Towards Automatic Recognition of Attitudes: Prosodic Analysis of Video Blogs

Noor Alhusna Madzlan[1 3], *JingGuang Han*[1], *Francesca Bonin*[1 2], *Nick Campbell*[1]

[1] CLCS, School of Linguistics, Speech and Communication Sciences, Trinity College Dublin
[2] SCSS, School of Computer Science and Statistics, Trinity College Dublin, Ireland
[3] ELLD, Faculty of Languages and Communication, UPSI, Malaysia

madzlann@tcd.ie, hanf@tcd.ie, boninf@tcd.ie, nick@tcd.ie

Abstract

Understanding of speakers' attitude is essential for establishing successful human interaction. In this paper we analyse attitude manifestations in video blogs. We describe the main features of this novel communication medium and focus our attention on its possible exploitation as a rich source of information for human-human and human-machine communication. We describe the manual annotation of attitudes and the prosodic analyses. Finally we present a preliminary automatic attitude annotation system that attains 65% accuracy.

Index Terms: video blog, prosody, attitude, automatic classification, statistical modeling, SVM.

1. Introduction

Social media is becoming a major form of interaction and of personal expression. While the popularity of content based platforms, such as web blogs, Twitter and Facebook, confirms that written text is still the major form of online interaction, new forms of expression are evolving. Conversational video blogs (in short vlogs) are becoming a widespread phenomenon of online social media, which create a huge amount of user generated content. Video blogs can be defined as personal diaries made available to the larger public in the form of self-recorded videos, where the users express themselves, their personality and share life events. They combine the best qualities of pre-recorded broadcasted speech with the naturalness of spontaneous conversations. However, at the same time, the speaker expresses himself in a de-contextualized situation, addressing an imagined audience without being influenced by the listener's reaction, as would happen in a face to face context.

Video blogs have piqued the interest of scholars in recent years. The main stream of research on vlogs involved the study of personality recognition [1, 2] and they have been widely studied with respect to non-verbal behavior and social attention. However, to our knowledge, the linguistic pragmatic aspect of this new form of expression has been given little attention. Video blogs are a unidirectional form of communication where the user intends to convey a message, an emotion or a personal opinion. In this work we are not interested in investigating personality as reflected in video blogger behavior in front of a camera, but rather the impression of what he wants to transmit at a purely pragmatic level. For this reason, we explore five attitudinal classes that appear to be representative of the videos in our corpus and analyse their prosodic characteristics. In this work we define attitude as social affective states that the video bloggers intend to transmit and we rely on the classification of attitudes in [3]. We are not interested in the inner emotion of the video blogger but in what he intends to express. Finally, we explore the predictive value of extracted prosodic cues for automatic recognition of attitudes in video blogs.
Main contributions of this work are:

- We focus on a new social media, describing the potentialities of this source of material and a qualitative analysis of its characteristics.
- We describe a novel corpus of vlogs and its manual annotation with respect to attitudes.
- We analyse prosodic features of attitude impressions in video blogs.
- We address the task of automatically predicting video bloggers' attitude impressions using multimodal nonverbal cues and machine learning techniques.

The paper is structured as follows: Section 2 describes some of the previous literature on video blogs. Section 3 outlines the characteristics of the dataset. Section 4 explains the annotation process and adaptation of the schema. Section 5 presents selection and analysis of prosodic features. Findings and results are presented in Section 6. Section 7 explains and discusses the analysis in detail. We conclude the study in Section 8.

Figure 1: Example of Video Blogs

2. Related work

Numerous studies with relevance to various fields of research have been conducted on video blog analysis. Biel et.al [1] conducted a study on recognition of the Big 5 Personality Traits

10.21437/SpeechProsody.2014-6

[4] of video bloggers represented through non-verbal signals. They were able to predict personality traits including Extraversion, Openness, Agreeableness, Conscientiousness and Emotional Stability from analysis of non-verbal characteristics in video blogs. Prosodic and visual feature extraction was conducted to identify labels of personality annotated by five people using Amazon Mechanical Turk. Findings suggest that speakers who demonstrate higher pitch range and higher motion activity possess Extroversion and Openness traits.

