ANNOTATION AND DETECTION OF CONFLICT ESCALATION IN POLITICAL DEBATES

Samuel Kim[1,2]*, Fabio Valente*[2] *and Alessandro Vinciarelli* [2,3]

[1] DSP Lab., Yonsei University, Seoul, Korea
[2] Idiap Research Institute, Martigny, Switzerland
[3] University of Glasgow, Glasgow, United Kingdom
samuel.kim@dsp.yonsei.ac.kr

ABSTRACT

Conflict escalation in multi-party conversations refers to an increase in the intensity of conflict during conversations. Here we study annotation and detection of conflict escalation in broadcast political debates towards a machine-mediated conflict management system. In this regard, we label conflict escalation using crowd-sourced annotations and predict it with automatically extracted conversational and prosodic features. In particular, to annotate the conflict escalation we deploy two different strategies, i.e., indirect inference and direct assessment; the direct assessment method refers to a way that annotators watch and compare two consecutive clips during the annotation process, while the indirect inference method indicates that each clip is independently annotated with respect to the level of conflict then the level conflict escalation is inferred by comparing annotations of two consecutive clips. Empirical results with 792 pairs of consecutive clips in classifying three types of conflict escalation, i.e., *escalation*, *de-escalation*, and *constant*, show that labels from direct assessment yield higher classification performance (45.3% unweighted accuracy (UA)) than the one from indirect inference (39.7% UA), although the annotations from both methods are highly correlated ($\rho = 0.74$ in continuous values and 63% agreement in ternary classes).

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

1. INTRODUCTION

A conflict in conversations can be defined as *an interaction that occurs between individuals when salient values or self-interests are threatened or challenged* and it is largely expressed by means of non-verbal cues such as interruptions [1]. Considering the conflicts in a conversation as particular hot-spots [2], automatic analysis of conflicts using non-verbal cues can find various applications in multimedia processing domain, such as indexing and summarization, just as other social phenomena such as dominance [3], agreement/disagreement [4], and acceptance and blame [5].

In our previous work [6], we formularized the problem of automatic detection of the levels of conflict in conversations. There 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 (it is also appeared as one of sub-challenges in INTERSPEECH 2013 Computational Paralinguistics Challenge [7]). In [8], we continued the study particularly focusing on *conflict escalation*, i.e., an increase in the intensity of conflict during a conversation, and investigated whether the conflict escalation can be detected by means of statistical classifiers trained on automatically extracted non-verbal features.

Since conflicts have negative effects on communication and detecting whether they increase or decrease may have several applications, e.g., machine-mediated human communication systems, in this work, we extend our approach to further investigate automatic detection of conflict escalation. In particular, we focus on annotation process to collect reliable labels on this subjective matter using crowd-sourced annotations. In our previous work, assigning labels with respect to conflict escalation is somewhat heuristic; clips from the debate database have been individually annotated and quantized, then the levels of two consecutive clips are compared in order to label conflict escalation.

In this work, we conduct a comparative study of the two different methods in annotating conflict escalation: *indirect inference* and *direct assessment*. In the indirect inference method, each clip is independently annotated with respect to the level of conflict then the level of conflict escalation is inferred by comparing annotations of two consecutive clips. This is similar to our previous work [8] but different in the sense that the levels of conflict remain as continuous values rather than quantized into classes. On the other hand, the direct assessment method indicates that annotators directly watch and compare two consecutive clips during the annotation processes. We hypothesize that the direct assessment method appropriately annotates the subjects' perception of conflict escalation while the indirect inference method may approximate the perception by comparing different subjects' perception of the level of conflict. To validate the hypothesis, we perform classification tasks using automatically extracted non-verbal features, i.e., conversational and prosodic features [6].

The remainder of the paper is organized as follows. Section 2 describes the database and two different annotation methodologies. In Section 3, we describe the feature extraction procedure followed by the classification tasks and their results. Finally the papers is concluded in Section 4.

2. ANNOTATION OF CONFLICT ESCALATION

2.1. Database

We use Canal9 broadcast political debates in French language. Each debate includes one moderator and two coalitions opposing one another on the issues of the day and we use a subset of the database, i.e., 45 debates, composed with four guests (two guests in each group) plus one moderator (see [9] for more details). The chosen debates have been segmented into 30-second non-overlapping clips assuming that the levels of conflict are stationary within the time period.

