AUTOMATIC DETECTION OF CONFLICTS IN SPOKEN CONVERSATIONS: RATINGS AND ANALYSIS OF BROADCAST POLITICAL DEBATES

Samuel Kim[1], *Fabio Valente*[1], *Alessandro Vinciarelli*[2]

[1]Idiap Research Institute, CH-1920 Martigny, Switzerland
[2]University of Glasgow, G12 8QQ Glasgow, United Kingdom
{samuel.kim,fabio.valente}@idiap.ch, vincia@dcs.gla.ac.uk

ABSTRACT

Automatic analysis of spoken conversations has recently searched for phenomena like agreement/disagreement in collaborative and non-conflictual discussions (e.g., meetings). This work adds a novel dimension investigating conflicts in spontaneous conversations. The study makes use of broadcasted political debates where conflicts naturally arise between participants. In the first part, an annotation scheme to rate the degree of conflict in conversations is described and applied to 12 hours of recordings. In the second part, the correlation between various prosodic/conversational features and the degree of conflict is investigated. In the third part, we perform automatic detection of the level of conflict based on those features showing an F-measure of 71.6% in three-level classification tasks.

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

1. INTRODUCTION

Automatic analysis and understanding of spoken conversations has been an active research field over the last years with a number of applications including indexing, retrieval and summarization. One of the most common type of data studied in this field is the meeting scenario, i.e., a small group's face-to-face conversations either spontaneous or scripted [1]. Phenomena studied in this setting consist of social dominance [2], engagement and hot-spots [3], behavioral codes (e.g., acceptance and blame) [4] as well as agreement/disagreement [5]. Besides meetings, broadcasted conversations, e.g., talk shows, also have been automatically analyzed searching how phenomena like agreement/disagreement are expressed in different languages [6, 7]. Statistical models used in these studies are trained on various lexical, prosodic and conversational features [3, 5, 4, 6, 7]. However, most of the conversational data used in the literatures represent collaborative, formal and non-conflictual scenario discussions.

This work adds a novel dimension to automatic analysis of human conversations by studying how *conflicts* can be automatically modeled and detected. Besides analysis, indexing and summarization, the detection of conflicts can find various applications, e.g., machine-mediated communication. Conflicts can be considered as particular hot-spots in a conversation or an extreme form of disagreement. Disagreement in conversations is a difference of opinion and typically expressed by means of verbal and non-verbal communication. Consequently, previous works have addressed the recognition of disagreement using features like words and dialog-acts together with other features like prosody or turn-taking statistics [3, 5, 4, 6, 7]. On the other hand, a conflict can be also considered as a collision of interests or intentions among people or groups where people are competing to win or trying to force an "adversary" into submission. Conflicts in conversations are largely expressed by means of non-verbal messages (interruptions, facial expressions, intensity and prosody, posture) which become more or less frequent depending on how intense the conflict is. For instance, overlapping speech intervals tend to become longer and more frequent, and more speakers attempt to interrupt each other more frequently during conflicts [8].

Spoken conversation corpora like the AMI corpus or the ICSI corpus contain disagreement instances which do not necessarily lead to conflicts between participants. In order to study conflicts in spontaneous conversations in a fairly-controlled scenario, this work makes use of a database of political debates [9] broadcasted in between 2005 and 2008[1]. In contrast with other benchmarks, political debates are real-world competitive multi-party conversations where participants do not act in a simulated context, but participate in an event that has a major impact on their real life (for example, in terms of results at the elections). Thus, even if the debate format imposes some constraints, the participants are moved by real motivations leading to highly spontaneous conflicts. As no standard coding scheme for rating conflicts is available in literature, Section 2 describes and motivates the rating protocol used to attribute a conflict score to a short conversation clip and to assign a level of conflict (Low, Medium or High) to them. The remainder of the work (section 3) then studies correlations between those scores and easily extractable features from the speech signal (f0 statistics, intensity, speech rate) and from the speaker segmentation (average speaker time, overlap, interruptions). Section 4 then studies how conflicts in conversations can be automatically detected by statistical classifiers trained on the speech and conversation features. The paper is concluded in Section 5.