Other related work involves multimodal sentiment analysis of opinion videos among Spanish speakers on YouTube [5]. Automatic feature extraction and prediction using audio, visual and textual features were conducted to identify sentiments of the speakers. Videos were labelled according to three positive, neutral and negative sentiments. Results showed that the smile offers the best feature prediction, while number of pauses and voice intensity followed suit. Positive sentiments are indicated by increased number of smiles and pauses, whereas negative sentiments are represented by higher voice intensity.

Treating attitudes as expressions of opinions made by the speaker towards related issues during interaction with their interlocutor, Mac et.al [6] investigated audio-visual prosodic attitudes involving cross-cultural relations. Their study examines perceptions of Vietnamese and French participants in identifying attitudinal expressions in the Hanoi standard dialect. A perception test was conducted and results showed that native listeners recognised attitudes better than foreign listeners. For Admiration, however, foreign listeners were able to successfully recognise the state better than native listeners.

Analysis on prosodic characteristics in speech not only facilitates but further enhances understanding of speakers' attitudes towards a particular topic or notion [3]. Henrichsen and Allwood [3] analysed the NOMCO speech corpus containing eight dialogues by automatically extracting multimodal features to identify attitudes. A standard set of ten attitudes called the A10-based annotation was developed. Multimodal features were extracted for automatic prediction of attitude categories and results showed that attitude labels could be predicted by the trained model.

3. Video Blog Dataset

This section introduces the dataset of videos that we use throughout the paper. We first describe the data collection process and then provide a high level analysis of the contents and of the type of speech typical of video blogs.

A total of 100 video blogs were selected from the YouTube channels of four different speakers [1]. The speakers have been selected according to the following characteristics:

- Native English speaker
- American English
- Male speaker
- Aged between 18-25 years old

Among the videos of these speakers, we selected the ones with the larger number of visualizations. On average each speaker is represented by about 25 videos. The videos were downloaded using a free add-on tool for Mozilla Firefox browser [2]. The audio tracks were extracted from each video using the same tool. Annotation and labeling was conducted manually using WaveSurfer [7], as in Fig.3.1. Sample rate for all videos is set at 44100bit rate with mono sound. Total duration of the corpus is 286 minutes, and the mean duration of each video is 2.88 minutes (sd=1.09).

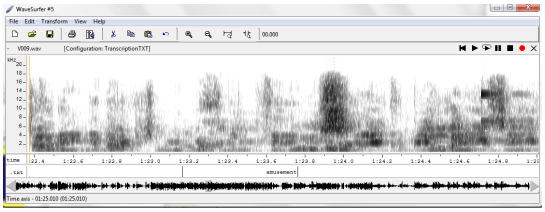

Figure 2: Annotation with Wavesurfer

3.1. Qualitative overview of the dataset

The video blogs in our dataset are typical of the video blog genre. They are pre-recorded monologues, recorded with the speaker facing the camera. They represent an expression of asynchronous communication with a delayed feedback provided by the comments of the audience, and they are usually stored in reverse chronological order. The speech is semi spontaneous. While in broadcast recordings, speech is typically prepared and scripted with a rigid format that the speaker needs to adhere to, in video blogs, the speech is more flexible. Video blogs may be characterised as prepared speech, because a substantial amount of time is given for preparation and planning prior to recording, but video bloggers do not strictly conform to the pre-written texts. As a result of this, utterances in video blogs resemble unprepared or spontaneous speech, particularly with regards to disfluencies, such as ungrammaticality, filled pauses, repetitions, repairs and false starts [8].

These two elements (the preparation of the script and the spontaneity) create a very interesting linguistic genre that preserves features of broadcast speech as well as features of natural spontaneous speech. The following is an example of video blog speech:

```
Hey guys!
If there's one thing I can't stand,
it's people who judge others.
Whether it's based on looks,
or what you heard about them, it
doesn't matter. If you don't know
the person personally, you have
no right to judge them. High
schools are the worst.
```

4. Annotation process

In the present work we are interested in analysing video blogger attitudes during the recording. In the unfolding of video blogs, the user is "playing" different roles and showing different attitudes to the audience. In order to conduct our experiments, we first labelled the corpus with respect to 5 main attitudes shown in Table 1. We base our annotation schema on the A10-set annotation identified in Henrichsen and Allwood [3].

