of inter-IPUs and intra-IPUs, we created a steady time line and binary representation (presence or absence of speech information) segments for both the channels. We then defined these segments as *steady conversation segments*. The labels of each segment were then defined by a set of rules. Labels of the segments are as follows:

- Turn (T): Maximal sequences of IPUs where one single speaker has the floor, and none of the IPUs from the interlocutor are present (Beňuš et al., 2011). T_A and T_C represent agent and customer's turns respectively.
- Pause (P): Gaps between the turns of the same speaker with no less than 0.5 sec. P_A and P_C represent agent and customer's pauses respectively.
- Overlap Types Ov= $\{Cmp, Ncm\}$: Overlapping turns between the two interlocutors with competitive or non-competitive intention (see section 2.3 for details).
- Lapse between speakers (L_B): Floor switches between the speakers with a silence duration of 2 sec or more.
- Lapse within speaker (L_W): Gaps between a speakers' turns with a silence duration of 2 sec or more.
- Switch (S): Floor switches between the speakers with silence less than 2 secs or with overlapping frames, not more than 20 ms.

3.1.2 Alignment: Turns and Emotional Segments

For the turn level analysis, it is important to align the turn sequences with the boundary of emotional segments. It is evident from manual annotation that an emotional segment consist of different turn types and not all the turns start inside the boundary. There are some cases where the start/end of emotional episode can be at the middle of a turn. We solved this problem using a rule-based approach. For example, if half of a mismatched turn fall inside an emotional segment we considered that as a part of emotional segment.

3.2 Feature Extraction

3.2.1 Lexical Features

We extracted lexical features from automatic transcriptions from an in-house developed Automatic Speech Recognition (ASR) System (Chowdhury et al., 2014). The word error rate of the system is 31.78% on the test set. To understand the utility of the automated transcriptions with such as error rate, in a different study we compared the performance between automatic and manual transcriptions for a automatic classification of emotions. The results show that performance differences are very low, only 1.2% drop with automated transcriptions (Alam et al., 2016). Therefore, we found that the use of automatic transcriptions are reasonable for the experiment given that manual transcriptions are not available in call cases. For the experiments, the transcriptions of each segment were converted into bag-of-words vectors weighted with logarithmic term frequencies (tf) multiplied with inverse document frequencies (idf). We also reduced the size of the dictionary by removing stop-words and lower frequent words.

3.2.2 Psycholinguistic Features

Psycholinguistic features were extracted from the transcriptions, using Linguistic Inquiry Word Count (LIWC) (Pennebaker et al., 2001). It has been used to study personality, the role of speakers in overlaps (Alam and Riccardi, 2014; Chowdhury et al., 2015b) among other social behaviors in order to understand the correlation between these attributes and word uses. The feature category includes linguistic (e.g., preposition, verb, word count), psychological (affect, positive, negative emotion, anxiety), personal concern (e.g., work, home, money), swear words, relativity among others. The LIWC is a knowledge-based system, which was designed using a set of dictionaries for different languages including Italian.

In the dictionary, each word was labeled with feature categories mentioned above. During the feature extraction process the word in the transcriptions was matched with the dictionary. Then, the matched category was computed as frequency or relative frequency. The Italian version of the dictionary contains 85 word categories (Alparone et al., 2004). We also extracted 5 general and 12 punctuation categories constituting a total of 102 features. We then removed LIWC features that are not observed in our training dataset.

3.2.3 Turn-Taking Features

The turn-taking features were generated using the turn sequence output of the Turn-Taking Information Extraction System, described in Section 3.1.1. The sequences were first aligned with each corresponding emotional segment (see Section 3.1.2). To understand the impact of the choice of turn-taking behavior, we divided the feature sets, at both segment and individual speaker levels, into two groups. A brief description of extracted features, in the segment, are as follows:

- General information about emotional segment (G1):
  - Participation equality, $P_{eq} = 1 - (\frac{\sum_i^N (T_i - T)^2/T}{E})$ where T is the average speech duration of the speakers. T_i is the total speech duration for each speaker. E represents the total speech duration. $N = 2$, represents two speakers as agent and customer inside the emotional segment.
  - Percentage of overlaps.
  - Percentage of Cmp and Ncm on total overlap duration.
- Length of different turn types (G2):
  - Median duration of T_A, T_C, P_A, P_C, Cmp, Ncm, L_W and L_B, inside emotional segment normalized by the median of speaker's respective turn in the whole conversation.