Table 1. Questionnaire provided to the annotators. Questions with (-) are designated to be inversely proportional to the other questions.

The atmosphere is relaxed	(-)
People argue	
People show mutual respect	(-)
One or more people are aggressive	
The ambience is tense	
People are actively engaged	

Assuming that clips containing only monologues or interactions between a single guest and a moderator are not conflictual, only 1496 clips (approximatively 12.5 hours) were selected for individual conflict annotations. To study conflict escalation, furthermore, we only consider the clips that are consecutively selected for individual conflict annotations. Thus, we deal with 792 clips (approximately 6.5 hours) in this work.

2.2. Questionnaire-based crowd-sourced annotations

We use a crowd-sourcing strategy to annotate the whole dataset. Specifically, we use the Amazon Mechanical Turk service[1] to easily manage a crowd for the annotation process. We have prepared a questionnaire that consists of 15 questions which reflect different aspects 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 [6]. In particular, in this work, we consider only the questions in the inferential layer listed in Table 1.

To assess the level of conflict escalation, we use two different strategies: indirect inference and direct assessment as illustrated in Fig. 1(a) and (b) respectively. The indirect inference method represents that each clip is individually annotated then the level of conflict escalation is inferred by comparing annotations of two consecutive clips. During the individual annotation processes, the annotators are asked to select one answer out of five possible alternatives in an ordinal scale [*strongly disagree, disagree, neither agree nor agree, agree*, and *strongly agree*]. 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 averaged across the questionnaire and the annotators, i.e.,

$$l_t = \frac{1}{Q}\sum_{q=1}^{Q} \frac{\sum_{j=1}^{R}\sum_{k=1}^{K} v_q(k)\delta(y_t^{j,q},k)}{\sum_{j=1}^{R}\sum_{k=1}^{K}\delta(y_t^{j,q},k)} , \tag{1}$$

where Q, R and K represent the number of questions, the number of annotators and the number of possible answers, respectively; $\delta(y_t^{j,q},k)$ represents a delta function, i.e.,

$$\delta(y_t^{j,q},k) = \begin{cases} 1 & \text{if } y_t^{j,q} = k; \\ 0 & \text{otherwise} , \end{cases}$$

and $y_t^{j,q}$ and $v_q(k)$ denote an index of chosen answer for question q by annotator j considering t-th video clip and the assigned value of k-th answer for question q, respectively. Note that the questions with (-) in Table 1 are designed to be inversely proportional to the other questions. Consequently, the values are assigned reversely for those questions, i.e.,

$$v_q(k) = \begin{cases} -(k-\zeta) & \text{if } q \in \{1,3\}; \\ k-\zeta & \text{otherwise} , \end{cases} \tag{2}$$

[1]https://www.mturk.com/

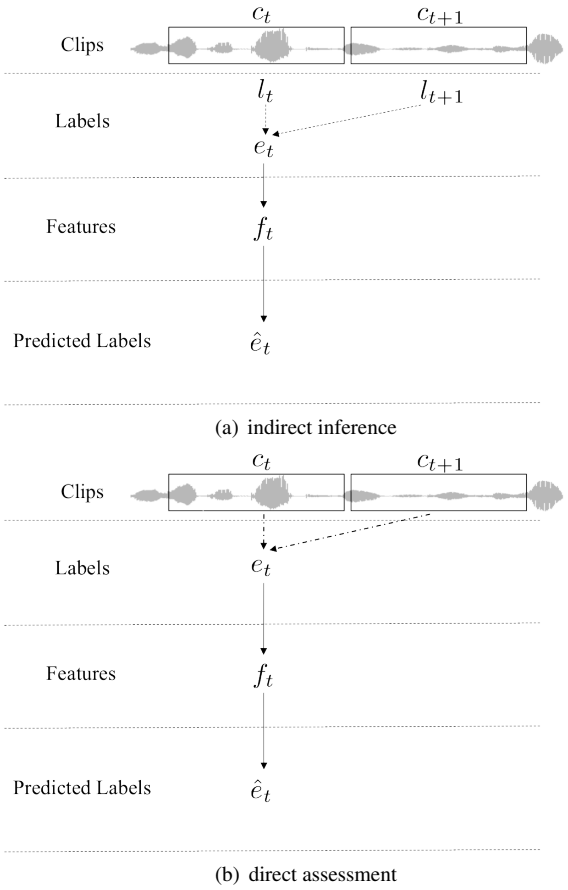


Fig. 1. Diagram of labeling and detecting the conflict escalation (a) indirect inference and (b) direct assessment.