[1] http://www.youtube.com/user/nigahiga
http://www.youtube.com/user/kevjumba
http://www.youtube.com/user/JustinJamesHughes
http://www.youtube.com/user/uncuthashbrown

[2] https://addons.mozilla.org/en-US/firefox/addon/easy-youtube-video-download/

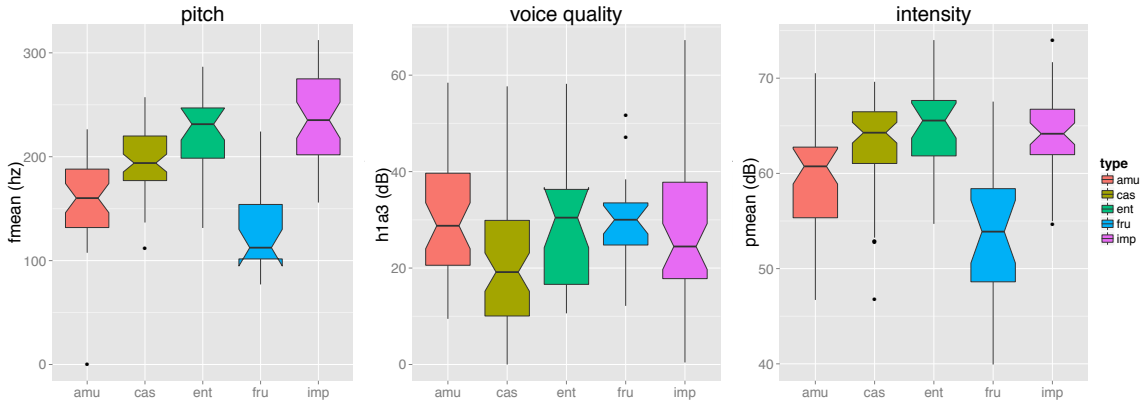

Figure 3: Pitch, voice quality and intensity distributions over the five attitude states (**amu**sement, **cas**ual, **ent**husiasm, **fru**stration, **imp**atience). The figure shows good separation in both Pitch and Intensity with clear differences in Voice Quality for Casual utterances.

The manual annotation took six weeks to complete. We annotated through several stages. A preliminary observation was devoted to understanding attitude expressions of video bloggers. After this qualitative analysis of the corpus, we selected from the A-10 the most representative attitudes for our dataset (Amusement, Impatience, Casual, Enthusiasm, Frustration). In a preliminary phase we listened to the audio files and marked segments of speaker higher activity. Subsequently, two raters were asked to label the segments using the annotation scheme presented in Table 1. Inter annotator agreement was calculated using Cohen's Kappa, and resulted in k=0.75 [9].

After labeling, the chunks' start and end times were extracted using a TCL/TK script.

Attitude	Description
Amusement	speaker laughs, chuckles
Impatience	speaker shouts, appears annoyed, harsh
Casual	speaker informally addresses the audience
Enthusiasm	speaker appears excited
Frustration	speaker appears 'defeated'

Table 1: Attitude Annotation schema

5. Feature extraction

The prosodic analysis was performed after extracting the following acoustic parameters from the video blogs audio channels, using a TCL/TK script:

- Fundamental Frequency (f0) - pitch level (high/low) [max,mean,min,median]
- Pitch target (fpct) - peak position of pitch (rising/falling)
- Voicing (fvcd) - percentage of voicing (vocal fold vibration) within each utterance
- Power/Intensity (dB) - loudness of the voice [max,mean,min,median]
- Power/Intensity movement (ppct) - peak position of power (rising/falling)
- Voice Quality (H1/a3) - tenseness of the voice (creaky/breathy) [h1h2,h1,a3] [10]
- Duration (dn) - length of the utterance (short/long)

In addition to the traditional prosodic parameters; pitch, intensity, and duration of the attitude segments, voice quality is included as a relevant acoustic parameter for communicative speech analysis. This parameter has been shown to have significant correlates with the interlocutor, speaking style and speech act [11].

5.1. Prosodic features analysis

In Table 2, we report the average and standard deviation of the prosodic values for each category. We notice that the attitudinal category Impatience shows the highest pitch and Frustration the lowest. Figure 3 (left and middle) shows the distributions of pitch and voice quality respectively. We observe that Frustration and Amusement differ significantly in terms of pitch from the other categories (lower pitch, $p<0.005$), while Casual significantly differs from the others with respect to voice quality ($p<0.005$). Impatience and Enthusiasm show similar distributions in terms of pitch, with a similar high average pitch. Figure 3 (right) shows the distributions of intensity. Again, Frustration is represented by speech with low intensity (significantly differing from the others, $p<0.005$), while, as expected, Impatience is the attitude characterised by a higher intensity. Significance was tested with a one tailed T-Test (alternative less).

