Towards Affective State Modeling in Narrative and Conversational Settings

Bart Jochems[1], *Martha Larson*[2], *Roeland Ordelman*[1], *Ronald Poppe*[1], *Khiet P. Truong*[1]

[1] Human Media Interaction Group, University of Twente, Netherlands
[2] Multimedia Information Retrieval Lab, Delft University of Technology, Netherlands

b.e.h.jochems@student.utwente.nl, m.a.larson@tudelft.nl,
{ordelman, poppe, k.p.truong}@ewi.utwente.nl

Abstract

We carry out two studies on affective state modeling for communication settings that involve unilateral intent on the part of one participant (the evoker) to shift the affective state of another participant (the experiencer). The first investigates viewer response in a narrative setting using a corpus of documentaries annotated with viewer-reported narrative peaks. The second investigates affective triggers in a conversational setting using a corpus of recorded interactions, annotated with continuous affective ratings, between a human interlocutor and an emotionally colored agent. In each case, we build a "one-sided" model using indicators derived from the speech of one participant. Our classification experiments confirm the viability of our models and provide insight into useful features.

Index Terms: affect, speech recognition, audio analysis, natural language communication

10.21437/Interspeech.2010-205

1. Introduction

Models of affective states are useful in a number of areas, including frustration detection, stress detection, highlight extraction, multimedia genre classification, and emotionally-enabled conversational interfaces. We carry out two studies on affective state modeling in communication settings that involve unilateral intent. In such settings, one participant, the *evoker*, strives to change the affective state of another participant, the *experiencer*. The goal of these studies is to determine the extent to which "one-sided" models can capture affective patterns in these settings. We use the designation "one-sided" models to refer to models that use the speech signal from either only the evoker or only the experiencer. In each setting, we investigate the ability of a one-sided model to capture peaks in emotional intensity.

A key contribution of our work is that it identifies a novel domain in affective state modeling, namely, communication settings with unilateral intent. This domain is interesting because it involves natural communication, yet the motivations of the speakers are simple and stable, facilitating both the creation and interpretation of models. In this respect, it contrasts with communication settings such as meetings, where a speaker may jump between evoker and experiencer roles and where affective intent changes over time. In the unilateral intent setting, we know which participant is the evoker and the nature of the evoker's goal. This knowledge allows us to assume that the evoker is following a set of strategies designed to attain that goal and the experiencer is reacting to these strategies. A one-sided model will thus capture a stimulus with a well-understood affective purpose or an affective reaction to that stimulus. If models are able to capture the essence of basic affective triggers and responses, they stand to achieve sufficient generality to be easily transferable to new domains. The studies reported in this paper set the first step towards this long-term final goal.

We concentrate on two particular cases: a narrative setting and a conversational setting. For each of these settings we choose a publicly available corpus with affective annotations that encode information about change in affective state. For the narrative setting we use the VideoCLEF 2009[1] *Beeldenstorm* dataset consisting of short-form documentaries annotated with viewer-reported narrative peaks. A narrative peak is a point at which a viewer feels a rise of dramatic tension or a heightened sense of involvement. For the conversational setting, we use the SEMAINE corpus[2] of emotionally colored character interactions consisting of recorded conversations between a human interacting fully naturally and a human playing an agent with a particular emotional style. The corpus is annotated with continuous affective ratings from which we extract ground truth for emotional intensity peaks. Table 1 illustrates how the participants in the corpus scenarios map onto the roles of evoker and experiencer, i.e., the participant roles in a communication setting involving unilateral intent.

Table 1. *Participant roles in communication settings (Participants modeled in our studies are shown in bold.)*

	Narrative setting	*Conversational setting*
Evoker	**Narrator**	Agent
Experiencer	Viewer	**Interlocutor**

In the narrative setting, the narrator's intent is to maintain interest in the content of the documentary by providing moments where viewers feel an intensified sense of involvement. Here, our study involves building an *evoker model* of the narrator that allows us to predict moments at which viewers report experiencing a peak in affective state corresponding to a perceived rise in dramatic tension. In the conversational setting, the agent's intent is to influence the interlocutor's affective state towards a particular emotion (e.g., happy, angry). Here, our study involves building an *experiencer model* of the interlocutor that allows us to detect peaks in the interlocutor's emotional intensity. As shown by the boldface in Table 1, we focus on two "one-sided" models. We must necessarily leave a narrative-setting experiencer model and a conversational-setting evoker model to future work, since our corpora lack either data or annotations to build these models.

The remainder of the paper is organized as follows. First we review related work. Then we give details on the feature sets and the classifiers we use to build our model. We describe each study in turn including experiments and results. Finally, we close with general discussion and outlook on future work.