2. DATA DESCRIPTION AND ANNOTATION PROTOCOL

The database used in this study consists of broadcasted political debates in French language [9]. Each debate includes one moderator and two coalitions opposing one another on the issues of the day. The debates are manually segmented into speaker turns. Furthermore the mappings between speakers and their roles (moderator, guest-group 1, guest-group 2) are available. Let us consider a subset of this database composed of 45 debates with four guests (two in each group) plus a moderator. The statistics related to this subset are summarized in Table 1. The recordings have been segmented into 30-second long uniform, non-overlapping clips. The clips where at least two guests speak are retained as potential conflict samples. The reason behind the choice is that monologues or interactions involv-

[1]The databased is downloadable at http://canal9-db.sspnet.eu/

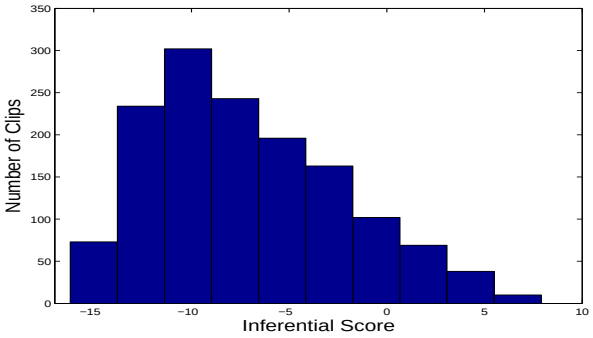

Fig. 1. Histogram of scores obtained using inferential questions only. Most of the clips have negative scores indicating on average a low level of conflict in the dataset.

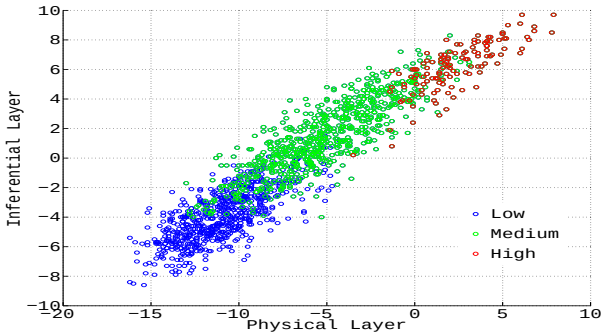

Fig. 2. Scatter of inferential and physical scores indicating that the perception of a conflict is higly correlated with the perception of certain behavior in the conversation.

ing the moderator are not, at least in principle, conflictual. Thus, only 1430 clips (approximatively 12 hours) are retained for annotation/ratings. These clips are annotated by 10 persons who answer a questionnaire (see Table 2) after watching individual 30 second clips. The questionnaire aims at attributing a conflict score to each clip. It includes 15 statements accounting for two *layers*: an *inferential layer* which includes questions about the interpretation of the group discussion and a *physical layer* that includes questions about the behavior being observed. The rationale used to generate the inferential statements is based on [10] which defines a conflict as a "*mode of interaction*" where "*the attainment of the goal by one party precludes its attainment by the others*". The statements are motivated by the perception of the "*competitive processes*" [10] typically resulting from conflicting goals, possibly leading to attempts of limiting, if not eliminating, the speaking opportunities of others in conversations. The questions of the physical layer are motivated by the literature on conflicts showing what are the most frequent behaviors observed (see, e.g., [8]) like interruptions, overlapping speech and other cues related to turn-organization in conversations but also head nodding, fidgeting and frowning. The questionnaire is multiple choice and each rater must select one answer out of five possible alternatives: agreement with the statement for inferential layer (Strongly Agree, Agree, Nor Agree neither Disagree, Disagree, Strongly Disagree) and frequency of a given behavior for physical layer (Never, Once or Twice, Sometimes, Often, Always). A numerical value in [-2,-1,0,1,2] is then assigned to each of the five levels thus converting answers into a numerical score which is summed up across inferential/physical questions and averaged across the 10 raters. The raters were not aware of the layers and questions belonging to the two types were mixed in the questionnaire. In summary, ratings provide two scores: an inferential one related to how raters perceive conflict and a physical one related to what behavior annotators notice in the clips. Fig. 1 plots the distribution of the inferential scores showing that most of them have negative values indicating on average a low degree of conflict. Fig. 2 plots the physical score versus the inferential score for the 1430 clips, indicating that the inferential score is highly correlated with what observed in the physical layer. In other words, the perceived level of conflict is highly correlated with the perceived

Table 1. Statistics of the corpus subset used in this study.

Number of recordings	45
Number of speakers per debates	5
Average debate length	40 minutes
Amount of overlap speech	8%
Average turn duration	9 seconds

Table 2. Questionnaire provided to the annotators divided according to an inferential and a physical layer. Before score computing, answers to questions 1,2 and 8 take a sign change.

