Annotation and Multimodal Perception of Attitudes: A Study on Video Blogs

Noor Alhusna Madzlan[1 3], *Justine Reverdy*[2], *Francesca Bonin* [2],
Loredana Cerrato[2], *Nick Campbell*[2]

[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, reverdyj@tcd.ie, boninf@tcd.ie, cerratol@tcd.ie, nick@tcd.ie

Abstract

We report the set-up and results of an experiment designed to verify to what extent attitudes can be identified and labelled by using an *ad hoc* annotation scheme. Respondents were asked to label the multimodal expressions of attitudes of a number of video bloggers selected from a vlog corpus. This study aims at measuring respondents' attitude choice as well as the difference in their attitude judgments. We investigate the contribution of different modalities to the process of attitude choice (audio, video, all). The results are analysed from three perspectives: inter-annotator agreement, contribution level for each modality and certainty level of attitude choice. Participants showed to perform better in perceiving attitudes when they were presented with the combined audio-visual stimuli in comparison to the audio only and video only stimuli. Participants showed to be more certain in selecting "Friendliness" than the other attitudes.

Index Terms: multimodal perception, video blogs, annotation,affective states.

1. Introduction

Communicative content in human communication involves the expression of social attitudes, defined as social affective states that the speakers intend to transmit to the audience as defined in [1]. Differently from emotions, attitudes might not correspond to the truth inner psychological state of the speaker, but represent what the speaker intentionally wants to show to the outside. Understanding how speakers express their social attitudes is a fundamental step in the process of successful communication both in human-human and human-machine interactions. While many researchers focus on detection of emotion in human-human conversations [2][3][4][5][6], less attention has been given to the analysis of social attitudes.

Nevertheless, understanding how speakers express their attitude by means of different verbal and visual feature is essential to establishing a successful communication and it is particularly useful when it comes to implementing better systems for Human Machine Interactions and Human Robot Interactions, because it can provide the machine with knowledge related to the socio-affective states of the participant.

Understanding of the rich communication content in terms of social signals provides invaluable skills in technologies such as companion systems, socially aware interaction systems, conversational agents. [7].

Previous studies in the field of Linguistics, Social Signal Processing and Affective Computing have highlighted the importance of integrating the information carried out by social signals, in particular emotions, affective states and attitudes in the process of analysing and interpreting the communicative content of interactions [8][9][10]. In this study we consider a specific communicative situation: video blogs (VLOGS) where speakers tend to have a dynamic representation of attitude expression in a specific scenario of social interaction. We focus our attention on how to define and label attitude expressions in a corpus of video blogs selected from Youtube. In order to label attitudes we defined an annotation scheme to annotate the vlog corpus. Our annotation scheme, named N5, is a derivation of the standard A10 attitude annotation proposed by Henrichsen and Allwood [11].

In this paper we present the results of an experiment in which we asked respondents to label multimodal expressions attitude of video bloggers. The aim of this study is to see to what extent attitudes can be identified and labelled by using our *ad hoc* annotation scheme.

2. Related Work

Recent studies explore communicative content, which includes affect and attitudes with its relation to their perceptual meanings [12][13] [14] [15].

Morlec et al. [12] suggest that attitudes strongly reflect in the prosody of the speaker. Their study introduced six attitudes expressed in French from the inter-perceptual-center group (IPCG) melodic curve corpus, which consist of 322 utterances for each of the six attitudes, which are Assertion, Question, Exclamation, Incredulous Question, Suspicious Irony and Evidence. They conducted a perception study among 20 participants to validate the six attitudes using training and testing sentence modules. Results suggest that there exist confusions between Incredulous, Question and Suspicious Irony despite clear prosodic distinctions.