where $k \in \{1, \cdots, K\}$. K and ζ are set 5 and 3 respectively in this work.

Inferring the level of conflict escalation can be done by subtracting the levels of conflict of two consecutive clips, i.e.,

$$e_t = l_{t+1} - l_t \ . \tag{3}$$

On the other hand, the direct assessment method indicates that during individual annotation processes the annotators watch and compare two consecutive clips, namely A and B, and are asked to select one answer out of five possible alternatives in an ordinal scale (A>>B, A>B, A=B, A<B, and A<<B, where inequalities represent comparative senses). Two consecutive video clips are arranged side-by-side (as illustrated in Fig. 2) and the second video clip can be played only after the first video clip is finished. Like the indirect inference method, a numerical value in [-2,-1,0,1,2] is assigned to each of the five levels, as in Eq. 2, then directly convert answers into the level of conflict escalation, i.e.,

$$e_t = \frac{1}{Q}\sum_{q=1}^{Q} \frac{\sum_{j=1}^{R}\sum_{k=1}^{K} v_q(k)\delta(y_t^{j,q},k)}{\sum_{j=1}^{R}\sum_{k=1}^{K}\delta(y_t^{j,q},k)} . \tag{4}$$

The primary difference between these two strategies is whether the annotators are able to observe two consecutive clips to assess differences between these clips. Furthermore, two consecutive clips are

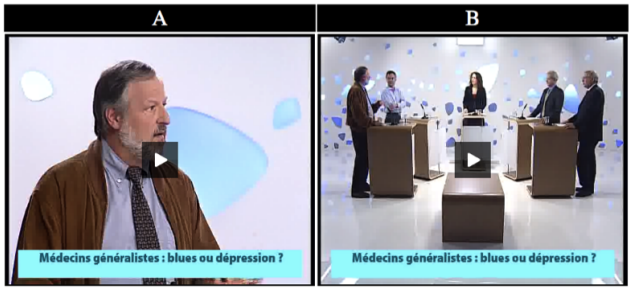


Fig. 2. An example of user interface for direct assessment. Two consecutive video clips are arranged side-by-side and the second video clip can be played only after the first video clip is finished.

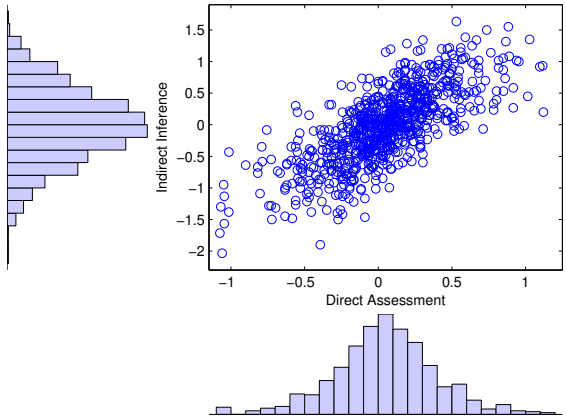


Fig. 3. Scattered plot and histograms of the conflict escalation levels through direct assessment method and indirect inference method.

Table 2. Statistics of collected crowd-sourced annotations for indirect inference and direct assessment.

	Indirect inference	Direct assessment
Number of clips	792	
Number of annotators	615	279
Annotators per clip	10	11
Clips per annotator	14	31

rarely annotated by the same annotators. Table 2 shows the statistics of collected crowd-sourced annotations through two different methods and Fig. 3 illustrates the scattered plot and histograms of the conflict escalation levels consolidated from the collected crowd-sourced annotations. As seen in the figure, there is a strong correlation ($\rho = 0.74$) between the conflict escalation levels through the indirect inference method and the direct assessment method.