4 Analysis and Results

For different feature sets, we investigated different experimental configurations. For the study of lexical similarities, our experimental conditions include: 1) lexical features from paired (i.e., agent and customer channel from same conversation) speakers' non-overlapping *vs* overlapping turns, 2) lexical features from non-paired (i.e., agent and customer channel extracted from unrelated conversation) speakers' non-overlapping *vs* overlapping turns. Where as for psycholinguistic features, we investigated features obtained from non-overlapping *vs* overlapping turns. For turn-taking features, we have not made any such distinctions. The non-overlapping turns include all the turns of the speakers excluding the overlaps. Where as the overlapping turns includes competitive (Cmp) and non-competitive (Ncm) overlaps.

4.1 Lexical Similarities

For the analysis, we computed cosine similarity of the agent and customer aligned segment representing different emotional states. For the lexical similarity we designed feature vector for agent $\overrightarrow{V_{S_A}}$ and customer $\overrightarrow{V_{S_C}}$ emotional segment using bag-of-word model and transformed them into tf-idf. Then, we computed cosine similarity, $sim(S_A, S_C) = \frac{\overrightarrow{V_{S_A}} \cdot \overrightarrow{V_{S_C}}}{\left|\overrightarrow{V_{S_A}}\right| \cdot \left|\overrightarrow{V_{S_C}}\right|}$ between the feature vector of the agent and customer's segment. For a pair-wise comparison of emotional states, then, we computed mean and standard deviation with statistical significance using t-test.

As mentioned earlier, we have four different experimental configurations for the analysis of lexical similarities. As a baseline, we computed the similarities between non-paired speakers using the lexical features from non-overlapping turns for different emotional segments. The results are presented in a form of similarity map in Figure 2. From the results, we observed that the interlocutors entrain each other in non-overlapping turns when the customer is expressing anger, and the value of similarity ($sim = 0.181$) is significantly ($p < 0.05$) higher than the similarities in any other emotional segment.

In the experiment with competitive overlapping turns, we observed the highest similarity of 0.035 and 0.031 in preceding-anger and anger segments, respectively. In the case of non-competitive overlapping

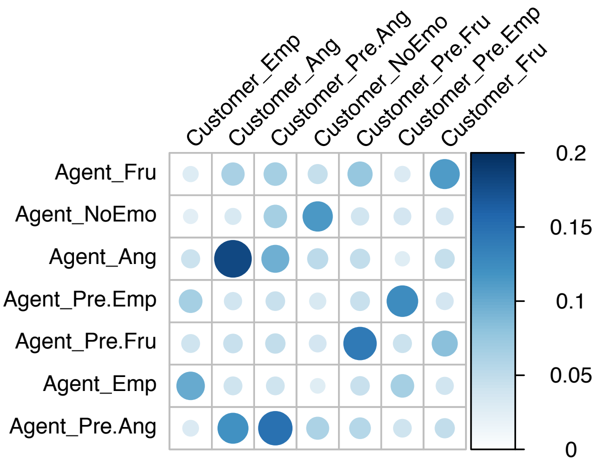


Figure 2: Lexical similarity between the emotional segment of the agent and the customer channel. Pre. represents preceding segments. Ang - anger, Fru - frustration, Emp - empathy, NoEmo - no-emotion

turns, a similarity of 0.034 was observed between the interlocutors in frustration segments. The results on overlapping turns are insignificant.