Rilliard et al. [14] conduct a perceptual study of the prosodic characteristics of attitudes (defined as prosodic attitudes) through audio-visual modalities. Extending work on six prosodic attitudes developed by Morlec et al. [12], they included audio-visual recording of the six attitudes from two French speakers and developed a perception test to present different modalities to 32 French listeners. Results show that the Audio-Visual modality prove most helpful for listeners to identify these prosodic attitudes, particularly Obviousness and Suspicious Irony. Despite attaining good recognition rates for each of the 6 attitudes, an interesting approach of analysis is the application of a cluster analysis to understand confusions between these attitudes. Analysis found that Doubt-Incredulity and Surprise-Exclamation are confused in the audio modality, while Question and Doubt-Incredulity are confused when presented in video stimuli. For the audio-video stimuli, video helps in distinguishing Exclamation from Doubt-Incredulity. This work is helpful in providing clear distinctions between the six prosodic attitudes by conducting a perception study and cluster analysis through different modes of stimuli.

Similar to [14], Allwood et al. [16] conducted a perception study on attitude (defined here as Affective-Epistemic States (AES)) using multimodal stimuli. The study involves 12 Swedish participants presented with recordings from the NOMCO First Encounter. Gestures are annotated based on the MUMIN annotation scheme [17]. Participants are shown a two-minute long clip of the corpus and are required to choose any words that describe both affective-epistemic and behavioural states. Results from semantic analysis lead to seven types of AES: happiness, interest, nervousness, confidence, disinterest, thoughtfulness and understanding. Audio-visual modality shows most attributions for nervousness, interest and thoughtfulness. Further analysis suggests that AES expression may be conflicting or complementing according to different modalities. Happiness, for instance, is expressed best through the audio modality but not vividly shown in video modality. This finding claims that multimodal expressions of AES are more complex to perceive.

Findings from these research works suggest that perception studies, through several methods, are typically used to validate the choices of attitudes. With reference to [16], our work elaborates on a similar method of validating our attitude choices through an online perception study.

3. N5 Attitude Categories

Our past work on developing an attitude recognition system [18][19] conducts data annotation using an adaptation of an attitude annotation scheme derived by Henrichsen and Allwood [11]. This annotation scheme consists of ten attitudes named A10, as listed in Table 1.

A10	
Amused	Bored
Casual	Confident
Enthusiastic	Friendly
Impatient	Interested
Thoughtful	Uninterested

Table 1: *Standard A10-based Annotation Scheme*

On the basis of A10, we developed a new annotation scheme, hereafter, N5, constituted by 5 categories, presented in Figure 1. Our hypothesis is that those categories are more representative of the attitudes present in our corpus of video blog. Four of the categories in our N5 annotation scheme are taken from the A10 annotation scheme and the category ”Frustration” is added because it was considered to be appropriate for our vlog corpus.

In order to validate our hypothesis, we asked two Linguist experts to annotate a total of 250 vlogs [19] using the N5 scheme. We then calculated their inter-rater agreement, which resulted to 0.75 Cohen’s Kappa. The reasonably high Kappa shows how the 5 categories are a good representation of the attitude in the corpus. However, in order to have a further validation, we also run a perceptual test involving a group of anonymous non-expert public participants.

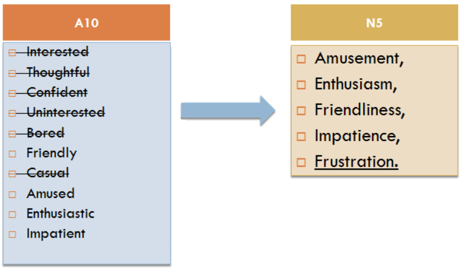


Figure 1: *The N5 attitude categories*

4. Perceptual Test Setup

We designed and run a perceptual test with three different aims:

i.) Validate the choice of the five attitude categories in the N5 annotation scheme
ii.) Investigate which of the modality (audio,video,combined) mostly contributes to the attitude selection task
iii.) Investigate the certainty level of the participants

Twenty participants, recruited among Trinity College Dublin (TCD) staff and students, took part anonymously in the experiment on a voluntary basis. They were requested to provide age and gender information and to read participation information before starting the test. A clearance from the SCSS Research Ethics Committee was obtained previous to the study.

Participants were provided with a link to an online survey and were given 20 minutes to answer all questions. The online survey was developed in-house using PHP5 with an MVC architecture associated with a MySQL database.

