Predicting the Conflict Level in Television Political Debates: an Approach Based on Crowdsourcing, Nonverbal Communication and Gaussian Processes

Samuel Kim[1]
Maurizio Filippone[2]
[1]Idiap Research Institute
CP592 - 1920 Martigny (CH)
samuel.kim@idiap.ch
maurizio.filippone@glasgow.ac.uk

Fabio Valente[1]
Alessandro Vinciarelli[1,2]
[2]University of Glasgow
Sir A.Williams Bldg. - G12 8QQ Glasgow (UK)
fvalente@idiap.ch
vincia@dcs.gla.ac.uk

ABSTRACT

One of the most recent trends in multimedia indexing is to represent data in terms of the social and psychological phenomena that users perceive. In such a perspective this article proposes an approach for the automatic detection of conflict level in television political debates. The proposed approach includes the use of crowdsourcing techniques for modeling the perception of data consumers, the extraction of (language independent) nonverbal behavioral cues and the application of regression techniques based on Gaussian Processes. The experiments have been performed over 1430 clips of 30 seconds extracted from 45 political debates (roughly 12 hours of material). The results show that a correlation up to 0.8 can be achieved between the actual and predicted conflict level.

Categories and Subject Descriptors: H.3.1 [Content Analysis and Indexing].**General Terms:** Experimentation. **Keywords:** Conflict, Social Signal Processing, Nonverbal Vocal Behavior, Multimedia Indexing

1. INTRODUCTION

Face-to-face social interactions are one of the most common subjects in multimedia data (e.g., television programs, Youtube videos, movies, etc.). Social and psychological phenomena are one of the most salient features of this type of material [7] and they influence, to a significant extent, what people perceive and remember about the data [3]. Hence, it is not surprising to observe that recent trends in multimedia indexing try to capture how people perceive data from a social and psychological point of view [9]. In such a perspective, this work considers the detection and measurement of conflict in multiparty conversations. The proposed approach includes three main aspects. The first is the use of crowdsourcing for modeling the perception of potential data consumers, especially when it comes to nonverbal behavioral cues that constitute the physical, machine detectable trace of conflict. The second is the extraction of features that account for the cues identified above, and the third is the prediction of the conflict level with different regression approaches, from simple linear techniques up to methods based on Gaussian Processes.

Conflict is the focus of this work because it is recognized as one of the main dimensions along which an interaction is perceived and assessed [4]. For what concerns nonverbal communication, Social Signal Processing [10] has shown that behavioral cues are one of the main keys towards automatic analysis and understanding of social phenomena. Furthermore, language independence is a major advantage when dealing with repositories of multilingual material.

The experiments have been peformed over a corpus of 1430 clips extracted from 45 political debates televised in Switzerland (see Section 2 for more details), for a total of 11 hours and 55 minutes of material. The data was annotated with Mechanical Turk (the Amazon crowdsourcing facility) in terms of conflict level, a continuous score that accounts for how competitive and aggressive is the exhange between debate participants. The prediction of the conflict level has been performed with several regression approaches and it has been assessed in terms of correlation between actual and predicted conflict level. The results show that a correlation up to 0.8 can be achieved.

The rest of the paper is organized as follows: Section 2 shows how the data has been modeled in terms of conflict perception, Section 3 presents the cues extraction approach, Section 4 illustrates the regression approaches, and Section 5 presents experiments and results. The final Section 6 draws some conclusions.

2. DATA COLLECTION AND PERCEPTION MODELING

The experiments of this work have been performed over a collection of clips extracted from 45 political debates. The clips have been identified as follows: first, the debates have been split into uniform, non-overlapping, 30 seconds long segments. Then, all segments where at least two people talk have been retained. The resulting 1430 clips (11 hours and

MM'12, October 29–November 2, 2012, Nara, Japan.

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

Table 1: The table shows the questionnaire used to annotate the conflict database of the case study. The first column reports the question ID, the second column shows the question with its sign and the third column says whether the question belongs to the Inferential (I) or Physical (P) layer.

