Automatic Detection of Conflict Escalation in Spoken Conversation

Samuel Kim[1]*, Sree Harsha Yella*[1,2] *and Fabio Valente*[1]

[1] IDIAP Research Institute, Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
{samuel.kim, sree.yella, fabio.valente}@idiap.ch

Abstract

This paper investigates the automatic recognition of conflict escalations during spontaneous conversations. In our previous work, we studied if the level of conflict in a segment of conversation can be automatically inferred by means of prosodic and conversational features. This work investigates the possibility of automatically recognizing if the conflict is increasing, i.e., escalating, or not. The dataset used for the study consists of political debates where short clips are classified into escalation, de-escalation and constant labels. Results show a Weighted Accuracy (WA) equals to 69.6% and an Unweighted Accuracy (UA) equals to 49.5% thus revealing lower accuracies compared to the simple conflict detection task (WA 86.1%, UA 71.1%). While the task appears more difficult compared to conflict detection, results are significantly better than chance level showing the feasibility of this approach. Furthermore, the paper investigates the use of a speaker diarization algorithm to extract features in a completely automatic fashion highlighting some limitations of diarization system.

Index Terms — Spoken Language Understanding, Conflicts, Paralinguistic, Spontaneous Conversation, Prosodic features, Turn-taking features

1. Introduction

Several works in automatic analysis and understanding of spoken conversations have investigated in recent years phenomena like social dominance [1], engagement and hot-spots [2] or agreement/disagreement [3]. However, most of the conversational data used in the literature represent collaborative, formal or non-conflictual scenario discussions like meetings or broadcast conversations.

In our previous work [4], we started to study the problem of detecting the levels of conflict in conversations. Conflicts are *mode of interaction* where *the attainment of the goal by one party precludes its attainment by the others* and are largely expressed by means of non-verbal cues such as interruptions, facial expressions, intensity and prosody, posture which became more or less frequent depending on how intense the conflict [5]. In [4], we showed that it is possible to detect the level of conflict in a conversation using statistical classifiers trained on conversational and prosodic features extracted from manual segmentation. The investigation was carried on a database of political debates [6] where spontaneous conflicts naturally arise between participants.

This paper continues the previous study focusing particularly on the detection of *conflict escalations* (and *de-escalation*). The phenomenon has been widely studied in social science and several models of conflict escalation have been proposed in literature (see [7, 8] for a review) and investigated both in spoken conversations and text conversations like email communication [9]. This work investigates whether *conflict escalation*, referred as an increase in the intensity of conflict during a conversation, can be detected by means of statistical classifiers trained on non-verbal features. As conflicts have negative effects on communication, detecting them before they arrive at the apex can have several applications, for example, machine-mediated communication. The clips from the debate database used in our previous work have been annotated with three levels of conflicts (high, medium, low), thus it is possible to address the problem comparing the levels of two consecutive clips in order to study cases of spontaneous escalation and de-escalation in conversations.

As the second contribution, the paper investigates the use of features extracted from an automatic segmentation system comparing results with manual segmentation used in [4]. Turn-taking patterns and overlaps between speakers provide rich information on the presence or the absence of conflicts in the conversation. However, the automatic estimation of speaker segmentation and especially regions of overlap speech between participants is prone to large errors [10]. Thus we study the use of a state-of-the-art speaker diarization method to extract this information.

The remainder of the paper is organized as follows. Section 2 briefly describes the database and its annotation, Section 3 describes the operative definition of conflict escalation and de-escalation, Section 4 describes the features and the speaker diarization system, Section 5 presents experimental results and finally the papers is concluded in Section 6.

2. Database

The database used in this study consists of broadcasted political debates in French language [6]. Each debate includes one moderator and two coalitions opposing one another on the issues of the day. The data are annotated into speaker turns, i.e., who spoke when, including overlapped region and a mappings between speakers and their roles, i.e., moderator or guest, is available. A subset of this database composed of 45 debates with four guests (two guests in each group) plus one moderator has been annotated in terms of conflicts.