4.2 Psycholinguistic Features

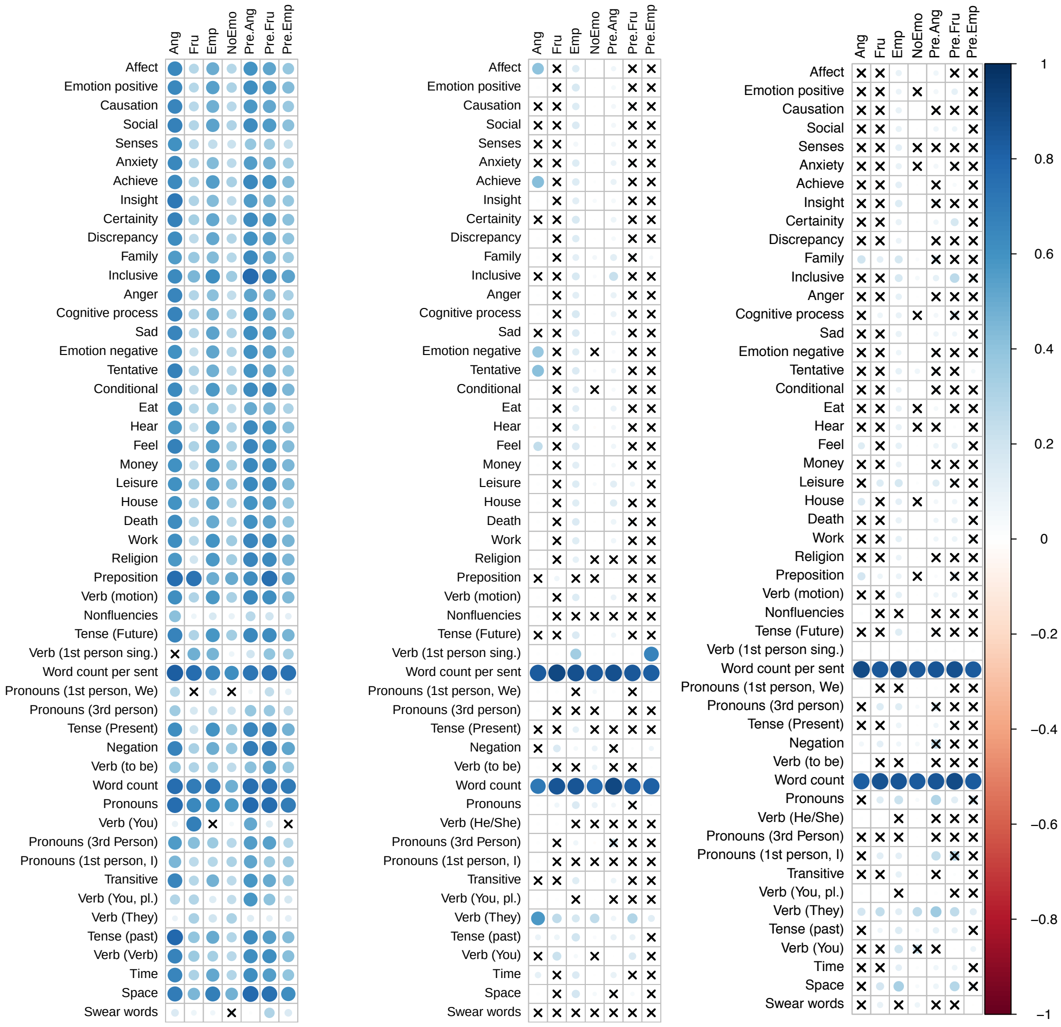
We explored the degree of coordination using Pearson correlation coefficient (r) between the interlocutors' behaviors by correlating psycholinguistic features obtained from overlapping and non-overlapping turns, presented in Figure 3. For the sake of simplicity, the magnitude of r values are presented using colors where as '✕' symbol represent the corresponding r is not significant. These analyses are based on entire emotion segments from the agent and customer channels, irrespective of turns. The r is calculated for each psycholinguistic feature by correlating the agent and customer feature vectors of the conversations. We calculated the significance of the correlation coefficient r using t-test with a degree of freedom equal to $n - 2$, where n represent the total number of instances.

From the correlation plot, it is apparent that the non-overlapping turns of the interlocutors in anger (Ang) segments has high correlation values compared to other emotional segments non-overlapping turns and also compared to overlapping turns (Ncm and Cmp). Not surprisingly the magnitude of the correlation is significantly higher for psychological features like anxiety, affect, and sad between anger segments compared to frustration and empathy segments. Looking at the preceding-anger segments, we observed that the magnitude of r for personal concern along with psychological features are also stronger. It indicates that the cues of anger segment can be found in its preceding segments. The results also show that the uses of pronouns or negation words is directly proportional to the another speaker's usage. We also observed similar patterns in the uses of tenses. The magnitude of r is much higher for past-tense uses in anger compared to others emotional segment and preceding emotional context.

In the case of frustration, the strength of r decrease compared to the preceding segment of frustration. Unlike preceding-frustration segment, we observed that in frustration, there is less coordination between the interlocutors with an exception in preposition and word count features. Though a slight increase in r is observed in verb (they) feature. It is also observed that the interlocutors seem to be more coordinated in the use of swear words in preceding-frustration segments compared to all other segments.

In empathy segments, the coordination of the agent and customer improves compared to preceding-empathy, frustration, and no-emotion segments but the magnitude of coordination is not as impressive as anger segments.