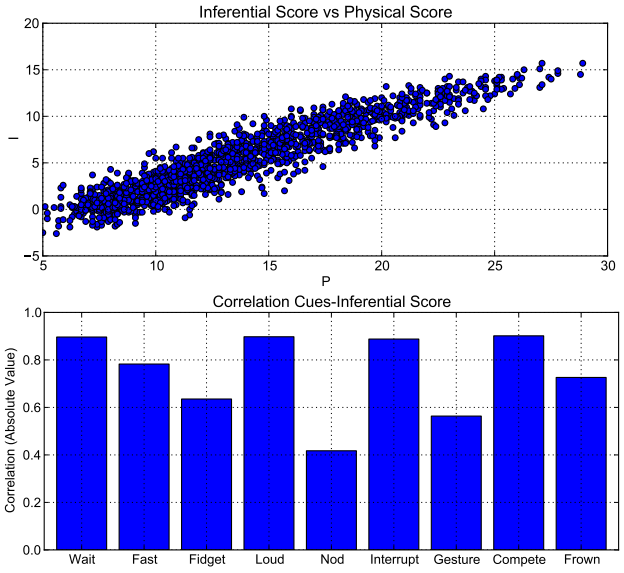

Figure 1: The upper plot illustrates the relation between Inferential and Physical scores. Each point is a clip and the correlation between I and P is 0.95. The lower plot shows the correlation between individual questions of the physical layer and I.

55 minutes in total) have been annotated with Mechanical Turk, the Amazon crowdsourcing system[1].

The annotation has been performed with the questionnaire of Table 1. Following common behavior observation methodologies [5], the questionnaire includes two *layers*: the first, called *physical* (denoted with "P" in the table) includes questions about objective observations about the behavior of the debate participants. The second, called *inferential* (denoted with "I" in the table) includes questions about the subjective interpretation that the annotators give about the scene they observe.

All questions are associated to 5 points Likert scales that range from "*Strongly Disagree*" to "*Strongly Agree*" for the inferential layer and from "*Never*" to "*Always*" for the physical layer. In both cases the scales are mapped into the interval $[0, 4]$. Thus, it is possible to calculate two scores (one per layer) by summing over the answers provided by the annotators. The sign changes (see Table 1) depending on whether the question is posed positively (e.g.,"people compete to talk") or negatively (e.g., "people wait for their turn before they talk") with respect to the presence of conflict. The number of assessors per clip is 10. The inferential and physical scores assigned to a given clip are the average of the corresponding scores assigned individually by each assessor.

The use of two layers aims at identifying the observable, machine detectable behavioral cues that actually explain the perception of the assessors. In the upper plot of Figure 1, the coordinates are physical (P) and inferential (I) score, respectively. Each dot corresponds to a clip and the correlation between the scores is 0.95 (90% of the variance in common). Hence, the behavioral cues targeted by the questions of the physical layer seem to explain to a large extent the perception of the assessors. However, not all cues have the same effectiveness. The lower plot of Figure 1 shows the individual correlation between the questions of the physical layer and the inferential score. The questions related to body, face and head cues (Q4 for fidgeting, Q7 for head nods and shakes, Q10 for gestures and Q15 for frown) appear to be less effective than those conveyed by voice and speech (Q2 for people waiting for their turn, Q3 for people talking fast, Q6 for people talking loud, Q9 for interruptions, Q13 for people competing to speak).

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

3. FEATURE EXTRACTION

The first step of the feature extraction process is the diarization, i.e. the segmentation of the data into intervals expected to correspond to one voice only (the diarization provides information on *who spoke when*). In this way, the feature extraction process can be applied not only to each clip as a whole, but also to individual debate participants. Three different approaches have been used: manual diarization, automatic speaker diarization, and automatic speaker diarization including overlapping speech detection. Based on the analysis of the annotations (see lower plot in Figure 1) the extraction process targets conversational (questions 2, 9 and 13) and prosodic (questions 3 and 6) features.

Conversational Features.

Four types of features account for conversational behavior. The first corresponds to **turn duration statistics**, namely mean, median, maximum, variance and minimum of speaker turns duration in the clip as well as the number of turns. The second type includes **total speaking time statistics**: mean, median, maximum, variance and minimum of total speaking time for individual speakers in the clip as well as the number of people speaking. The third type accounts for **turn-taking patterns**: each participant in the discussion is either the moderator or a participant that belongs to one of the two coalitions that oppose one another. In other words, each person is assigned a label $r \in (m, g1, g2)$. The bi-gram counts $p(r_t, r_{t-1})$, where r_t is the label of the

speaker that holds the floor of the conversation at turn t, are expected to change depending on whether there are conflict and competition or not.

The rest of the conversational features concern overlapping speech and include the **amount of overlap** relative to the clip duration; in addition to the total amount of overlap O_T, three types of role based overlaps are considered, i.e., overlap between moderator and guests OV_{MG}, overlap between guests belonging to the same group OV_{SG} and the overlap between guests belonging to opposite groups OV_{OG}. Overlaps between more than two participants are not considered. Finally the features include the **turn keeping/turn stealing ratio** in the clip, defined 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 choice consists in capturing aggressive interruptions aimed at grabbing the floor of the conversation.

Prosodic Features.

To extract prosody features, pitch and intensity are estimated using the Praat Toolkit (`http://www.praat.org/`) every 10 ms and two types of statistics are extracted: one from the entire clip (30 seconds) and one for each speaker turn in the clip. The first models the entire conversation while the latter models the prosodic behavior of individual speakers.