The test consisted of three phases. In order to validate the N5 scheme, *i)*, participants were provided with N5 categories and had an additional choice showing the remaining categories from the A10 scheme listed under a drop-down menu with the headings “Others”. The participants presented with the stimuli had to select one of the categories to describe the affective state.

In order to investigate which of the modality (audio,video, combined) mostly contributes in the attitude recognition task, *ii)*, participants had to label a total of 58 stimuli presented in three sections of 18 questions each. Section A consists of the audio only stimuli, Section B comprises video only stimuli (audio muted) and Section C presents both audio-video stimuli.

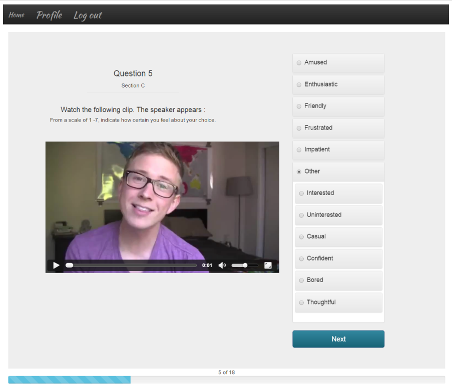
Finally, *iii)* to investigate the certainty level, after selecting an attitude, participants were asked to decide, on a scale ranging from 1 to 7 (going from Unsure to Very Certain), how certain they were about their judgments on their attitude selection. An example of this certainty scale is pictured in Figure 4.

5. Results

We analysed the results from three perspectives: inter-annotator agreement, contribution level for each modality and certainty level of attitude choice.

5.1. Inter-annotator agreement

Results achieved 100% agreement among all the participants for 37% of the stimuli. We further conducted inter-annotator agreement, and found a “fair agreement” between all 20 raters with a k-value of 0.27 using weighted Fleiss Kappa [20]. The



Home Profile Log out

Question 5
Section C

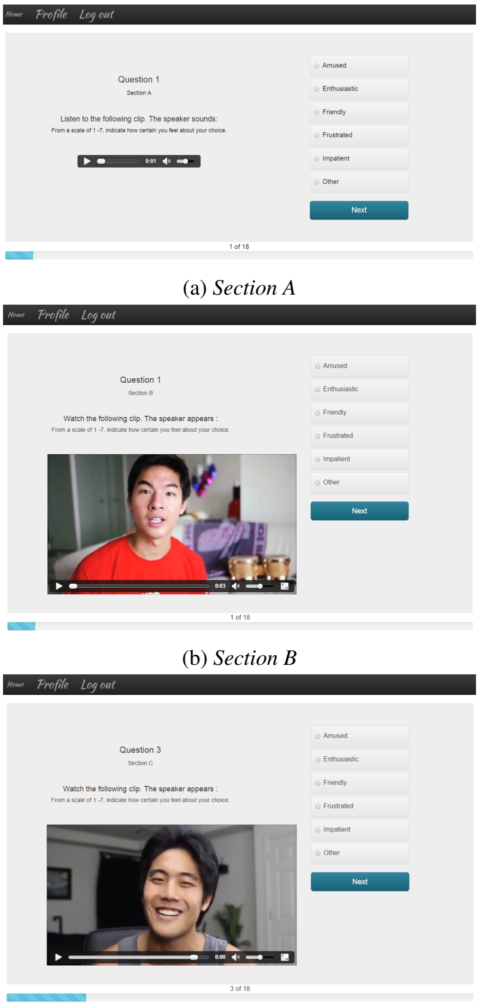
Watch the following clip. The speaker appears :
From a scale of 1 -7, indicate how certain you feel about your choice.

○ Amused
○ Enthusiastic
○ Friendly
○ Frustrated
○ Impatient
● Other
○ Interested
○ Uninterested
○ Casual
○ Confident
○ Bored
○ Thoughtful

Next

5 of 18

Figure 2: *N5 and "Other" attitude choices*



Home Profile Log out

Question 1
Section A

Listen to the following clip. The speaker sounds:
From a scale of 1 -7, indicate how certain you feel about your choice.