#	Question	Layer
1	The atmosphere is relaxed (-)	Inferential
2	People wait for their turn before speaking (-)	Physical
3	One or more people talk fast	Physical
4	One or more people fidget	Physical
5	People argue	Inferential
6	One or more people raise their voice	Physical
7	One or more people shake their heads and nod	Physical
8	People show mutual respect (-)	Inferential
9	People interrupt one another	Physical
10	One or more people gesture with their hands	Physical
11	One or more people are aggressive	Inferential
12	The ambience is tense	Inferential
13	One or more people compete to talk	Physical
14	People are actively engaged	Inferential
15	One or more people frown	Physical

Table 3. Distribution of clips according to the three level of conflict.

	Low	Medium	High	Total
Number of clips	611	694	125	1430

frequency of certain phenomena in the conversation (overlaps, interruptions, fidgeting and so on). The Pearson correlation coefficient of the two scores is 0.94. Section 3 will investigate how automatically or semi-automatically extracted features from the speech signal or the speaker segmentation, correlates with the perception of conflict.

After that, the paper investigates how conflicts in conversation can be automatically detected by means of statistical classifiers like SVM trained on those features. The continuous scores are quantized into three classes based on majority voting over statements and raters; let us refer to them as *level of conflict* (High, Medium, Low). The high and low classes are those where majority voting on physical and inferential layers agree in attributing a positive or a negative value to the clip. The remaining clips are assigned to the Medium class. The distribution of those clips is reported in Table 3 as well as depicted in Fig. 2 (blue, green and red dots). It can be noticed that, even if the data represents competitive conversations (debates) and the clips that contain only monologues have been removed, only 8% of the data are labeled as containing high degree of conflict which show that those are rare phenomena in fairly controlled conversations.

3. CONVERSATIONAL AND PROSODIC FEATURES

Agreement/disagreement can be automatically detected from features related to the structure of the conversation,. i.e., the way speakers organize in taking turns during the discussion [11, 6]. Similarly, let us study their correlations with the conflict clips rating. Correlations are computed w.r.t. the inferential layer scores. The set of features considered consists in :

1) the **turn duration statistics** which include mean, median, maximum, variance and minimum of speaker turns duration in the clip as well as the **average number of turns**.

2) the **turn-taking pattern** between speakers. Knowing that each participant in the discussion is either moderator, guest of group 1, or guest of group 2, i.e., $r = (m, g1, g2)$, the way participants take turn in the conversation could reveal patterns of conflict. For instance, during a conflict, we could expect several guests taking the floor of the discussion alternatively. This information can be modeled with a simple bi-gram counts, i.e., $p(r_t, r_{t-1})$ where r_t is the role of the speaker that holds the floor of the conversation at turn t.

3) the **amount of overlap** relative to the clip duration; we distinguish three types of overlaps based on the role that each speaker has in the debate, i.e., overlap between moderator and guests OV_{MG}, overlap between guests belonging to the same group OV_G and the overlap between guests belonging to opposite groups OV_{G12}. Overlaps between more than two participants are not studied.

4) the **turn keeping/turn stealing ratio** in the clip, defined according to the notation defined in [12], as the ratio between the number of times a speaker change happens and the number of times a speaker change does not happen after an overlap. The rationale behind this consists in capturing aggressive interruptions aiming at grabbing the floor of the conversation.

Prosodic features have been shown to correlate with a number of phenomena including the speaker level of engagement in the conversation [3, 4]. Speech rate, articulation rate, and their statistics (mean, median, standard deviation) are computed using pseudo-syllables over the clip and over the turns without any normalization. Pitch and intensity are also estimated using the Praat Toolkit (http://www.praat.org/) every 10 milliseconds and two types of statistics are extracted: one from the entire clip (30 second) and one for each speaker turn in the clip. The first models the entire conversation while the latter models the prosodic behavior of a particular speaker.

1) **Clip-based statistics**: they represent the mean, median, standard deviation, maximum, minimum and quantiles (0.01, 0.25, 0.75 and 0.99) of pitch and intensity obtained from the entire clip. Before computing those, frame-level prosodic features are speaker based normalized applying a *Z-norm* ($\bar{x} = (x - m_s)/\sigma_s$ where m_s and σ_s are speaker statistics obtained on the entire debate).