In competitive and non-competitive overlapping turns, a very few significant coordination has been observed. The experiment with non-competitive turns shows that the interlocutors coordinate in anger segment with the features such as affect, achieve, negative emotion, tentative, and verb (they). In the case of competitive overlaps, we observed weak positive correlations between the interlocutors in preceding-frustration segment with feature inclusive, preceding-anger segment with a verb (they), and in empathy segment with space feature.



(a) Non-overlapping segments (b) Ncm overlapping turns (c) Cmp overlapping turns

Figure 3: Correlation analysis at the non-overlapping segment, and overlapping segments, where '✕' symbol represents that the corresponding r is not significant. Pre. represents preceding segments. Ang - anger, Fru - frustration, Emp - empathy, NoEmo - no-emotion

4.3 Turn-taking Features

For the experiment with turn-taking features, we applied a multilevel logistic regression to understand the association of turn-taking features with emotional expressions and how they differ from one emotional state to another. The association of turn-taking features with emotional segments are presented in Table 1, in terms of regression coefficients. In Table 1 (a), the coefficients are reported with respect to the preceding segment of each emotion, where as in Table 1 (b), the coefficients represents the association of each turn-taking feature with the preceding emotion segment *vs.* no-emotion segments.

The results indicates that compared to the preceding context of empathy (Pre.Emp) and no emotion (NoEmo) segments, $participationEquality$, $MedianTurnC$ and $MedianPauseC$ has a negative effect on empathy (Emp) segment, where as $\%Overlap$ and length of non-competitive overlap ($MedianNcm$) has a significant positive effect. Thus indicating the importance of non-competitive overlap in the empathic segment (Emp). The results also hypothesize that during this emotional episode, agents tends to talk more allowing less participation equality between the agent and the customer. The duration of customer's turn and pause tends to be small.

The features $\%Overlap$, the length of overlaps ($MedianNcm$ and $MedianCmp$) has a positive effect for anger segment compared to no-emotion segment. We also observed similar findings for $MedianCmp$ for preceding-anger w.r.t to no-emotion segment. It is observed that the length of non-competitive overlaps ($MedianNcm$) has a positive association where as the length of the lapse between the speakers ($MedianLb$) has a negative effect on anger with respect to preceding context (Pre.Ang). From the result of comparing preceding-anger w.r.t to no-emotion segments, we noticed that the positive association of the length of competitive overlap is present from the preceding context as an indication of anger.

The features $\%Overlap$, the length of overlaps ($MedianNcm$ and $MedianCmp$) has a positive effect for anger segment compared to the no-emotion segment. We also observed similar findings for $MedianCmp$ for preceding-anger w.r.t to the no-emotion segment. It is observed that the length of non-competitive overlaps ($MedianNcm$) has a positive association where as the length of the lapse between the speakers ($MedianLb$) has a negative effect on anger with respect to preceding context (Pre.Ang). From the result of comparing preceding-anger w.r.t to no-emotion segments, we noticed that the positive association of the length of competitive overlap is present from the preceding context as an indication of anger.

Apart from the results presented in Table 1, we also compared the association of turn-taking features with empathy, anger, and frustration with respect to each other. We found no significant difference between anger and frustration segments. However, the preceding context of anger and frustration shows that compared to the preceding-frustration, decrease of pause length is positively associated with preceding-anger segment, especially in agent's side. It is observed that an increase in the length of competitive overlap duration, $MedianCmp$, is positively associated with anger segments w.r.t empathy segments.

Figure 4: Duration distribution of competitive and non-competitive overlaps in different emotional segments.

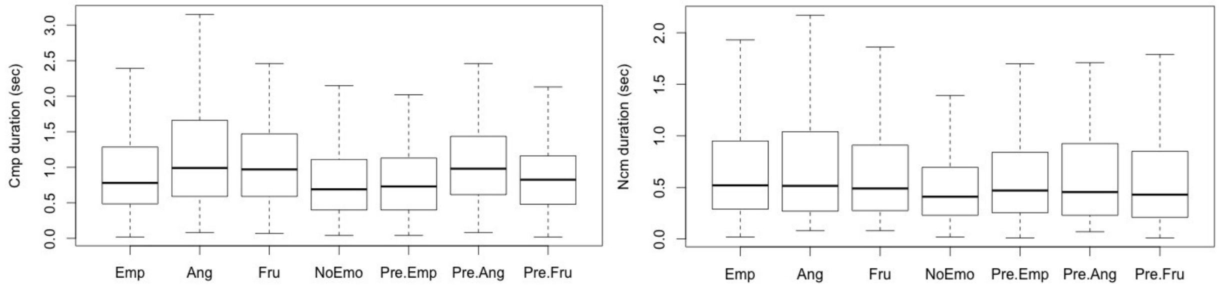


Table 1: Regression coefficient w.r.t preceding segment of each emotion and no-emotion segments.