The debates have been segmented into 30-second non-overlapping clips (3109 in total) assuming that the levels of conflict are stationary within the time period. The video clips have been shown to annotators who had to answer 15 questions concerning their perception of conflict. The questionnaire was designed to attribute scores in a conflict space, i.e. inferential layer and physical layer, for each clip. Details on the annotation process and the questionnaire can be found in [4]. Clips containing only monologues or interactions between a single guest and a moderator are not, at least in principle, conflictual. Thus, only 1496 clips (approximatively 12.5 hours) were selected for

Table 1: Number of instances in each class of levels of conflict

Low		Medium	High	Total
un-annotated	annotated			
2255		724	130	3109
1613	642			

annotation and the remaining clips (1613 clips, 13.5 hours) are considered as non-conflicted clips. A total of 10 annotations per clip were obtained and then converted in three levels of conflict: *high, medium*, and *low* by a majority voting method (see [4] for more details). The un-annotated monologue clips are labeled as low class. By doing so, we have labels for the whole dataset and the label distribution is reported in Table 1.

3. Labeling and Detecting Conflict Escalation/De-Escalation

Three possible situations can be considered in order to study the evolution of conflict in the conversation: *escalation*, *de-escalation*, and *constant*. Figure 1 illustrates a basic diagram of the proposed conflict escalation detection system. Let us consider two consecutive clips in a debate, c_t and c_{t+1}, labeled with their level of conflict (high, medium, low) l_t and l_{t+1}. As shown in the upper part of the figure, the escalation label at the current clip (e_t) is based on the level of conflict at the current clip (l_t) and the level of conflict at the following clip (l_{t+1}). We label as *escalation* the clips whose following clip contains higher levels of conflict and as *de-escalation* the clips whose following clip contains lower levels of conflict. We also assign *constant* for those clips whose following clip maintain the same level of conflict. Table 2 shows the number of instances how the levels of conflict change between adjacent clips (from labels in rows to labels in columns) in the dataset. Table 3 shows the distribution of those labels. Note that we do not assign any label to the last clip of each debate since there is no following clips.

Let us designate with f_t a set of features (conversational and prosodic) extracted form clip c_t. In order to infer whether there is conflict escalation, in a training phase, two different models are trained using features f_t; one is for inferring the levels of conflict at the current clip (l_t) and the other is for the levels of conflict at the following clip (l_{t+1}) assuming that the features extracted from the current clip should be related with the levels of conflict at the following clip. In a testing phase, as shown in the bottom part of the figure, we can obtain predicted levels of conflict both for the current clip ($\hat{l}_t$) and the following clip ($\hat{l}_{t+1}$) using two different models mentioned above. Finally, comparison between $\hat{l}_t$ and $\hat{l}_{t+1}$ will provide information on whether the conflict is escalating, de-escalating or staying constant.

Table 2: Transition matrix showing how the levels of conflict change between adjacent clips (from labels in rows to labels in columns).

	Low	Medium	High
Low	1831	367	25
Medium	355	287	70
High	31	63	35

Table 3: Number of instances in each class of conflict escalation.

De-escalation	Constant	Escalation	Total
449	2153	462	3064

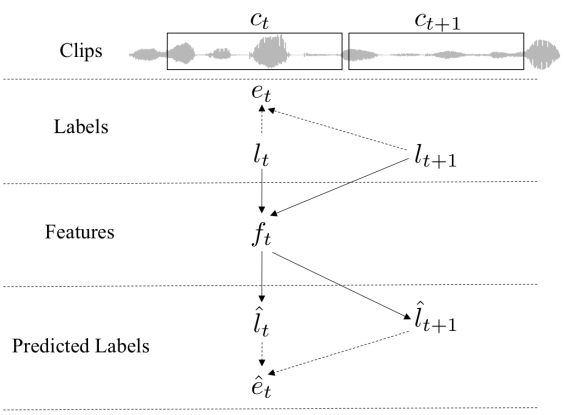

Figure 1: Diagram of labeling and detecting the conflict escalation.

Table 4: List of features: (a) prosodic and (b) conversational features.

