Let Me Finish: Automatic Conflict Detection Using Speaker Overlap

Félix Grèzes[1*], *Justin Richards*[2*], *Andrew Rosenberg*[1]

[1]Department of Computer Science, The CUNY Graduate Center, New York, USA
[2]Departement of Linguistics, The CUNY Graduate Center, New York, USA
{ fgrezes@gc.cuny.edu, jrichards@gc.cuny.edu, andrew@cs.qc.cuny.edu }

Abstract

The automated detection of conflict will be a crucial feature of emerging speech-analysis technologies, whether the purpose is to assuage conflict in online applications or simply to mark its location for corpus analysis. In this study, we examine the predictive potential of overlapping speech in determining conflict, and we find that this feature alone is strongly correlated with high conflict levels as rated by human judges. In analyzing the SSPNET debate corpus, we effect a 2.3% improvement over baseline accuracy using speaker overlap ratio as a predicted value, suggesting that this feature is a reliable proxy for conflict level.

In a follow-up experiment, we analyze the patterns of predicted conflict in the beginning, middle and end of an audio clip. Our findings show that the beginning and final segments are more predictive than the middle, which indicates that a primacy-recency effect is bearing on the perception of conflict. Since the beginning segment itself can be quite predictive, we also show that accurate predictions can be made dynamically, allowing for real-time classification during live debates.

Index Terms: conflict detection, paralinguistics

1. Introduction

The automatic detection of conflict among speakers is an important frontier in spoken language processing. Reliable conflict detection will also be a boon to customer service, automated mental health counseling, and the deployment of artifically intelligent agents in a variety of roles. Also, segments of a recorded audio database that exhibit high conflict may be of special interest for information retrieval in security and intelligence applications. However, conflict is a rather general phenomenon that manifests in various ways. It remains to be seen what features of speech are most likely to elicit a high-conflict judgment from raters.

We focused our research on feature identification, basing our study on the SSPNet Conflict Corpus provided in the INTERSPEECH 2013 ComParE conflict sub-challenge[1]. Previous work on recorded meetings has shown that prosodic cues can be somewhat indicative of hot spots, a term used generally to describe spikes of intensity[2]. Conflict is more narrowly defined than emotional outbursts or intensity. That said, the baseline results of this subchallenge indicate that prosodic features can predict conflict with above-chance accuracy [1].

To improve the automatic recognition of conflict, we look beyond the vocal features of individual speakers. Qualities of a single stream of speech, whether lower level features like speech rate, intensity, and pitch or higher level features such as emotional valence, can often tell us something about conflict level. However, they may also be qualities intrinsic to that speaker. An individual may talk fast or sound pressured by nature, not by circumstance.

We hypothesize that higher-order features, characterized by the dialogue at large, may be able to transcend speaker-specific idiosyncrasies. Such features as speaker turn duration and interruption have yielded success for others studying the same corpus[3]. The same study includes as a feature, when predicting conflict for a given sound clip, the conflict level of adjacent clips, and we draw inspiration from this technique as well.

The dialogue-level feature most closely associated with conflict seems intuitively to be speaker overlap — the phenomenon of one speaker talking over another rather than allowing the other to finish. In our first experiment, we explore the relationship between overlapping speech and user ratings of conflict (Section 2.1).

Finding a strong relationship between overlap and conflict, we examine those tokens that have a high overlap ratio and low conflict score. We find that these clips are charactized by either (a) shared laughter among the speakers or (b) a de-escalation of conflict toward the second half of the clip. This motivates a second phase of our investigation. Presuming that raters are likely to rate a sample as low-conflict if it exhibited a decline in conflict from beginning to end, we add what amounts to a conflict gradient to our feature set, supplying conflict scores for the first, middle, and final thirds of each clip.

2. Methodology and Experimental Results

The baseline results given by Schuller et al. [1] are quite high on the *Conflict Sub-Challenge*. A linear kernel Support Vector Machine (SVM) algorithm trained using Sequential Minimal Optimisation (SMO) yields an unweighted average recall (UAR) of 79.1%on the development set and 80.8% on the test set. These baseline experiments, as well as those to be outlined in this section, are all run using Weka 3.7.1 [4]. The model computes a zero-to-10 real-valued conflict score for each clip and assigns a label of either "low" or "high" according to a midpoint split of the conflict score.