○ Amused
○ Enthusiastic
○ Friendly
○ Frustrated
○ Impatient
○ Other

Next

1 of 18

(a) *Section A*

Home Profile Log out

Question 1
Section B

Watch the following clip. The speaker appears :
From a scale of 1 -7, indicate how certain you feel about your choice.

○ Amused
○ Enthusiastic
○ Friendly
○ Frustrated
○ Impatient
○ Other

Next

1 of 18

(b) *Section B*

Home Profile Log out

Question 3
Section C

Watch the following clip. The speaker appears :
From a scale of 1 -7, indicate how certain you feel about your choice.

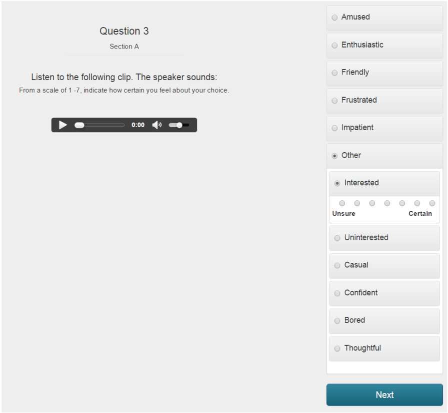
○ Amused
○ Enthusiastic
○ Friendly
○ Frustrated
○ Impatient
○ Other

Next

3 of 18

(c) *Section C*

Figure 3: *Examples of all sections*



Question 3
Section A

Listen to the following clip. The speaker sounds:
From a scale of 1 -7, indicate how certain you feel about your choice.

○ Amused
○ Enthusiastic
○ Friendly
○ Frustrated
○ Impatient
● Other
● Interested
○ ○ ○ ○ ○ ○ ○
Unsure Certain
○ Uninterested
○ Casual
○ Confident
○ Bored
○ Thoughtful

Next

Figure 4: *Example of certainty scale*

low value for agreement is not surprising considering the large number of raters (20 raters) involved in the test. In general it is possible to observe that Frustration is often preferred over other attitudes (see Figure 5).

This observation is in agreement with our justification for the inclusion of the "Frustration" state as an attitude class that is salient in the vlog dataset. Figure 5 shows also that category "Other" did not get enough choices to justify inclusion in our N5.

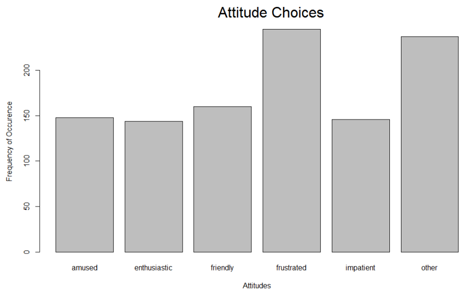


Figure 5: *Frequency of Occurrence of selected Attitudes*

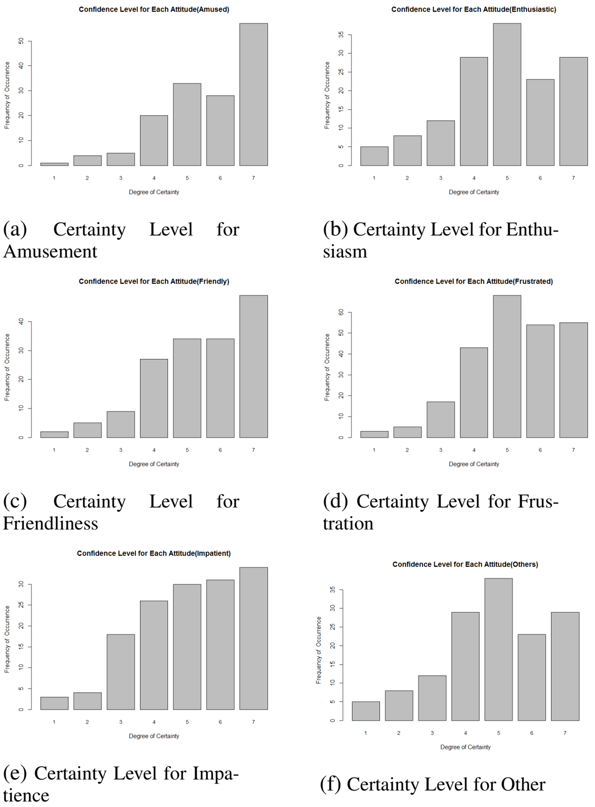
5.2. Contribution of modality in the attitude selection task