(a) Prosodic features

Low-level features	Attributes	
	Clip-based	Speaker turn-based
Pitch	mean, median, std, max, min, {1, 25, 75, 99} percentile	-
	mean, median, std	mean, median, std, max, min, {1, 25, 75, 99} percentile
	only consider overlapped regions	mean, median, std, max, min, {1, 25, 75, 99} percentile
Intensity	mean, median, std, max, min, {1, 25, 75, 99} percentile	-
	mean, median, std	mean, median, std, max, min, {1, 25, 75, 99} percentile
	only consider overlapped regions	mean, median, std, max, min, {1, 25, 75, 99} percentile

(b) Conversational features

Turn-based feature	Attributes
duration of turns	mean, median, std, max, min
individual speaking time	mean, median, std, max, min
amount of overlap	between opposing group between moderator within a group total amount
turn taking pattern	between opposing group between moderator within a group total number of turns turn stealing ratio
number of participants	-

4. Feature Extraction

The features used are similar to those introduced in our previous work [4] and they consist in conversational and prosodic features extracted at speaker and clip level. Conversational features are used to capture the structure of conversations,. i.e., the way speakers organize in taking turns during the discussion. Table 4 provides the list of features that we use in this work (prosodic features include 90 features and conversational features include 20 features).

Pitch and intensity values are normalized on a speaker-level basis so that their distributions over the entire debate with respect to each participants has zero mean and unit standard de-

viation. In contrary to our previous work [4], prosodic features (pitch and intensity statistics) are also extracted from overlap regions.

In order to model the dynamics of the conflict, i.e., whether the conflict is escalating or de-escalating, temporal differences between features are introduced. At first, a given clip is segmented into two non-overlapping parts by dividing at the half point, i.e., two 15 second long segments. After that, the conversational and prosodic features are extracted from each half and their differences are calculated. This procedure is similar to estimating delta features in Automatic Speech Recognition, thus we refer them as Delta features hereafter. In contrast, the features extracted from the whole clip will be referred as Static features.

4.1. Automatic Speaker Segmentation

Extracting features described above, either conversational or prosodic, requires speaker segment information, i.e. who speaks when. In our previous work, we used manual segmentation for extracting various statistics. Towards a fully automated system, the use of automatic speaker diarization method [11] and overlap speech detection method [10] are investigated in the following.

The baseline diarization is based on information bottleneck clustering framework [11]. This clustering starts with an initial uniform segmentation of input audio file and then performs agglomerative clustering at each step by combining segments that are closest according to a distance measure. The output of the clustering assigns each segment to a unique cluster, i.e., a speaker. Since it assigns only one cluster for each segment, overlap information is lost in case of multiple speakers talking simultaneously. In order to overcome this issue, an HMM-based overlap detection method is used to detect overlap speech [10]. Detected overlap regions are then assigned to the two speakers who are closest in time based on the diarization output. Table 5 shows the performance of speaker diarization methods in terms of Diarization Error Rate (DER) and F-measure for detecting overlapped regions. As shown in the table, automatically detecting overlapped regions improve the performance of the diarization in terms of DER although the overlap detection algorithm itself has a low F-measure. Otherwise, the diarization errors are comparable to those obtained on other spontaneous conversation data as meeting recordings [11].

Table 5: Performance of speaker diarization in terms of diarization error rate (DER). Performance regarding overlap detection is given in terms of F-measure.

(%)	w/o overlap detection	w/ overlap detection
DER	11.4	10.6
F-measure	-	32.0

4.2. Correlations between features and labels

We compute the Pearson correlation coefficients between the feature values and the labels for both levels of conflict at the current clip and the following clip to identify the most informative ones. The correlation coefficients computed using the features from manual segmentation and automatic speaker diarization are denoted as ρ and $\bar{\rho}$, respectively.

The most correlated features are 1) the *minimum pitch during overlap* amongst the prosodic ones with correlation values of $\rho = -0.65$; $\bar{\rho} = -0.66$ for the current clip and $\rho = -0.35$; $\bar{\rho} = -0.38$ for the following clip and 2) the *total amount of overlap* in the clip amongst the conversation features with correlation values of $\rho = 0.77$; $\bar{\rho} = 0.67$ for the current clip and $\rho = \bar{\rho} = 0.37$ for the following clip. Note that the most correlated features are the same whether manual segmentation or automatic segmentation is used.