Groups	**Features**	**(a) Compared to preceding segment**			**(b) Compared to noemo segment**					
		Emp	**Ang**	**Fru**	**Emp**	**Ang**	**Fru**	**Pre.Emp**	**Pre.Ang**	**Pre.Fru**
G1	participationEquality	**-0.948**	0.259	**-1.381**	-0.244	1.018	0.253	**0.823**	0.060	**1.669**
	% Overlap	**0.069**	0.059	**0.131**	**0.099**	**0.112**	**0.083**	0.035	0.046	-0.046
	% Cmp	0.002	0.008	-0.005	0.005	0.015	**0.011**	0.001	0.009	**0.015**
	% Ncm	0.007	0.000	-0.009	**0.010**	-0.002	0.000	0.002	0.000	**0.008**
G2	MedianTurnA	0.001	-0.002	-0.001	0.000	-0.001	-0.001	0.000	0.000	-0.001
	MedianTurnC	**-0.003**	0.001	**0.003**	-0.001	0.002	**0.003**	**0.002**	0.001	0.000
	MedianPauseA	0.001	0.001	-0.004	0.002	-0.005	-0.002	0.001	-0.004	**0.005**
	MedianPauseC	**-0.005**	-0.008	**-0.008**	**-0.003**	0.002	-0.002	0.002	0.003	**0.006**
	MedianCmp	0.001	0.002	0.001	**0.003**	**0.006**	**0.005**	0.002	**0.004**	**0.005**
	MedianNcm	**0.004**	**0.007**	0.002	**0.003**	**0.003**	0.002	0.001	0.001	0.000
	MedianLb	-0.002	**-0.011**	**-0.006**	**-0.005**	-0.004	-0.003	**-0.002**	0.000	0.000
	MedianLw	0.000	-0.002	-0.001	**-0.002**	-0.003	-0.001	-0.001	0.000	0.000

We also compared the duration of competitive and non-competitive overlap within different emotions and preceding emotional segments. In case of competitive, as shown in Figure 4, we observed that duration of mean competitive overlap in anger ($1.25s$) and frustration ($1.09s$) are significantly more compared to the empathy ($0.93s$), no-emotion ($0.80s$) while there is not significant difference between the duration of competitive in anger and frustration segment. In the case of preceding emotion segments, the duration of competitive overlap in frustration is significantly higher than that of preceding-frustration ($0.91s$), where as preceding-anger ($1.12s$) and preceding-frustration is significantly higher than no-emotion. It is also observed that competitive duration in empathy segment is also longer ($p < 0.05$) than no-emotion segments. As for non-competitive duration, shown in Figure 4, there is no significant difference between anger ($0.72s$), frustration ($0.68s$) and empathy ($0.69s$) segment. But it is observed that empathy has significantly longer non-competitive overlap compared to no-emotion ($0.53s$) and preceding-empathy ($0.61s$) segment. Even, the preceding context of empathy (Pre.Emp) has significantly longer non-competitive overlap duration than the non-competitive overlap where there is no emotion. While in anger and frustration, the non-competitive overlap length is significantly higher than the no-emotion segment.

5 Conclusions

In this study, we explored the coordination of interlocutors in different emotional segments using lexical, psycholinguistic and turn-taking features. We investigated such feature sets in terms of regression coefficients, cosine similarity and correlation analysis, respectively. We observed that the interlocutors match each other turns, in terms of lexical similarity and psycholinguistic features, significantly more in anger segment compared to other emotional segments. We also observed that in preceding segment of anger the speakers shows significant correlation with each other in terms of psycholinguistic features. In terms of turn-taking features, no significant differences between anger and frustration have been noticed, apart from the difference in length of pauses in the preceding segment of the emotion. It indicates that preceding context of anger has shorter pause with respect to frustration. Unlike anger, we found less coordination in the segment where the agent is empathic even though an increase in the percentage of non-competitive overlaps has been observed. This is our preliminary study towards utilizing these feature sets for the classification of emotional states and turn-taking discourse, which we will investigate in future.