2) **Speaker turn-based statistics**: they represent the mean, median and standard deviation of pitch and intensity obtained over individual speaker turns (similarly to the clip-base statistics). Before computing those, frame-level prosodic features are globally normalized applying a *Z-norm* (statistics are obtained on the entire debate). This second set of features is anticipated to capture partial dynamics between participants better then the first one as they model prosodic behavior of each individual speaker averaging the estimates over its turns.

Let us now consider the Pearson correlation coefficients between those feature values and the inferential conflict scores. Amongst conversational features, OV_{G12}, i.e., the amount of overlap between speakers belonging to opposing groups is the highest correlated feature ($\rho = 0.63$), consistently with what observed in [8]. The feature with the second highest correlation ($\rho = 0.37$) is the bigram count $(g1, g2)$, i.e., the number of times participants from different groups take turn one after each other and the third highest correlated feature ($\rho = 0.36$) is the turn stealing ratio. Other features correlation are below 0.3. In summary, the features that correlate better with the inferential scores are the ones that indicate competition for the floor.

Amongst prosodic features, the speech articulation rate computed over the entire clip is the highest correlated feature ($\rho = 0.51$). After that, mean, median and 0.25 quantile of intensity computed over the speaker turns show correlations between 0.33 and 0.41 as intuitively explained by the fact participants speak louder during a conflict. On the other hand, features with high negative correlations are those related with statistics of minimum pitch estimated and normalized over speaker turns whose values are in between -0.35 and -0.37 suggesting that low pitch values can be associated with non-conflictual conversations. Other features have absolute correlation values below 0.3.

Table 4. Per-class precision, recall, and f-measure when the SVM classifier is trained on conversational, prosodic features or both of them in case of two class (high/low) problem.

Measures	Features	Low	High
Precision (%)	Conversational	95.7	91.3
	Prosodic	96.7	83.7
	All	97.5	86.4
Recall (%)	Conversational	98.5	78.2
	Prosodic	96.6	84.1
	All	97.2	87.8
F-measure (%)	Conversational	97.0	84.2
	Prosodic	96.7	83.9
	All	97.3	87.1

4. DETECTION OF CONFLICTS

Previous sections have shown how conversational and prosodic features correlate with conflict scores attributed by human raters. Let us now study how those features can be used to automatically classify conflicts into three levels [High, Medium, Low] quantized as described in Section 2. Two types of experiments with increasing difficulty are performed; the first aims at distinguishing between High/Low conflicts only (a two-class problem discarding instances that belong to medium level), while the second aims at distinguishing between the three levels, i.e., High/Medium/Low. The entire dataset (1430 clips) is randomly split into 10 folds where 9 are used as training and the remaining is used for testing. The procedure is repeated until all the folds are used for testing. The classifier is a simple multi-class linear-kernel SVM. As the three classes are not equally distributed, classification performances are reported in terms of precision, recall and F-measure.

Table 4 reports the performances of an SVM trained on conversational, prosodic features and both of them in classifying high/low conflict levels. In general, conversational features have higher precision while prosodic features have higher recall in detecting high conflicts. Their combination improves the F-measure up to 87.1% thus appearing complementary to each other.

Table 5 reports the performances of a three-way SVM trained on conversational, prosodic features and both of them. On average, prosodic features outperform the conversational ones (F-measure 68.1% versus 63.8%). Their combination provides an F-measure of 71.6% thus they appear complementary. Fig. 3 plots (a) precision, (b) recall and (c) F-measure per each of the three classes when conversational and prosodic features are used. It can be noticed that con-

Table 5. Average precision, average recall, and average f-measure when the SVM classifier is trained on conversational, prosodic features or both of them in case of three level of conflict.

Features	Precision (%)	Recall (%)	F-measure (%)
Conversational	66.4	66.8	63.8
Prosodic	68.4	68.7	68.1
All	71.7	71.9	71.6