The first group of features includes then **clip-based statistics**: mean, median, standard deviation, maximum, minimum and quantiles (0.01, 0.25, 0.75 and 0.99) of pitch and intensity statistics obtained from the entire clip[2]. The second group includes **speaker turn-based statistics**: mean, median and standard deviation of pitch and intensity obtained over individual speaker turns (similarly to the clip-base statistics). The statistics above are estimated not only where only one person talks, but also over overlapping speech-segments. The accurate estimation of pitch in these regions is still an open issue, but prosody information during overlapping speech should not be neglected.

4. REGRESSION

This section briefly presents the regression models employed in this work to characterize the mapping between features and conflict level. In the remainder of this paper, the n training samples will be denoted by d-dimensional feature vectors $\mathbf{x}_i$ and the corresponding target values representing the conflict level by y_i. In order to keep the notation uncluttered, model parameters will be generally denoted by $\boldsymbol{\theta}$, all training samples by the $n \times d$ matrix X, and all corresponding target values by $\mathbf{y}$. For the sake of completeness, this paper focuses on a number of linear and nonlinear regression models based on different paradigms as discussed next.

Bayesian Linear Regression (BLR).

The BLR approach employed in this work models the target variable y as the sum of a linear combination of the features $\mathbf{x}$, a bias term, and a stochastic term $\varepsilon \sim \mathcal{N}(\varepsilon|0, \theta_{\sigma^2})$, so that $p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{y}_i|\mathbf{x}^{\mathrm{T}}\boldsymbol{\theta} + \theta_{\mathrm{bias}}, \theta_{\sigma^2})$. This work adopts a fully Bayesian treatment of this model, and assumes weakly informative priors on all model parameters that are conjugate to the likelihood $p(\mathbf{y}|X, \boldsymbol{\theta})$. This makes it possible to get the posterior distribution over $\boldsymbol{\theta}$ in closed form and to obtain a predictive distribution where the parameters are integrated out of the model (see section 3.3 of [1] for full details).

Gaussian Processes for Regression (GPR).

One limitation of linear regression is the difficulty to choose an appropriate set of basis functions to use in a given application. GPR addresses this by adopting a Gaussian Process (GP) [6] prior over the latent functions $f(\mathbf{x}, \boldsymbol{\theta})$ that are used to model $y = f(\mathbf{x}, \boldsymbol{\theta}) + \varepsilon$, with $\varepsilon \sim \mathcal{N}(\varepsilon|0, \theta_{\sigma^2})$. The GP prior is fully specified by the mean function, that in this work is assumed to be zero, and by the covariance function k, that is parametrized by $\boldsymbol{\theta}$. The choice of the covariance function reflects the properties of the functions that can be drawn from the GP. In this work, two covariance functions are tested that yield a nonlinear mapping between features and targets, namely the Radial Basis Function (RBF) covariance

$$k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta}) = \theta_a \exp\left[-\theta\|\mathbf{x}_i - \mathbf{x}_j\|^2\right].$$

and the RBF covariance with Automatic Relevance Determination (ARD) [1, 6]:

$$k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta}) = \theta_a \exp\left[-\sum_{r=1}^{d}(\boldsymbol{\theta})_{(r)}(\mathbf{x}_i - \mathbf{x}_j)^2_{(r)}\right].$$

Both covariances have an amplitude parameter θ_a; in the first case, there is one global length-scale parameter θ, whereas in the second case there is one length-scale parameter for each feature. This makes the RBF ARD covariance function suitable for interpretation of the importance of the individual features in the regression task.

In the case of the RBF covariance function employed here, it is not possible to obtain the posterior distribution over $\boldsymbol{\theta}$ in closed form. This work adopts a type II Maximum Likelihood approach [1], whereby $\log[p(\mathbf{y}|X, \boldsymbol{\theta})] = \log[\mathcal{N}(\mathbf{y}|\mathbf{0}, K(\boldsymbol{\theta}) + \theta_{\sigma^2} I)]$ is optimized with respect to $\boldsymbol{\theta}$.

Support Vector Regression (SVR).

SVR is a sparse nonlinear regression method making use of kernel functions to map the input space into a possibly infinite dimensional space where a linear regression task is performed. SVR is formulated as the optimization of a linear combination of the so called *empirical risk* which accounts for the error in fitting the training data and a complexity term. In the objective function, an inverse regularization parameter C multiplies the empirical risk.