Acknowledgments

The research leading to these results has received funding from the European Union - Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 610916 - SENSEI - http://www.sensei-conversation.eu/.

References

Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. 2009. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759.

Alessandro Vinciarelli, Maja Pantic, Dirk Heylen, Catherine Pelachaud, Isabella Poggi, Francesca D'Errico, and Marc Schröder. 2012. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87.

Björn Schuller, Anton Batliner, Stefan Steidl, and Dino Seppi. 2011. Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication*, 53(9):1062–1087.

Evgeny A. Stepanov, Benoit Favre, Firoj Alam, Shammur Absar Chowdhury, Karan Singla, Jeremy Trione, Frederic Béchet, and Giuseppe Riccardi. 2015. Automatic summarization of call-center conversations. In *In Proc. of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*.

Firoj Alam and Giuseppe Riccardi. 2014. Fusion of acoustic, linguistic and psycholinguistic features for speaker personality traits recognition. In *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 955–959, May.

Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2016. Can we detect speakers' empathy?: A real-life case study. In *7th IEEE International Conference on Cognitive InfoCommunications*.

F. Alparone, S. Caso, A. Agosti, and A. Rellini. 2004. The italian liwc2001 dictionary. Technical report, LIWC.net, Austin, TX.

James J Gross and Ross A Thompson. 2007. Emotion regulation: Conceptual foundations. *Handbook of Emotion Regulation*, 3:24.

James J Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3):271.

James J Gross. 2011. *Handbook of emotion regulation*. Guilford Press.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161.

Klaus R Scherer. 2000. Psychological models of emotion. *The neuropsychology of emotion*, 137(3):137–162.

Maja Pantic, Roderick Cowie, Francesca D'Errico, Dirk Heylen, Marc Mehu, Catherine Pelachaud, Isabella Poggi, Marc Schroeder, and Alessandro Vinciarelli. 2011. Social signal processing: the research agenda. In *Visual analysis of humans*, pages 511–538. Springer.

Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Martin L Hoffman. 2008. Empathy and prosocial behavior. *Handbook of Emotions*, 3:440–455.

Maria Koutsombogera, Dimitrios Galanis, Maria Teresa Riviello, Nikos Tseres, Sotiris Karabetsos, Anna Esposito, and Harris Papageorgiou. 2015. Conflict cues in call center interactions. In *Conflict and Multimodal Communication*, pages 431–447. Springer.

Morena Danieli, Giuseppe Riccardi, and Firoj Alam. 2014. Annotation of complex emotion in real-life dialogues. In Roberto Basili, Alessandro Lenci, and Bernardo Magnini, editors, *Proc. of 1st Italian Conf. on Computational Linguistics (CLiC-it) 2014*, volume 1.

Morena Danieli, Giuseppe Riccardi, and Firoj Alam. 2015. Emotion unfolding and affective scenes: A case study in spoken conversations. In *Proc. of Emotion Representations and Modelling for Companion Systems (ERM4CT) 2015,*. ICMI.

Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion*, 98:45–60.

Rafael A Calvo and Sidney D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *Affective Computing, IEEE Transactions on*, 1(1):18–37.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, pages 696–735.

Shammur Absar Chowdhury, Giuseppe Riccardi, and Firoj Alam. 2014. Unsupervised recognition and clustering of speech overlaps in spoken conversations. In *Proc. of Workshop on Speech, Language and Audio in Multimedia - SLAM2014*, pages 62–66.

Shammur Absar Chowdhury, Morena Danieli, and Giuseppe Riccardi. 2015a. Annotating and categorizing competition in overlap speech. In *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Shammur Absar Chowdhury, Morena Danieli, and Giuseppe Riccardi. 2015b. The role of speakers and context in classifying competition in overlapping speech. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Shammur Absar Chowdhury, Evgeny A. Stepanov, and Giuseppe Riccardi. 2016. Predicting user satisfaction from turn-taking in spoken conversations. In *Proc. of INTERSPEECH*.

Shrikanth Narayanan and Panayiotis G Georgiou. 2013. Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proceedings of the IEEE*, 101(5):1203–1233.

Štefan Beňuš, Agustín Gravano, and Julia Hirschberg. 2011. Pragmatic aspects of temporal accommodation in turn-taking. *Journal of Pragmatics*, 43(12):3001–3027.

Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58.