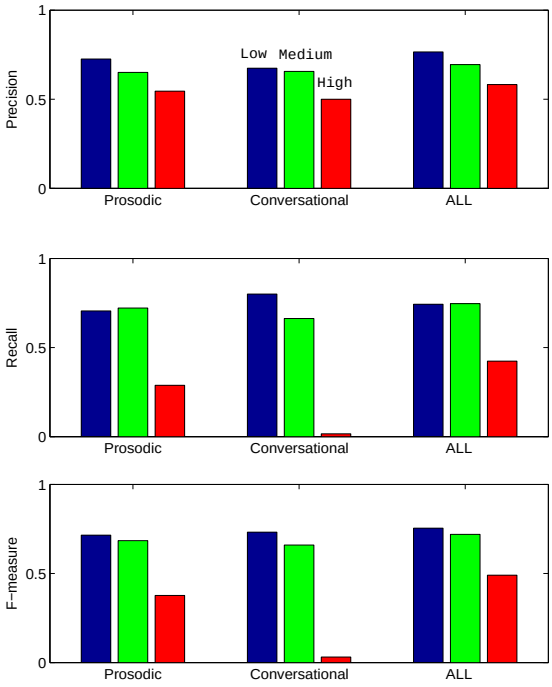

Fig. 3. Per Class precision, recall, and f-measure whenever the SVM classifier is trained on conversational, prosodic features or both of them in case of three level of conflict (Blue bars represent low level of conflict, Green bars represent medium level of conflict while red bars represent high level of conflict).

versational features have very low recall (thus very low F-measure) for high conflict levels and confusion matrix reveals that most of the the high conflict clips are assigned to the medium class. Reversely, prosodic features hold their performances also in the three class cases. This suggests that overlaps and interruptions can model part of conversations where conflict exists although perceived level or degree of conflict is rather modeled by the prosodic behavior of speakers (intensity and speech rate).

5. CONCLUSION

Most of the recent works on multi-party spoken conversations has focused on collaborative and non-conflictual scenarios like meetings [3, 5]. Among those studies, automatic detection of agreement/disagreement has been investigated extensively, including broadcast conversation scenarios like talk shows [6, 7]. This paper studies an extreme case of disagreement in a conversation which is represented by conflicts. As meeting corpora do not contain conflict instances, the study is carried on political debates where conflicts appear with significantly higher frequency than in cooperative scenarios investigated so far (e.g., meetings). The final purpose is automatic conflict detection using statistical classifiers trained on various speech and conversational features.

A coding scheme to assign a degree of conflict to a short conversation excerpt is introduced and 12 hours of data have been annotated according to this scheme. Correlation studies between those scores and a set of easily extractable features from the speech signal (f0 statistics, intensity, speech rate) and from the conversation (turns statistics, amounts of overlap, interruptions) revealed that the most correlated features with the level of conflict in the clips are: the amount overlap between participants (0.63) and their turn taking patterns (0.37) as well as the mean intensity (0.41) and speech rate (0.51) of the speakers involved in the conversation.

When those features are used to train SVM classifiers to automatically detect high/low degrees of conflicts (two class problem), conversational features produce higher precision while prosodic features produce higher recall and they appear complementary as their combination improves over the individual feature sets. As the problem is extended into three classes (High/Medium/Low), it is interesting to notice that conversational features cannot distinguish between Medium and High conflict levels suggesting that they can detect conversations where conflicts exists although the perceived degree of conflict is better detected by the prosodic features. Still they appear complementary to each other and they can achieve a F-measure equal to 71.6% in the tree-classes problem.

Interestingly, those results were obtained through the use of some very simple and easily extractable features discarding several sources of information raters were provided with, e.g., the video information and the verbal content. In future works, we plan to include approaches that utilize other sources of informations, e.g., lexical features.[2]

6. REFERENCES

[1] McCowan I., Gatica-Perez D., Bengio S., Lathoud G., Barnard M., and Zhang D., "Automatic analysis of multimodal group actions in meetings," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004.

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

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

[4] M. Black, A. Katsamanis, C.-C. Lee, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Automatic classification of married couples' behavior using audio features," in *In Proceedings of InterSpeech*, 2010.

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

[6] Wang W., Yaman S., Precoda P., and Richey C., "Automatic identification of speaker role and agreement/disagreement in broadcast conversation.," in *Proceedings of ICASSP*, 2011.

[7] Wang W., Precoda K., Richey C., and Raymond G., "Identifying agreement/disagreement in conversational speech: A cross-lingual study," in *Proceedings of Interspeech*, 2011.

[8] V.W. 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.

[9] Vinciarelli A. et al., "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.

[10] C.M. Judd, "Cognitive Effects of Attitude Conflict Resolution," *Journal of Conflict Resolution*, vol. 22, no. 3, pp. 483–498, 1978.

[11] Galley M., McKeown K., Hirschberg J., and Shriberg E., "Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies.," in *Proc. 42nd Meeting of the ACL*, 2004.

[12] Adda-Decker M. and et., "Annotation and analysis of overlapping speech in political interviews," in *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2008.

[2]This work was funded by the EU NoE SSPNet and Swiss National Fundation NCCR IM2.