2.1. Overlap

Our primary hypothesis states that the ratio of overlapping speech to non-overlapping speech in a debate is a useful feature for the detection of conflict levels. Since the SSPNet Conflict Corpus comes with hand-labeled meta-data containing the time and duration of speech for each speaker, as well as any overlapping speech, our first step was to incorporate this true ratio, henceforth referred to as*gold overlap*, alongside the 6,373 acoustic features in the baseline experiment.

*These authors contributed equally to this work

Table 1 shows the results of the baseline model on the development set, trained on the training set. Our results follow. The numbers are counts of each prediction on the data set.

	Predictions	
Labels	High	Low
High	88	25
Low	25	102

Predictions on development set
UAR = 79.1%

Table 1: *Baseline Experiment on Class detection*

Table 2 shows the results of two experiments based on the gold-overlap feature. The first was trained with gold-overlap as the unique feature, while in the second gold-overlap was used alongside the 6,373 acoustic features from the baseline experiment. The evaluation is done on the development set and uses the same experimental parameters as the baseline.

	Predictions	
Labels	High	Low
High	63	55
Low	12	115

Using only Gold-overlap
UAR = 74.2%

	Predictions	
Labels	High	Low
High	78	35
Low	13	114

Using Gold-Overlap and Baseline Features
UAR = 79.4%

Table 2: *Gold-Overlap Experiments on Class detection*

The first figure of Table 2 shows that, even when used alone, the gold ratio of overlapping speech to non-overlapping speech can predict the conflict class with meaningful accuracy. The second figure shows that when combined with other features, the ratio of overlapping speech will bolster the accuracy of the baseline experiment. These preliminary results encourage us to further explore our hypothesis.

Our next step is to build a classifier to predict this ratio of overlapping speech, so that it may be applied to unlabeled data. There are a variety of techniques that can be applied to recognizing overlapping speech. We hypothesize, however, that the baseline feature set with its 6,373 features might already be encoding some of this information through the spectral and intensity features. Therefore we train a regression model using this baseline feature set to predict the gold-overlap ratio score. We find that Weka's SMOreg algorithm, using complexity coefficient c=5.0E-4, produces an effective regression model with 80% correlation coefficient with the gold-overlap feature evaluated with 20-fold cross validation.

We then apply this regression model to the corpus to obtain a new feature: predicted overlap. To verify the usefulness of this new feature, we train a conflict classifier using solely the predicted overlap. To ensure that predicted overlap is the true cause of improvement over the baseline results, all the conflict classifiers are trained using the same parameters presented in the baseline experiments[1], i.e.: SMO with SVM complexity parameter C=0.1. The performance of this classifier on the devel set can be seen in Table 3.

Here we find that using a prediction of overlapping speech as the sole feature of an SMO classifier produces a model with *better* performance than the baseline experiment. This provides

	Predictions	
Labels	High	Low
High	80	33
Low	12	115

Predictions on development set
UAR = 80.5%

Table 3: *Experiment using only Predicted Overlap*

significant support for our hypothesis that overlapping speech is a reliable predictor of conflict.

Evaluating the predictions of this model on the test set, we obtain the results shown in Table 4. Here we find that this single feature classifier produces an improvement of 2.3% over baseline UAR. Moreover, it should be noted that this model is based only on the train partition, as opposed to the baseline results on the test set trained on both the train and development sets.

	Predictions	
Labels	High	Low
High	130	41
Low	22	204

UAR = 83.1%

Table 4: *Prediction on Test set*

Table 2 shows that the overlap ratio can be complemented with the baseline feature set to improve the prediction accuracy. Therefore our next step is to build upon the overlap-only classifier by adding more acoustic features. However, our attempts to add baseline features to the set do not improve the results of Table 3. In fact, adding even one feature, the best one according to Weka's Infogain dimensionality reduction technique, seems to reduce the quality of the classifier.