We further analysed the relevance of multimodalities for attitude perception and observed that a fusion of audio and visual information is most helpful for participants to perceive attitude expressions of vlog speakers. Specifically, annotators reached a precision of 35.5% when exposed to the Audio+Video stimulus, of the 33.1% while exposed to Audio only and of 31.6% while exposed to video only.

5.3. Certainty level for attitude choice

Following that, we conducted analysis on the certainty level of participants with their attitude choice. Figure 6 shows levels of certainty per attitude.

Participants showed to be most certain when selecting "Impatience" and "Friendliness", while they showed less certainty when selecting the categories listed under "Other".



(a) Certainty Level for Amusement

(b) Certainty Level for Enthusiasm

(c) Certainty Level for Friendliness

(d) Certainty Level for Frustration

(e) Certainty Level for Impatience

(f) Certainty Level for Other

Figure 6: Certainty levels for each attitude

This suggests that participants were not certain and most confused about their choice of the “Other” category.

6. Discussion and Conclusions

The main aim of this study was the validation of our novel attitude annotation scheme N5. To achieve this aim we performed a perception test with 20 participants who were asked to annotate a subset of our vlog corpus.

The low inter-annotator agreement is expected and in-line with Schuller’s [21] statement on the difficulty in obtaining reliability in the annotation of affective states due to the equivocal nature of affect data. Factors like age, gender and cultural backgrounds of participants may contribute to this variation. This finding is not unexpected as it is challenging to assign labels to these kind of phenomena since attitude perception is subjective.

“Frustration” was chosen most of all 11 attitudes, making this a relevant label to annotate attitude in the vlog data. On the other hand, there was not sufficient consistency in the “Other” category to justify inclusion of an extra attitude. This findings suggest that our N5 attitude categories seems to be a sufficient scheme to annotate attitudes in our vlog corpus.

Participants showed to perform better in perceiving attitudes when they were presented with the audio-visual signals in comparison to the audio only and video only stimuli. We found that the fusion of multimodalities from the vlog data is in agreement with Shochi et al. [22], who also report that Audio-Visual modalities have stronger influence in attitude perception.

To further understand which attitude categories are clearly detected and which of the attitudes participants have reservations about, we conducted a certainty test. We notice that participants were more certain in selecting “Friendliness” to the other attitudes. Another observation from this measure of certainty is that the participants showed uncertainty when selecting the attitudes from the drop down menu Others. This is interesting for us as the attitudes included in the “Others are those from the A10 Attitude annotation scheme which we decided not include in the N5 scheme, as we assumed they were not represented in our vlog corpus. This level of uncertainty among participants may be an indication that the attitudes from the“Others list are indeed not so representative of our vlog corpus.

7. Future work

Our work presents a perception study to validate the choice of attitude categories in our vlog dataset. As an extension of this work, the application of this results will be implemented in a predictive classifier in developing a computational framework for automatic attitude recognition. To further improve the current findings, we suggest plausible methods for measuring attitude perception. Due to varying results from multi-rater agreement test, we plan to analyse confusion matrix and/or perform cluster analysis to explain these discrepancies. Future work is also planned for an in depth analysis of gender and age effects to better understand factors that can contribute to attitude perception.

8. Acknowledgements

This work is supported by the English Language and Literature Department, UPSI, Ministry of Education Malaysia, Center for Excellence for Digital Content Technology (ADAPT) at TCD and the Speech Communication Laboratory at TCD.