2. Related work

Affect is generally analyzed as consisting of two components: an arousal dimension representing activeness vs. passiveness

[1] http://www.multimediaeval.org
[2] http://semaine-db.eu/

and a valence dimension representing positivity vs. negativity. With our models, we aim to capture a combination of valence and arousal, e.g., what is referred to by [1] as emotional intensity – the magnitude of the overall emotional reaction. Emotional intensity is characterized in [4] as the simplest description of an emotional state, the deviation from cool and neutral. In this section, we summarize work on how affect is expressed and on modeling affective states.

2.1. Expression of affect

Affect manifests itself both in the way people speak and the words they use. Emotion impacts physiology, which in turn affects speech production [15]. The body's reaction to emotional arousal results in loud, fast, high-energy speech [11]. Language style reflects social and psychological aspects of the world of the speaker [12]. The use of different word types, in particular, function words, emotion words and content words, has been identified as important for revealing psychological effects [12]. Pronouns are function words and one psychological aspect they reflect is speaker social engagement [3][12].

In [2], it is pointed out that in the Western intellectual tradition, emotive uses of languages were originally studied as rhetorical techniques. Speakers make use of rhetorical devices to enhance their impact on their listeners, and emotional speech, the *pathos* of Aristotle, is a key strategy. We assume a broad base of similarity between emotional communication, involving spontaneous outbursts of emotion, and emotive communication, involving the signaling of affect to communication partners as part of a consciously applied strategy, cf. e.g., [1]. For this reason, we expect evoker speech in the narrative context (makes use of rhetorical strategies) and experiencer speech in the conversational context (involves the spontaneous expression of emotion) to have similar characteristics.

2.2. Affective state modeling

Work on automatic detection and classification of affective states encompasses a variety of scenarios. In the meeting domain, affect has been studied in the form of hotspots, regions of increased participant involvement [17]. In the sports domain, affect modeling takes the form of highlight detection, the identification of points at which strong excitement is evoked in the viewer [6][10]. Highlight models resemble our models in that they aim to predict the intensity of the experiencer. In sports events, the experiencer is often present in the video in the form of the audience, simplifying the task [10]. Valence and arousal expressed in the speech of video game players has also been studied [13]. Here, the experiencer is the game player and the evoker is the game and the other game participants. Methods for affective modeling use audio alone (e.g., [17]) or audio and video features (e.g., [10]). Closest to our own work are methods combining acoustic and lexical features, whereby acoustic features are more effective for predicting arousal and lexical features for valence [13].

3. Experimental framework

We build and test two models: the evoker model in a narrative setting predicts peaks of viewer response and the experiencer model in a conversational setting, detects peaks in interlocutor emotional intensity. Recall that our models are one-sided, i.e., trained on a single participant role. One-sided models are useful in domains such as telephony, where privacy reasons might restrict full recordings. They are also well suited for entertainment settings where the reaction of the listener/viewer is minimal or difficult to record. This section discusses our choices of features and classifiers for our models.

3.1. Features

To choose our feature set we look at indicators from the literature and select features that perform well and can be straightforwardly extracted from the (windowed) speech signal.

Acoustic features (Table 2) We use acoustic features corresponding to speech characteristics triggered by the physiology of emotion (cf. Section 2.1). In particular, we chose popular features related to pitch, energy and speech rate [18].

Table 2. *Acoustic features used for the models*

Pitch	mean, minimum, maximum, range, standard deviation, Δ with previous, Δ with next
Intensity	mean, minimum, maximum, range, standard deviation, Δ with previous, Δ with next.
Speed	speech rate, syllables per segment

Pitch and intensity related features were extracted using Praat (in the same manner as in [8]) and speed features were extracted using the speech transcripts and also a Praat script [5].

Lexical features (Table 3) We choose to focus on emotional words and functional words, in particular on pronouns (cf. Section 2.1). We leave content words out of consideration due to issues related to their topic dependence, mentioned in [12].

Table 3. *Lexical features used for the models*

Emotional vocabulary	intensity score (cf. Eq. 1) calculated using affective ratings of words
Functional vocabulary	raw count of first and second person pronouns ("I", "you", "we")

The standard approach to emotional vocabulary, and the one adopted here, is to make use of an affective dictionary. We make use of Whissell's Dictionary of Affect in Language [16], which provides evaluation and activation scores (both in the interval [-1,1]), corresponding to valence and arousal. Recall that we are interested in intensity, which here refers to the overall emotional impact. In order that the affective score reflects intensity, we take arousal into account only in those cases that it is active, i.e., positive. A negative arousal corresponds to passive affect – a lack of involvement or engagement and should not contribute to intensity. We calculate an affective intensity score for each word using Eq. 1.

$$intensity = \sqrt{\left(\frac{arousal + |arousal|}{2}\right)^2 + valence^2} \qquad (1)$$

The overall emotional vocabulary score for a given segment of speech signal is the average affective score for each of the words that it contains.