2.2. Escalation or De-escalation

Rarely does a debate consist of constant arguing, shouting, and interruptions from start to finish. Instead we imagine conflict might escalate during controversial topics and de-escalate afterwards. Another way of picturing this is that conflict is a local feature in a debate, being low most of the time and peaking when the contentious topics are discussed. Our task, meanwhile, is to predict the judgments of human raters, and intuition tells us that those raters are more likely to deem an exchange contentious if the disagreement escalates throughout the exchange. Thus, we hypothesize that escalating conflict will correlate more closely with a global conflict label than de-escalating conflict, even when the global predicted conflict level is held constant. Anticipating that conflict-level difference is a linear relationship that can be learned automatically by the SMO classifier, we simply add to each feature vector the predicted conflict for each of three segments.

2.2.1. Corpus Preparation

We first divide each audio file into three segments of equal (10-second) duration. Four files, three in the training set and one in the test set, are shorter than 30 seconds and not compatible with this implementation. These are recordings of the end of the debate program, containing large proportions of music, and

we discard them for these experiments. Since these files contain less speech time, they have a low conflict score. Thus, for the anomalous instance in the test set, we manually set the prediction to low.

We divide the rest of the files into thirds because smaller segments might not contain enough information, and three segments of 10 seconds should be enough to reveal the escalation or de-escalation of conflict. On each segment we use OpenSMILE's [5] feature extractor, producing the same 6,373 features as the original .arff files supplied in the challenge data package. We then apply our predictors, learned on the non-segmented data set (Section 2.1), for overlapping speech, conflict score and conflict class in order to get local predictions.

2.2.2. *Preliminary Analysis of the Segmented Corpus*

We generated a predicted class (high,low) label for each of three segments for each file. In order to understand what these predictions may tell us about a clip's "true" global conflict rating, we generated association frequency counts for each pattern of three labels. Table 5 shows the relative frequency for each of eight possible patterns. This proportion is calculated as the number of associations between a segment pattern and the "high" label divided by the total number of occurences for that pattern. Some of the data are predictable — high-high-high is the best predictor, low-low-low is the worst — but other patterns yield unexpected results. Our hypothesis regarding escalation and de-escalation is not robustly supported. Low-high-high, for example, is not more strongly correlated with high than is high-high-low. One thing the data do strongly suggest is that when a class label is positioned both at the beginning and the end of the clip, the clip is especially likely to be globally labeled as such. That is, low-high-low is more predictive of a low global label than either low-low-high or high-low-low, and the same is true for the high label.

Conflict Predictions				High Label
Start	Mid	End	Instances	Proportion
Low	Low	Low	347	10%
Low	Low	High	114	40%
Low	High	Low	85	34%
Low	High	High	65	77%
High	Low	Low	160	37%
High	Low	High	63	79%
High	High	Low	78	75%
High	High	High	118	90%

Table 5: *Statistics on the segmented corpus*

2.2.3. *Experiment Results*

After our premiliminary analysis of the segmented data, we proceeded to train a concflict class predictor with the set of features built from the segments. For this experiment we extracted and used the following new features: predicted ratio of overlapping speech on the start, middle and end segment; predicted class of each segment and finally the confidence of eash of these class predictions. The models used to generate these predictions are those described in Section 2.1 .

Our results with this approach were only partially successful. The UAR of our predictions was just as high as to those given by our models built on the whole audio file, up to 80% when trained and cross-validated on the development and training set together. Since this accuracy is as high as our best model, it suggests that escalation and de-escalation are indeed useful features to detect conflict. While we have not shown that it performs better than global features, it remains an interesting concept for future work.

We also analyzed the predictive power of the first segment of the debates alone, as well as the first and second segments together, in each case ignoring the features from the third segment. Table 6 shows the predictions on the development set of two models: the first trained only on the features extracted from the first segment of the audio file, and the second using features from both the first and middle segments.

	Predictions	
Labels	High	Low
High	71	42
Low	29	98

Using Solely the first segment
UAR = 70.4%

	Predictions	
Labels	High	Low
High	68	45
Low	10	117

Using the first and middle segments
UAR = 77.1%

Table 6: *Class Detection using Starting Segments*

While the predictions are naturally less accurate than those from a model learned on the whole file, they do suggest that conflict detection can be performed in an online fashion, implemented while listening and reacting to live debates.