This work uses the standard ε-insensitive loss function [8] as implemented in the `LIBSVM` library [2]. The kernel used in the experiments are the linear (LIN) kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j$$

and Radial Basis Function (RBF) kernel

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left[-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2\right].$$

In the experiments, C and γ (in the case of the RBF kernel) have been tuned using 5-fold cross-validation within the training set.

[2]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).

	BLR	GPR RBF	GPR RBF ARD	SVR LIN	SVR RBF
Manual	0.809	0.808	0.817	0.809	0.810
Automatic	0.712	0.765	0.710	0.709	0.739
Auto w.o.s.	0.776	0.783	0.774	0.772	0.777
Manual	2.27	2.27	2.22	2.29	2.27
Automatic	2.71	2.44	2.70	2.71	2.58
Auto w.o.s.	2.42	2.38	2.43	2.44	2.41

Table 2: Correlation coefficients (upper part) and Root Mean Square Errors (lower part) achieved with different regression approaches for manual diarization, automatic diarization, and automatic diarization with overlapping speech.

5. RESULTS

This section reports the results obtained applying the regression models to the data set presented in the previous section. Let m be the number of test samples, and let $\hat{y}_i$ represent the prediction for the i-th test point with actual target value y_i. Also, let $\hat{\mu}$ and μ be the corresponding mean values and $\hat{\sigma}^2$ and σ^2 the corresponding variances across the test set. The two evaluation metrics used to assess the performance in predicting the conflict level are the Correlation Coefficient (CC)

$$\mathrm{CC} = \frac{1}{m\,\sigma\,\hat{\sigma}} \sum_{i=1}^{m} (y_i - \mu)(\hat{y}_i - \hat{\mu}) \qquad (1)$$

and the Root Mean Square error (RMSE)

$$\mathrm{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2} \qquad (2)$$

Results are reported in Table 2; the values represent the average of the two scores computed using 5-fold cross-validation. The split in 5 folds was done in such a way that the experiments are debate and speaker independent (no samples corresponding to the same debate or person appear in both training and test set).

All regression models seem to achieve a similar performance both in terms of CC and RMSE. Also, the use of covariance/kernel functions with ARD or not, does not lead to differences in performace, but the ARD covariance allows a direct interpretation of the relative importance of the features; a thorough analysis and discussion of this aspect cannot be reported here for lack of space. The use of manual diarization allows one to make predictions with CC around 0.8 and a corresponding RMSE of about 2.25. The reduction in performance due to the use of automatic diarization is significant, but the CC score remains around 0.75 with a corresponding RMSE of about 2.5. The use of automatic diarization with overlapping speech, which is one of the unique features of the present work, allows one to improve with respect to automatic diarization achieving results half-way between manual and automatic diarization.

6. CONCLUSIONS

This paper has presented experiments on automatic detection and measurement of conflict level in political debates. The results show that (i) the extraction of features accounting for nonverbal communication in speech (prosody, speaker adjancency patterns, overlapping speech, etc.) and (ii) the application of regression approaches allows one to predict the conflict level with correlation up to 0.8. Furthermore, the use of crowdsourcing to model the perception of data consumers seems to be a suitable method to identify the nonverbal cues (hence the features) that can lead to a satistfactory performance.

The results are interesting under two main respects. The first is that the conflict level can be used as a content descriptor that accounts for what users actually perceive in the data. The second is that the experiments provide indications on the physical traces of conflict in conversations. This is important for any technology expected to interact with people like, e.g., social robots, artificial agents, intelligent interfaces, etc.

7. ACKNOWLEDGMENT

This work is supported by the European Community's Seventh Framework Program (FP7/2007-2013), under grant agreement no.231287 (SSPNet). Furthermore, it is supported by the Swiss National Science Foundation through the National Centre of Competence in Research on Interactive Multimodal Information Management (IM2). The authors would like to thank Marcello Mortillaro for scientific advice.

8. REFERENCES

[1] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1:27, 2011.

[3] S. Dumais, E. Cutrell, J.J. Cadiz, G. Jancke, R. Sarin, and D.C. Robbins. Stuff i've seen: a system for personal information retrieval and re-use. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 72–79, 2003.

[4] J.M. Levine and R.L. Moreland. Small groups. In D. Gilbert and G. Lindzey, editors, *The handbook of social psychology*, volume 2, pages 415–469. Oxford University Press, 1998.

[5] P. Martin and P. Bateson. *Measuring Behaviour. An Introductory Guide*. Cambridge University Press, 2007.

[6] C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[7] B. Reeves and C. Nass. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press, 1996.

[8] V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

[9] A. Vinciarelli and M. Pantic. Implicit Human-Centered Tagging. *IEEE Signal Processing Magazine*, 26(6):173–180, 2009.

[10] A. Vinciarelli, M. Pantic, and H. Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing Journal*, 27(12):1743–1759, 2009.