The overall functional word score for a given segment of speech signal is the raw count of first and second person pronouns it contains. These pronouns reflect social engagement and personal involvement. Intensity score and pronoun count have worked well in our previous work [9], which, in contrast to this work, made use of an unsupervised approach to narrative peak detection.

3.2. Classifiers and classification setup

We chose to carry out our experiments with a Naïve Bayes classifier and a decision tree (J48). The Naïve Bayes classifier estimates a generative model and imposes the assumption that features are independently distributed. The decision tree is a non-parametric classifier that uses features to learn recursive splits of the data. Prediction of intensity peaks was performed by moving a 10-second sliding window over the video stream, advancing the window by one second at each step. The three top scoring peaks were returned for each episode (narrative setting) or interaction session (conversational setting). For our

experiments we used the implementations from Weka.[1] Evaluation was done by leave-one-out cross-validation. In each fold, one episode/session was left out. We measure performance in terms of recall, proportion of ground-truth peaks that are correctly identified. We report performance in terms of recall and precision at three (R@3 and P@3), since the classifier always identifies three peaks.

4. Evoker in a narrative setting

4.1. VideoCLEF 2009 narrative peak corpus

The VideoCLEF 2009 narrative peak corpus contains 45 episodes of a Dutch-language documentary on the visual arts called *Beeldenstorm*. Automatic speech recognition transcripts [7] with word-level time markers accompany the corpus. The documentary host, Henk van Os, is known for his narrative skill, ensuring the sophistication and quality of strategies deployed to hold the viewer's attention.

Three annotators annotated each episode by marking the top three narrative peaks, defined as moments within each episode at which they felt the most extreme emotional reaction. The episodes are short enough (ca. 8 min.) to allow annotators to watch them in their entirety, ensuring that emotional response is reported within its context. The annotators set a start and end point of each peak, under the constraint that a peak can last no longer than 10 seconds.

Here, we are interested in "general" peaks, peaks that viewers will tend to agree on. For this reason, we merge the annotations of all three annotators and use only peaks that were identified as all three annotators as ground-truth peaks. Annotator peaks were merged if they had a two-second overlap. There are 22 general peaks in total in the 45 episodes. Our choice of general peaks means that we have perfect inter-annotator agreement for our ground truth. However, our system performance measured with respect to this ground truth will provide only a conservative estimate of the utility of our classifier. In real world settings, the system might generate peaks that are acceptable to a large number of users, without necessarily satisfying all. Understanding the personal component of peak prediction is a topic we leave to future work.

4.2. Evoker model experimental results

In Table 4, we present the results of the classification experiments for models trained on narrator (evoker) speech from VideoCLEF 2009 narrative peak corpus using classifiers and features described in Section 3.1.

Table 4. *Evoker model in narrative setting: intensity peaks*

	Naïve Bayes		J48	
	P@3	R@3	P@3	R@3
Acoustic	0.01	0.09	0.01	0.09
Lexical	0.03	0.18	0.07	0.41
Acous.+Lex.	0.04	0.23	0.06	0.36

Note the theoretical maxima: P@3=0.16 (22 correct out of 45x3 hypothesized peaks) and R@3=1. Note also a classifier that selects peaks randomly achieves P@3=0.01 and R@3=0.09. Two features sets, lexical and acoustic+lexical, outperform this random baseline, confirming that one-sided models are viable for narrative peak prediction. Further, lexical features outperform acoustic features. Their combination does not, however, yield a clear advantage over lexical features alone. Also, the decision tree outperforms Naïve Bayes.

[1] http://www.cs.waikato.ac.nz/ml/weka/

After the experiments we analyzed a selection of cases manually in order to better understand our results. We found the properties of narrative peaks in the corpus to be highly variable, reflecting a broad palette of creative narrative strategies. For example, peaks can be characterized by either fast speech or slow speech. This sort of diversity offers a possible explanation for the strength of the J48 decision tree classifier, which imposes no assumptions concerning the existence of underlying distributions. We noticed that acoustic properties tend to be diverse and often opposed (i.e., either loud or soft speech characterizes peaks) in a way that lexical properties are not (i.e., high, but not low-intensity emotional words characterize peaks), a possible account for lexical feature superiority.