5. Experiments

Experiments are performed using a 5-fold cross validation to provide speaker and debate independent training/testing subsets. The entire dataset is split into 5 folds where 4 are used as training and the remaining is used for testing. The procedure is repeated until all the folds are used for testing. Note that we carefully design the folds so that they exclusively contain speakers and debates in a way the same speakers would not appear in both training and testing data. A simple debate-independent folds would not be speaker-independent since there are speakers who participated in multiple debates. Since it is required to have data for training the overlap detector on speaker diarization, we share the same folding information to train models for the overlap detection and extract the set of features according to the speaker diarization results.

The classification is based on a simple multi-class linear-kernel SVM. As the number of classes is not equally distributed, classification performances are reported in terms of Unweighted Accuracy (UA) as well as Weighted Accuracy (WA) which are commonly used in paralinguistic classification tasks [12]. For comparison, we provide chance level performance as well. The chance level performance is evaluated using randomly generated labels with the prior probabilities of individual classes learnt in a training fold.

In the first experiment, we perform classification tasks with respect to the levels of conflict at the current clip (using notation introduced in Section 3, this corresponds to estimating l_t) and report the results in Table 6. It can be observed that detecting the level of conflict at the current clip achieves a WA value of 86.1% and a UA value of 71.1% when manual segmentation is used. We can also observe that the use of dynamic features (delta features) degrades the performance. When manual segmentation is replaced with automatic segmentation, the performances degrade to 83.8% and 62.3% showing that imprecise boundaries from diarization system actually affect the feature extraction and consequently the conflict detection.

Table 7 shows the performance of classifying the levels of conflict at the following clip reporting WA/UA values of 74.2% and 37.7% (using the notation introduced in Section 3, this corresponds in estimating l_{t+1}). Although the performance is lower than classifying levels of conflict at the current clip, it still significantly outperforms the chance level ($p < 10^{-10}$)[1]. As before a degradation in performance is verified when manual segmentation is replaced with automatic segmentation. However it is interesting to notice that the use of delta features improves the performance to 41.5% in case of UA measure indicating that changes in conversational and prosodic features within the clip carry information on the levels of conflict at the following clip.

The detection of conflict escalation or de-escalation can then be obtained comparing results obtained by the two previously described classifiers. The difference between $\hat{l}_{t+1}$ and $\hat{l}_t$ provides information on the fact the conflict is increasing, decreasing or staying constant. Results are reported in Table 8.

[1]The McNemar's test is used to show significancy.

Table 6: Performance of classifying levels of conflict at the current clip.

		WA (%)	UA (%)
Manual Segmentation	Static	86.1	71.1
	Static + Delta	84.4	67.6
Speaker Diarization	Static	83.8	62.3
	Static + Delta	82.8	61.4
Chance Level		59.1	34.0

Table 7: Performance of classifying levels of conflict at the following segment.

		WA (%)	UA (%)
Manual Segmentation	Static	74.2	37.7
	Static + Delta	73.7	41.5
Speaker Diarization	Static	73.3	35.8
	Static + Delta	73.6	38.7
Chance Level		58.4	33.3

Table 8: Performance of classifying conflict escalation by comparing two classification results, i.e., the levels of conflict at the current clip and the following clip.

		WA (%)	UA (%)
Manual Segmentation	Static	69.6	49.5
	Static + Delta	68.9	47.8
Speaker Diarization	Static	69.0	47.9
	Static + Delta	68.4	45.5
Chance Level		54.8	34.1

The classification achieves WA/UA values of 69.6% and 49.5% whenever static features are used. As before, a degradation in performance is observed if the output of a speaker diarization systems replaces the manual segmentation.

Comparing results from Table 6, Table 7 and Table 8, it is possible to notice that the problem of detecting escalation/de-escalation is more difficult than detecting the intensity of the conflict. Comparing results with chance levels, however, it is notable that performance is still significantly better than chance level ($p < 10^{-10}$) thus showing the feasibility of this approach to detect escalation/de-escalation.