In this work, we consider three possible situations in order to study the evolution of conflict in the conversations: *escalation*, *de-escalation*, and *constant*. Based on consolidated levels of conflict escalation, we split the clips into those three classes using quantiles

Table 3. Confusion matrix of assigned labels using indirect inference (rows) and direct assessment (columns) in terms of number of clips.

	De-escalation	Constant	Escalation	Sum
De-escalation	176	67	9	254
Constant	67	132	69	268
Escalation	13	68	189	270
Sum	258	267	267	792

so that the number of clips are as equivalent as possible, i.e.,

$$c_t = \begin{cases} \text{Escalation} & q_{\frac{1}{3}} \leq e_t; \\ \text{Constant} & q_{\frac{2}{3}} \leq e_t < q_{\frac{1}{3}}; \\ \text{De-escalation} & e_t < q_{\frac{2}{3}}, \end{cases}$$

where $q_{\frac{1}{3}}$ and $q_{\frac{2}{3}}$ represent the first and the second tertiles of score distribution. Table 3 shows the confusion matrix of assigned labels using indirect inference and direct assessment in terms of the number of clips. The agreement between the two methods (in terms of whether the labels are the same or not) is 63%.

3. DETECTION OF CONFLICT ESCALATION

3.1. Feature Extraction

The features used in this work are similar to those introduced in our previous work [6, 8] and they consist of 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, while prosodic features are used to capture the speaking styles of conversations, i.e. the way speakers convey their speech. These features have shown promising results in automatic detection of agreement/disagreement [10, 11], social roles [12], level of engagement [2, 5].etc.

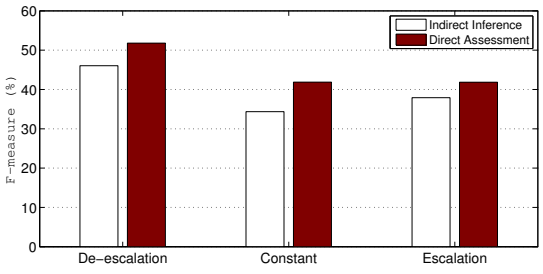
Extracting features described above, either conversational or prosodic, requires speaker segment information, i.e. who speaks when for how long. In our previous work, we used manual segmentation for extracting various statistics. In fact, the Canal9 database is annotated into speaker turns, i.e., who spoke when, including overlapped regions and a mappings between speakers and their roles, i.e., moderator or guest. Towards a fully automated system, we use an automatic speaker diarization method [13] and an overlap speech detection method [14] (see [8] for more details).

3.2. Classification Results

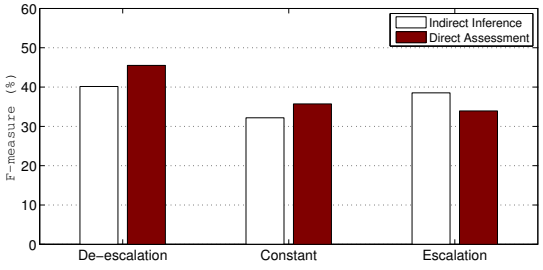
As we discussed earlier, we focus on three possible situations in the evolution of conflict: *escalation*, *de-escalation*, and *constant*. Experiments are performed using 5-fold cross validation to provide speaker and debate independent training/testing subsets. The entire dataset is split into 5 folds 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 fold 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

Table 4. Performance of classifying conflict escalation in terms of WA and UA according to annotation methods and feature extraction strategies. The performance for chance level is computed by assigning the majority class to all the classification.