6. Experiments and Results

In this section, we address the task of automatically predicting video bloggers' attitude impressions. Specifically, we were interested in assessing the prediction performances using the selected features for the annotated attitude classes. We will describe the experimental settings and present the results obtained. The study is based on the ground truth data collected (see Section 3) to evaluate classification performance; a user study focusing on the subjective assessment of the quality of the automatically extracted annotations is planned as future work.

We conducted our experiments using the data collection described in Section 4 in a 16 dimensional feature matrix. With the two different features sets as in Table 2, we trained a Support Vector Machine (SVM) with radial basis function (Gaussian) kernel as the classifier for the 5 attitude categories. We used a 10-fold cross validation approach to evaluate the trained model. We evaluated different feature sets:

Feature set ALL: All 16 features described in Table 2.

Type	fmean	fmed	fmax	fmin	fpct	fvcd	pmean	pmed	pmax	pmin	ppct	h1h2	h1a3	h1	a3	dn
Amused	162.97 *(36.09)*	172.24 *(50.94)*	228.04 *(59.84)*	103.01 *(30.43)*	0.43 *(0.25)*	0.55 *(0.19)*	59.86 *(5.75)*	60.51 *(5.88)*	75.37 *(5.38)*	36.71 *(10.43)*	0.49 *(0.27)*	6.49 *(4.74)*	30.00 *(12.17)*	-25.28 *(9.53)*	-55.28 *(5.81)*	1.07 *(0.31)*
Impatient	234.39 *(39.72)*	242.39 *(49.92)*	311.76 *(47.04)*	139.05 *(44.12)*	0.43 *(0.27)*	0.61 *(0.16)*	64.09 *(4.13)*	66.05 *(3.99)*	77.90 *(2.67)*	36.48 *(13.09)*	0.38 *(0.25)*	6.00 *(6.34)*	27.96 *(14.88)*	-26.06 *(11.18)*	-54.03 *(7.41)*	1.26 *(0.50)*
Casual	196.14 *(32.2)*	193.97 *(30.02)*	248.40 *(43.85)*	147.33 *(40.80)*	0.26 *(0.16)*	0.73 *(0.13)*	63.23 *(4.69)*	66.75 *(3.28)*	76.60 *(2.97)*	32.08 *(15.61)*	0.32 *(0.20)*	4.34 *(4.67)*	20.65 *(15.28)*	-34.72 *(17.54)*	-55.37 *(5.53)*	0.55 *(0.16)*
Enthusiastic	220.1 *(38.65)*	242.62 *(40.23)*	310.56 *(47.55)*	115.04 *(50.65)*	0.39 *(0.28)*	0.72 *(0.18)*	64.70 *(4.7)*	66.50 *(4.51)*	79.28 *(3.08)*	35.66 *(14.28)*	0.35 *(0.24)*	6.08 *(6.74)*	29.54 *(13.54)*	-26.72 *(11.87)*	-56.27 *(6.28)*	1.20 *(0.52)*
Frustrated	124.78 *(35.15)*	136.84 *(40.16)*	175.89 *(55.96)*	86.67 *(33.09)*	0.40 *(0.30)*	0.49 *(0.23)*	53.65 *(7.32)*	54.49 *(7.24)*	70.35 *(5.64)*	30.62 *(14.80)*	0.57 *(0.26)*	3.59 *(5.02)*	29.22 *(9.81)*	-27.43 *(9.72)*	-56.64 *(4.51)*	1.27 *(0.57)*

Table 2: Mean values for each attitude category with standard deviation in brackets

Feature set SEL_1: Select only the features that are not highly correlated (with Pearson's correlation coefficient $r < 0.7$): fmean, fmin, fpct, fvcd, pmean, ppct, h1h2, h1a3, h1, a3, dn. A correlation study of the feature set revealed that some of the features are highly correlated (correlation coefficient $r > 0.7$ with $p < 0.01$ in T-test). Only one of the features in the highly correlated feature pairs was selected and a new 11 dimensional feature set: SEL_1 was generated.