3. Discussion

In Section 2.1, we describe experiments that show that a single feature — predicted overlap — is a more reliable predictor of conflict than a large and diverse set of acoustic-prosodic features. What makes this result still more remarkable is that the predictor of overlap that we use in this work is trained on *the same* set of acoustic-prosodic features. Essentially, this process is replacing the ground-truth label of CONFLICT with OVERLAP, then learning a mapping from OVERLAP to CONFLICT. We note that this process does not necessarily add any information to the conflict prediction system, however, it demonstrates that either overlap is 1) easier to predict or 2) more reliably labeled than conflict directly. One possible explanation for this is that overlap is a relatively objective measure. The speaker diarization annotation is fairly conservative in its indication of overlap, ignoring backchannels and overlap at smooth turn changes. The conflict rating, on the other hand, is based on aggregates of user responses. These responses likely have more noise than the overlap scores, making them more difficult to predict directly.

It may be valuable to consider the relationship between overlap ratio and conflict rating as a linear chain generative model, where conflict (c) is independent of acoustics (a) given overlap (o), i.e. $p(c|o,a) \approx p(c|o)p(o|a)$ ac.Under this generative perspective, conflict ratings are a noisy observation conditioned on overlapping speech. By observing this conditioning variable, we are able to more accurately model the overall relationship between acoustics and conflict.

These results open the possibility of predicting overlap based on other, more specifically engineered approaches. There is a wide range of techniques for the prediction of overlapping speech. Yamamoto et al. [6] use support vector regression al-

beit with more compact feature representation than used here. Quinlan and Asano [7] describe an approach originally developed to identify, track, and count the number of speakers in a room. Geiger et al. [8] describe an approach wherein each speaker's speech is projected into a unique basis space. Each of these seek to identify specific regions of overlap, while our approach considers each token as a whole and tries to predict overlap ratio directly.

Here it's important to highlight one of the more striking observations from our overlap experiment: Predicted overlap ratio performs better than actual overlap on the development set. We can imagine a few explanations for this. Actual overlap is based on manual annotation of a sound file, and it does not include small particles of overlap, such as failed interruptions. A regression model trained on actual overlap may find evidence of these failed interruptions (a fairly common occurence in heated discussion) in the acoustic signal of a debate file. An alternate or additional explanation is that, by training on actual overlap using thousands of vocal-signature features, we may have identified a quality of an individual's speech that manifests when that individual is being talked over. Occasionally, that quality may emerge in monologue speech, perhaps when the individual feels especially challenged or pressured. This phenomenon would also add more information to the predicted overlap regression model. Thus, while the abovementioned approaches could refine the detection of actual overlapping speech, other techniques might be developed to detect the vocal signature of what might be called "competitive" speech. We see no reason that a combination of these approaches or other approaches couldn't be employed to improve conflict detection in future work.

In the realm of segmentation, our research reveals some interesting patterns that point toward future work. Our class predictions on file segments clearly show correlation between certain patterns and global conflict labels. Our first observation aligns with intuition: If the majority of segments associate with a certain label, that label is most likely the global one. Within this majority subset, though, we find something interesting. The co-occurence of a label in the first and last position is more predictive than its consecutive occurence in any two positions. This may be evidence of the primacy-recency effect in psychology, which indicates that the first and last elements in a series make the strongest impressions on a human mind[9]. These kinds of co-occurences, meanwhile, appear not to have been learned by our classifier. The label on each segment, observed alone, is not in fact a very predictive feature. Therefore the classification model assigns only low weights to the prediction on each segment. Because our work focuses largely on overlap ratio, we incorporate segment prediction features only naïvely. Future work could begin by adding two binary features, each indicating whether a high or low prediction, respectively, occupies either the first and last position.

The fact that such a model would involve analyzing and anticipating the judgment of raters does not compromise its usefulness. Indeed, it is difficult to say that there is a crucial difference between intrinsic conflict and the perception of conflict. The latter may be just as important as the former, if not indistinguishable from it. In counseling, customer service, and other online applications, whether an individual *remembers* an experience as contentious is likely to be the most important aspect of conflict detection.