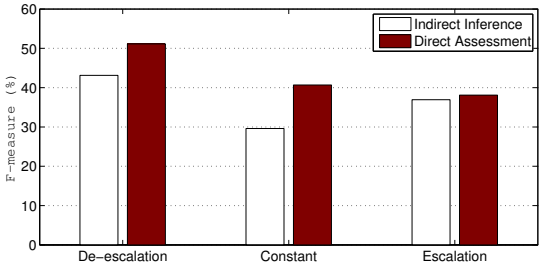
	Indirect Inference		Direct Assessment	
	WA (%)	UA (%)	WA (%)	UA (%)
Manual Segmentation	39.5	39.7	45.2	45.3
Speaker Diarization	37.6	37.8	42.4	42.5
Speaker Diarization w/ overlap detection	36.9	37.1	43.7	43.8
Chance level	34.1	33.3	33.7	33.3



(a) manual segmentation



(b) speaker diarization



(c) speaker diarization w/ overlap detection

Fig. 4. Per-class F-measure for classification tasks using features extracted from (a) manual segmentation, (b) speaker diarization and (c) speaker diarization with overlap detection.

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 using the LIBSVM toolkit [15]. The classification performances are reported in terms of unweighted accuracy (UA) as well as weighted accuracy (WA) which are commonly used in paralinguistic classification tasks [16]. Table 4 shows the performance of classification tasks according to annotation methods and segmentation strategies and the performance for chance level is computed by assigning the majority class to all the classification. For further investigation, Fig. 4 provides per-class F-measure of the classification tasks. They clearly show that labels that are consolidated by the answers of the direct assessment method can yield higher performance in classification tasks. That proves the hypothesis, i.e., the direct assessment method appropriately annotate the subjects' perception of conflict escalation rather than indirect inference method, by showing the labels are correlated with the non-verbal features to yield higher classification performance.

It also shows the consistent results with our previous work [8] that utilizing an automatic speaker diarization algorithm instead of manual segmentation can degrade performance. This is reasonable because errors from automatic speaker diarization can propagate by providing imprecise (missing or adding) speaker segment information which is crucial to extracting most features mentioned above. Although automatic overlap detection can bring some benefits especially with the labels from direct assessment, crucial information such as speaker roles, i.e., moderator or participants, is still missing, which consequently motivates our future work on role recognition.

4. CONCLUSIONS

We studied annotation and detection of conflict escalation in multiparty spontaneous conversations, particularly broadcast political debates. For annotation, we compared two different strategies in conflict escalation assessment: indirect inference and direct assessment. We showed that the labels from both methods are highly correlated ($\rho = 0.74$ in continuous values and 63% agreement in ternary classes). However, empirical results with 792 pairs of consecutive clips in classifying three types of conflict escalation, i.e., *escalation*, *de-escalation*, and *constant*, showed that labels from direct assessment yielded higher classification performance than the one from indirect inference (39.7% unweighted accuracy for indirect inference and 45.3% for direct assessment). This suggests that perceiving actual difference between two consecutive clips is required to annotate conflict escalation.

In the future, as we discussed, we will study automatic role recognition methods (e.g., [17]) to incorporate with the automatic speaker diarization methods. This is expected to compensate for the missing role information of participants through the automatic speaker diarization methods. We will also investigate regression tasks, similarly done in [18], to regress the level of conflict escalation.

5. ACKNOWLEDGEMENT

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

6. REFERENCES

[1] 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.

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

[3] 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.

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

[5] 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 *Proceedings of InterSpeech*, 2010.

[6] S. Kim, F. Valente, and A. Vinciarelli, "Automatic detection of conflicts in spoken conversations: ratings and analysis of broadcast political debates," in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing*, Mar. 2012, pp. 5089–5092.

[7] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings of Interspeech*, 2013.

[8] S. Kim, S. H. Yella, and F. Valente, "Automatic detection of conflict escalation in spoken conversations," in *Proceedings of INTERSPEECH*, Sep. 2012.

[9] 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.

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

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

[12] F. Valente and A. Vinciarelli, "Language-independent socio-emotional role recognition in the ami meetings corpus," in *Proceedings of Interspeech*, 2011.

[13] 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.

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

[15] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

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

[17] H. Salamin and A. Vinciarelli, "Automatic role recognition in multiparty conversations: an approach based on turn organization, prosody and conditional random fields," *IEEE Transactions on Multimedia*, 2012.

[18] S. Kim, M. Filippone, F. Valente, and A. Vinciarelli, "Predicting the conflict level in television political debates: an approach based on crowdsourcing, nonverbal communication and gaussian processes," in *ACM Multimedia*, 2012.