Table 3 shows the results of 10-fold cross validation SVMs with the different feature sets.

Feature Set	Accuracy
ALL	61.85
SEL_1	**65.46**

Table 3: Results for the different feature sets.

Results show that the feature set selected after removing the highly correlated features attained the best prediction accuracy. This result has also been compared with other feature set experiments by reducing the threshold of r to 0.65 or increasing to 0.75. SEL_1 resulted to be the feature set showing better prediction performance with a 65.46% accuracy rate.

7. Discussion

Understanding of speakers' attitude is not only essential for establishing successful human-human interaction, but could significantly contribute to the development of a robust system for human-robot interaction. Video blogs, spontaneous and intimate conversations, represent a rich source of natural attitude expressions.

We have reported the prosodic analyses of a collection of video blogs, for a better understanding of the acoustic dynamics behind five different attitudes including Amusement, Enthusiasm, Casual, Impatience and Frustration. Pitch and intensity show a strong discriminative value, and voice quality emerges as a feature characterising Casual friendly talk. We also investigate the predictive performance of prosodic cues and results from the automatic classification experiments show a prediction accuracy of 65.46%.

Although preliminary, results of this work are in line with [3]. In building conversational agents, attitude management (recognition and synthesis) is a key aspect. Given the sensitivity of conversational partners to an inadequate attitude response, attitude recognition is required to be reliable and robust. We believe that the combination of three main prosodic components such as pitch, intensity and voice quality can provide a robust attitude classification. In order to create a natural and spontaneous interaction, a conversational agent should also be able to synthesise the prosodic dynamics representative of the different attitudes. Our study goes in the direction of a better understanding of these dynamics.

8. Conclusion

In this paper we have explored attitude manifestations in video blogs. We have described the main features of this novel communication medium and focused attention on its possible exploitation as a rich source of information in human communication. We have presented a novel corpus of video blogs, its collection and annotation according to five attitudes, and analysed the main prosodic features characterising these classes. Finally we have presented a machine learning approach for the automatic detection of attitudes in video blogs and reported its preliminary results. Future work will be dedicated to extending the corpus and the annotations, and to exploring other feature selection approaches such as Principal Component Analysis.

9. Acknowledgements

This work is supported by the English Language and Literature Department, UPSI, Ministry of Education Malaysia, the Innovation Bursary of Trinity College Dublin, the Speech Communication Lab at TCD, and by the SFI FastNet project 09/IN.1/1263.

10. References

[1] J.-I. Biel, O. Aran, and D. Gatica-Perez, "You are known by how you vlog: Personality impressions and nonverbal behavior in youtube." in *ICWSM*, 2011.

[2] J.-I. Biel and D. Gatica-Perez, "The good, the bad, and the angry: Analyzing crowdsourced impressions of vloggers," *Ethnicity*, vol. 16, no. 4.8, pp. 0–7, 2012.

[3] P. J. Henrichsen and J. Allwood, "Predicting the attitude flow in dialogue based on multi-modal speech cues," *NEALT PROCEEDINGS SERIES*, 2012.

[4] L. R. Goldberg, "An alternative" description of personality": the big-five factor structure." *Journal of personality and social psychology*, vol. 59, no. 6, p. 1216, 1990.

[5] V. Rosas, R. Mihalcea, and L. Morency, "Multimodal sentiment analysis of spanish online videos," 2013.

[6] D.-K. Mac, V. Aubergé, A. Rilliard, and E. Castelli, "Cross-cultural perception of vietnamese audio-visual prosodic attitudes," in *Speech Prosody*, 2010.

[7] K. Sjlander and J. Beskow, "Wavesurfer - an open source speech tool," 2000.

[8] R. Dufour, V. Jousse, Y. Estève, F. Béchet, and G. Linarès, "Spontaneous speech characterization and detection in large audio database," *SPECOM, St. Petersburg*, 2009.

[9] J. Cohen *et al.*, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[10] H. M. Hanson, "Glottal characteristics of female speakers: Acoustic correlates," *The Journal of the Acoustical Society of America*, vol. 101, p. 466, 1997.

[11] N. Campbell and P. Mokhtari, "Voice quality: the 4th prosodic dimension," in *15th ICPhS*, 2003, pp. 2417–2420.