We chose 10 false alarms for further examination, selecting cases in which the J48 classifier made the same prediction using acoustic and lexical features. None of the cases were implausible as a narrative peak. About half could be characterized as a sort of "narrative plateau", a segment with noticeable heightened intensity, but too long to be a peak. In four cases, these segments were transitions between two topics. In two cases, the classifier picked up the final sentence of the episode, which often adds a new twist or outlook. The ground truth contains several cases of all three annotators choosing a closing remark as one of the top three peaks, supporting the conclusion that these peaks were false alarms with respect to the ground truth, but might not be perceived by users as serious errors. Indeed, both had been chosen by one of the three annotators as a top three peak during the annotation process.

5. Experiencer in a conversational setting

5.1. SEMAINE emotional color interaction corpus

The SEMAINE corpus of emotionally colored character interactions contains recordings of interaction sessions between a human user and an agent, i.e., a human playing a particular Sensitive Artificial Listener (SAL). The corpus was created by the SEMAINE project, which aims to build SALs, multimodal dialogue systems with the social interaction skills needed for sustaining conversation with humans. Each agent (evoker) has a particular emotional agenda and a conversational goal of shifting the user (interlocutor) towards that state. We use 23 sessions from the corpus involving four agents (cf. Table 6) and six users. The sessions are 2-10 minutes in length. Human-generated transcripts with sentence-level time markers accompany the corpus. We estimated word-level time markers by spreading words evenly over the time span of the sentence.

Each session in our dataset is annotated with continuous valence and arousal levels by up to four raters using Feeltrace [4]. Raters use emotional projection and focus on experiencer speech. To identify intensity peaks, we average the continuous annotations of all annotators on segments of experiencer speech. Averaging the traces eliminates potential individual biases and achieves a more general view [4][10]. We then calculate the *intensity* (cf. Eq. 1) every 0.5 seconds using the average arousal and valence values of the continuous annotation. Parallel to the ground truth for the VideoCLEF narrative peak corpus, the three highest maxima within the video are used as the ground truth intensity peaks. The resulting total is 69 ground truth intensity peaks in the 23 interaction sessions.

5.2. Experiencer model experimental results

In Table 5, we present the results of the classification experiments for models trained on interlocutor (experiencer) speech from the SEMAINE corpus using classifiers and features described in Section 3.1. Note the theoretical maxima: P@3=1 and R@3=1. Note also a classifier that selects peaks randomly

achieves P@3=R@3=0.145. Nearly all conditions outperform the random baseline, confirming the viability of the one-sided model. Here, acoustic features outperform lexical features and again we see no clear benefit in the combination. Also note that the Naïve Bayes classifier is superior to the decision tree.

Table 5. *Experiencer model in conversational setting Detected intensity peaks (P@3=R@3)*

	Naïve Bayes	J48
Acoustic	0.29	0.17
Lexical	0.20	0.13
Acoustic+Lexical	0.28	0.17

Again, after the experiments we performed manual analysis. We found peaks to be characterized by conflicting views (e.g., agent says, "You are a doormat" and interlocutor contradicts) and by laughter. Our models were able to find such peaks. We attribute the relatively lower performance of the lexical features to the conversational style of the sessions. If pronouns and emotion words are overall characteristics of conversational style, they are less suited to discriminate individual peaks. Recall that the word-level time markers were estimated and note that lexical features may prove (marginally) more useful if exact time codes are available. Finally, Naïve Bayes performs well, presumably due to the utility of assuming underlying distributions for the features.

We chose 10 false alarms for further examination, selecting cases in which the Naïve Bayes classifier made the same prediction using acoustic and lexical features. Of these cases, seven gave the impression of plausible peaks, e.g., the interlocutor is contradicting/correcting or telling a happy story.

Table 6 presents an agent character breakdown of the performance of the Naïve Bayes classifier.

Table 6. *Detected intensity peaks by agent (P@3=R@3)*

Agent name	*Emotional color*	*Naïve Bayes*
Obadiah	Depressive	0.33
Prudence	Sensible	0.22
Poppy	Happy	0.33
Spike	Angry	0.20

Strategies used by the agents are exaggeration and encouraging the interlocutor to tell stories that induce a certain mood. Reactions to exaggeration are difficult to detect on a lexical level because of topical variation, which serves, in part, to account for low peak detection performance in Spike-sessions. Upbeat stories, however, had characteristic word usage (e.g., "pleasant" and "wonderful"), reflected in the relatively good performance achieved for the Poppy-sessions. We would like to note the existence of peaks where the agent broke role and shared hilarity ensued. These peaks were well detected, but are unrelated to the underlying emotional color of the agent.