6. Conclusion

This paper continues the study of detecting conflicts during a spontaneous spoken conversation and particularly focus on conflicts escalations. The study is performed on the Canal 9 database of political debates where conflicts between participants naturally arise during the conversation. In our previous related work, we showed that it is possible to detect the levels of conflict in a short segment of conversation using prosodic and conversational features. This paper extends the previous study investigating if it is possible to detect when the conflict is escalating, i.e., when the level of conflict is increasing. Predicting conflict escalation can have important applications into analysis as well as machine mediated communication.

The database has been annotated in terms level of conflicts (low, medium, high) thus allowing the investigation of cases in which the level is increasing, decreasing or staying constant which are referred as escalation, de-escalation and constant. Correlation studies between labels and the speech-related (prosodic and conversational) features revealed that the most correlated features are common across the labels and segmentation methods. Specifically, *minimum pitch during overlap* amongst prosodic features and *total amount of overlap* amongst conversational features are the highest correlated features.

Classification experiments based on SVM classifier reveal that it is possible to detect the levels of conflict with an Unweighted Accuracy equals to 71.1% and a Weighted Accuracy equals to 86.1%. On the other hand, conflict escalations/de-escalations can be detected with an Unweighted Accuracy equals to 49.5% and a Weighted Accuracy equals to 69.6% showing that this task is more complex than simply detecting the intensity. We also show that dynamic features (delta features) are informative in predicting the conflict level of a following clip.

Furthermore the use of an automatic speaker diarization algorithm is investigated. Errors measure in terms of DER and F-measure overlap detection show state-of-the-art performances (DER 10.6%; overlap F-measure 0.32) thus comparable to numbers obtained on meeting [10, 11]. However, when the manual segmentation is replaced with the output of an automatic speaker diarization system for detecting speaker boundaries and overlap regions, the overall performances degrade in both tasks suggesting that further improvements need to be obtained on the diarization system.

In all the cases, the results obtained by the detection systems are statistically significant outperforming chance levels showing the feasibility of automatic escalation detection.

7. Acknowledgement

This work was funded by the EU NoE SSPNet, SNF-RODI and SNF-IM2.

8. References

[1] D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations from non-verbal activity cues," *IEEE Transactions on Audio, Speech and Language Processing*, Mar 2009.

[2] D. Wrede and E. Shriberg, "Spotting "hotspots" in meetings: Human judgments and prosodic cues," in *Proceedings of Eurospeech*, 2003.

[3] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of agreement vs. disagreement in meetings: training with unlabeled data," in *Proceeding NAACL*, 2003.

[4] S. Kim, F. Valente, and A. Vinciarelli, "Automatic detection of conflicts in spoken conversations: ratings and analysis of broadcast political debates," in *IEEE International Conference on Acoustic, Speech, and Signal Processing (ICASSP)*, Mar. 2012.

[5] V. Cooper, "Participant and observer attribution of affect in interpersonal conflict: an examination of noncontent verbal behavior," *Journal of Nonverbal Behavior*, vol. 10, no. 2, pp. 134–144, 1986.

[6] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, "Canal9: A database of political debates for analysis of social interactions," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, September 2009, pp. 1–4.

[7] K. W. Thomas, "Conflict and conflict management: Reflections and update," *Journal of Organizational Behavior*, vol. 13, 1992.

[8] J. Allwood, "Cooperation, Competition, Conflict and Communication," *Gothenburg Papers in Theoretical Linguistics*, vol. 94, pp. 1–14, 2007.

[9] R. A. Friedman and S. C. Currall, "Conflict escalation: Dispute exacerabiting elements of email communication," *Human Relations*, 2003.

[10] K. Boakye, O. Vinyals, and G. Friedland, "Improved overlapped speech handling for speaker diarization," in *Proceedings of Interspeech*, 2011.

[11] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Transactions on Audio Speech and Language Processing*, vol. 17, no. 7, 9 2009.

[12] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The interspeech 2011 speaker state challenge," in *Proceedings of Interspeech*, 2011.