9. References

[1] Y. Lu, V. Aubergé, A. Rilliard *et al.*, “Do you hear my attitude? prosodic perception of social affects in mandarin,” *Proceedings of Speech Prosody 2012*, pp. 685–688, 2012.

[2] P. Ekman, “Are there basic emotions?” 1992.

[3] A. Kappas and N. Krämer, *Face-to-Face Communication over the Internet: Emotions in a Web of Culture, Language, and Technology*, ser. Studies in Emotion and Social Interaction. Cambridge University Press, 2011. [Online]. Available: http://books.google.ie/books?id=ofM_AHampHsC

[4] J. A. Hall and D. Matsumoto, “Gender differences in judgments of multiple emotions from facial expressions.” *Emotion*, vol. 4, no. 2, p. 201, 2004.

[5] C. Yoo, J. Park, and D. J. MacInnis, “Effects of store characteristics and in-store emotional experiences on store attitude,” *Journal of Business Research*, vol. 42, no. 3, pp. 253–263, 1998.

[6] P. Ekman, W. V. Friesen, M. O’Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, P. E. Ricci-Bitti *et al.*, “Universals and cultural differences in the judgments of facial expressions of emotion.” *Journal of personality and social psychology*, vol. 53, no. 4, p. 712, 1987.

[7] P. Persson, J. Laaksolahti, and P. Lönnqvist, “Understanding socially intelligent agents-a multilayered phenomenon,” *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 31, no. 5, pp. 349–360, 2001.

[8] A. Vinciarelli and G. Mohammadi, “Towards a technology of nonverbal communication: Vocal behavior in social and affective phenomena,” igi-global, Tech. Rep., 2010.

[9] M. D. Pell, "Judging emotion and attitudes from prosody following brain damage," *Progress in brain research*, vol. 156, pp. 303–317, 2006.

[10] D. Ballin, M. Gillies, and I. Crabtree, "A framework for interpersonal attitude and non-verbal communication in improvisational visual media production," 2004.

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

[12] Y. Morlec, G. Bailly, and V. Aubergé, "Generating the prosody of attitudes," in *Intonation: Theory, Models and Applications*, 1997.

[13] J.-M. Blanc and P. F. Dominey, "Identification of prosodic attitudes by a temporal recurrent network," *Cognitive Brain Research*, vol. 17, no. 3, pp. 693–699, 2003.

[14] A. Rilliard, J.-C. Martin, V. Aubergé, T. Shochi *et al.*, "Perception of french audio-visual prosodic attitudes," *Speech Prosody, Campinas, Brasil*, 2008.

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

[16] J. Allwood, S. Lanzini, and E. Ahlsén, "Contributions of different modalities to the attribution of affective-epistemic states," in *Proceedings from the 1st European Symposium on Multimodal Communication University of Malta*, pp. 1–6.

[17] J. Allwood, L. Cerrato, L. Dybkjaer, K. Jokinen, C. Navarretta, and P. Paggio, "The mumin multimodal coding scheme," in *Proc. Workshop on Multimodal Corpora and Annotation*, 2005.

[18] N. A. Madzlan, J. Han, F. Bonin, and N. Campbell, "Towards automatic recognition of attitudes: Prosodic analysis of video blogs," *Speech Prosody, Dublin, Ireland*, pp. 91–94, 2014.

[19] N. Madzlan, J. Han, F. Bonin, and N. Campbell, "Automatic recognition of attitudes in video blogs - prosodic and visual feature analysis," in *INTERSPEECH*, 2014.

[20] J. L. Fleiss, J. Cohen, and B. Everitt, "Large sample standard errors of kappa and weighted kappa." *Psychological Bulletin*, vol. 72, no. 5, p. 323, 1969.

[21] B. Schuller, "Multimodal affect databases: Collection, challenges, and chances," *The Oxford Handbook of Affective Computing*, pp. 323–333, 2014.

[22] T. Shochi, D. Erickson, A. Rilliard, V. Aubergé, J.-C. Martin *et al.*, "Recognition of japanese attitudes in audio-visual speech," in *Speech prosody*, vol. 2008, 2008, pp. 689–692.