6. Conclusion and Outlook

We have presented two studies on affective state modeling, one involving an evoker model in a narrative setting and the other an experience model in a conversational setting. The models that we build are one-sided, in the sense that they contain features extracted from the speech of only one participant role. Our experimental results confirm that models can be trained that outperform the random baseline and demonstrate that both acoustic and lexical features make contributions. Our settings are both examples of communication settings involving unilateral intent, a novel domain for affective state modeling. In the narrative setting, our models were able to capture a range of strategies deployed with the intent to hold audience attention, including dramatic pauses and rhetorical questions. In the communication setting, our models captured strategies intended to shift affective state, including exaggeration and humor. Future work will involve expanding our understanding of unilateral intent settings, especially with respect to the possibility, already mentioned above, of training a model on one domain and using it in another with minimal extra effort.

7. Acknowledgements

The research received funding from EC FP7 NoE PetaMedia (grant agreement n° 216444) and from the SEMAINE project.

8. References

[1] Banse, R. and Scherer, K., "Acoustic profiles in vocal emotion expression", J. Personality Social Psychol, 70:614-636, 1996.

[2] Caffi, C. and Janney, R.W., "Toward a pragmatics of emotive communication", Journal of Pragmatics, 22(3-4):325-373, 1994.

[3] Campbell R.S. and Pennebaker J.W., "The secret life of pronouns: flexibility in writing style and physical health", Psychol Sci. 14(1):60-5, 2003.

[4] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M. and Schroder, M., "Feeltrace: An instrument for recording perceived emotion in real time", Proceedings of the ISCA Workshop on Speech and Emotion, pp.19–24, 325-373.

[5] Jong, N. de and Wempe, T. "Praat script to detect syllable nuclei and measure speech rate automatically", Behavior Research Methods, 41:385, 2009

[6] Hanjalic, A., "Generic Approach to Highlights. Extraction from a Sports Video," Proceedings of the International Conference on Image Processing, 1:2-4, 2003.

[7] Huijbregts, M., Ordelman, R. and de Jong, F., "Annotation of heterogeneous multimedia content using automatic speech recognition", Semantic Multimedia, LNCS 4816, pp. 78-90, 2007.

[8] Huang, Z., Chen, L., and Harper, M., "An open source prosodic feature extraction tool", Proceedings Language Resource and Evaluation Conference, 2006.

[9] Larson, M., Jochems, B., Smits, E. and Ordelman, R., "Exploiting speech recognition transcripts for narrative peak detection in short-form documentaries", Experiments in Multilingual Information Access Evaluation, Vol. 2, LNCS, 2010, to appear.

[10] Liu, C., Huang, Q., Jiang, S., Xing, L., Ye, Q. and Gao, W., "A framework for flexible summarization of racquet sports video using multiple modalities", Computer Vision and Image Understanding, 113(3): 415-424, 2009.

[11] Oudeyer, P.-Y., "The production and recognition of emotions in speech: features and algorithms", International Journal of Human-Computer Studies, 59(1-2):157-183, 2003.

[12] Pennebaker, J.W., Mehl, M.R. and Niederhoffer, K.G., "Psychological aspects of natural language use: our words, our selves", Annual Review of Psychology, 54(1):547-577, 2003.

[13] Truong, K.P. and Raaijmakers, S., "Automatic Recognition of Spontaneous Emotions in Speech Using Acoustic and Lexical Features", Proceedings of 5th Joint Workshop on Machine Learning and Multimodal Interaction, pp. 161-172, 2008.

[14] Valstar, M. F., McKeown, G., Cowie, R. and Pantic, M., "The SEMAINE corpus of emotionally coloured character interactions", Proc. Inter. Conf. Multimedia & Expo 2010, to appear.

[15] Ververidis, D. and Kotropoulos, C., "Emotional speech recognition: Resources, features, and methods", Speech Communication, 48(9):1162-1181, 2006.

[16] Whissell C. and Charuk, K. "A dictionary of affect in language: II. Word inclusion and additional validation", Perceptual and Motor Skills, 61(1):65-66, 1985.

[17] Wrede, B. and Shriberg, E., "Spotting hotspots in meetings: Human judgments and prosodic cues", Proceedings European Confenence on Speech Communication and Technology (Eurospeech), pp. 2805-8, 2003.

[18] Zeng, Z., Pantic, M., Roisman, G.I. and Huang, T.S., "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions", IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(1):39-58, 2009.