4. Related Work

Other researchers have explored a broad range of techniques to detect disagreement, frustration, and polarity, and many of these techniques can be joined with ours to detect conflict in a different context. Liscombe, Riccardi, and Hakkani-Tür found that the recognition of a caller's frustration can be substantially improved by conditioning on the same judgment made on that caller's previous two speaking turns [10]. In a corpus with sequentially ordered clips, using an effective speaker diarization model, this technique could bolster results.

While the work by Kim[3] and Wrede[2] has laid groundwork for conflict detection in recorded audio, other studies point to features that can be added with the inclusion of more data. Bousmalis, Mehu, and Pantic outlined a series of nonverbal cues, e.g. arm-crossing and head-nodding, that correlate with agreement or disagreement[11], while Charfuelan and Schröder[12] use the SentiWordNet software to determine sentiment in online commentary. The inclusion of such lexical information is likely to improve results, to the extent that there is consistency in user labels. The above work suggests what can be done with data that includes more information, such as a visual signal, or has been preprocessed by effective speaker diarization and speech recognition technology.

5. Conclusions

We have improved upon the baseline experiment results by over 2% (from 80.8% to 83.1%), using solely our predicted ratio of overlapping speech as a feature of the classifier. Our hypothesis, that overlap ratio would boost the performance of the classifier by supplementing the baseline features, was fulfilled and exceeded. In fact, predicted overlap alone was the best feature set we observed, obviating all the others, suggesting that overlap is a reliable proxy for conflict.

These results suggest a number of avenues for future work. There are a variety of sophisticated approaches to identifying overlap in speech. Using more robust speaker diarization and overlap identification tools trained on larger corpora may be able to generate more reliable measure of speaker overlap and therefore conflict.

We have noticed that there are positional effects to conflict. We find that the beginning and ending of stimuli seem to have a greater effect on perceptions of conflict than internal regions. In this work we look at static thirds of each stimulus. Future work will look at a more dynamic identification of regions based on speaker diarization or speech segmentation.

Arguably the most important contribution of our work is to reduce the amount of data needed to detect conflict. High accuracy can be achieved merely by measuring intervals of overlapping speech, or a vocal signal associated therewith, as well as some meta-information about their positioning.

6. Acknowledgements

This work was partially supported by DARPA FA8750-13-2-0041 under the Deep Exploration and Filtering of Text (DEFT) Program.

7. References

[1] 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 *in Proc. Interspeech 2013, ISCA, Lyon, France, 2013*, 2013.

[2] B. Wrede and E. Shriberg, "Spotting "hot spots" in meetings: Human judgments and prosodic cues," in *in Proc. Eurospeech*, 2003, pp. 2805–2808.

[3] S. Kim, S. H. Yella, F. Valente, and D. Lausanne, "Automatic Detection of Conflict Escalation in Spoken Conversation," in *in Proc. Interspeech 2012*, 2013.

[4] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009.

[5] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *ACM Multimedia*, 2010, pp. 1459–1462.

[6] K. Yamamoto, F. Asano, T. Yamada, and N. Kitawaki, "Detection of overlapping speech in meetings using support vector machines and support vector regression," *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. E89-A, no. 8, pp. 2158–2165, Aug. 2006.

[7] A. Quinlan and F. Asano, "Detection of Overlapping Speech in Meeting Records Using the Modified Exponential Fitting Test," in *in Proc. of the 15th European Signal Processing Conference*, no. Eusipco, 2007, pp. 2360–2364.

[8] T. Geiger, R. Vipperla, N. Evans, and G. Rigoll, "Speech overlap detection using convolutive non-negative sparse coding: New improvements and insights," in *in Proc. of the 20th European Signal Processing Conference*, no. Eusipco, 2012, pp. 340–344.

[9] E. A. Feigenbaum and H. A. Simon, "A theory of the serial position effect," *British Journal of Psychology*, vol. 53, no. 3, pp. 307–320, 1962.

[10] J. Liscombe and et al., "Using context to improve emotion detection in spoken dialog systems," in *in Proc. Interspeech 2005*, 2005, pp. 1845–1848.

[11] K. Bousmalis, M. Mehu, and M. Pantic, "Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools," in *The Third International Conference on Affective Computing and Intelligent Interaction*, 2009.

[12] M. Charfuelan and M. Schr, "Correlation analysis of sentiment analysis scores and acoustic features in audiobook narratives," 2002.