UNIVERSITEIT TWENTE.

Human Media Interaction
University of Twente
P.O. Box 217, 7500 AE Enschede
The Netherlands

Detecting Emotional Intensity Peaks in Narrative and Conversational Settings

by

Bart Jochems
b.e.h.jochems@student.utwente.nl
July 2, 2010

Committee

R. Ordelman	K. Truong	R. Poppe	M. Larson	D. Heylen

Preface

Right now, you are reading the thesis about my graduation assignment. During my time at the University of Twente, I found that the field of interaction between humans and computers was the most interesting field in Computer Science. Therefore I was happy to be able to perform my graduation assignment at the faculty of Human Media Interaction.

Before starting with the final assignment, I followed different courses from this faculty. My internship originated in the same faculty. During this internship my job was to make council meetings more accessible using spoken document retrieval.

During my capita selecta research, called "Narrative peak detection in short-form documentaries", I already became familiar with the research area of affective video content analysis. It was a good introduction to this master thesis, which started in November 2009.

I would like to thank my graduation committee, Roeland Ordelman, Khiet Truong, Ronald Poppe, Martha Larson and Dirk Heylen, for their time, work and comments. Furthermore I would like to thank Joe Laufer for his review of my thesis. And last but not least, I thank my parents for giving me the opportunity to do this study.

I hope you enjoy reading this thesis,

Bart Jochems
Enschede
July 2010

Samenvatting

De laatste jaren wordt er vanuit de informatica veel onderzoek gedaan naar het analyseren van affect in videobeelden. Modellen die affect van kijkers kunnen voorspellen zijn namelijk bruikbaar in een groot aantal gebieden. Denk hierbij aan bijvoorbeeld frustratie detectie, stress detectie, hoogtepunten extractie, automatisch genre classificatie en gebruikersinterfaces die gebruik maken van emoties om video's te tonen.

In dit verslag doen we onderzoek naar het automatisch detecteren van pieken in emotionele intensiteit. In dit onderzoek kijken we naar video content waarbij een deelnemer (de *evoker*) de affectieve toestand van een andere deelnemer (de *experiencer*) beïnvloedt. Hierbij concentreren we ons op twee gebieden, namelijk verhalen en conversaties.

Om deze pieken in emotionele intensiteit te detecteren wordt er eerst een literatuuronderzoek uitgevoerd waarbij wordt gekeken hoe andere studies affectieve patronen herkennen en hoe deze patronen zich voordoen. In dit literatuuronderzoek ligt de focus vooral op sport video's en films. Aan de hand van dit literatuuronderzoek worden de features, denk hierbij aan spraak snelheid, toonhoogte enz., bepaald die gebruikt worden om pieken in emotionele intensiteit te herkennen.

Nadat de feature set is bepaald gaan we evalueren hoe goed onze modellen presteren in het herkennen van deze pieken. Voor deze evaluatie gebruiken we twee datasets. Voor de verhalende setting gebruiken we VideoCLEF 2009. In de verhalende setting probeert de *evoker* de *experiencer* geïnteresseerd te houden door deze een gevoel van betrokkenheid te geven. Voor de conversationele setting gebruiken we de SEMAINE dataset. In deze dataset probeert de *evoker* de affectieve toestand van de *experiencer* richting een bepaalde emotie (bijvoorbeeld blij of angstig) te sturen. Uit de resultaten van deze evaluaties blijkt dat het automatisch detecteren van emotionele intensiteit daadwerkelijk mogelijk is.

Tot slot doen we een aantal voorstellen tot mogelijke implementaties voor een emotionele intensiteit browser. Deze browser moet mensen helpen met het vinden van emotionele momenten in video content. Met behulp van zo'n browser zouden mensen in staat moeten zijn om sneller door de content heen te kunnen navigeren, doordat mogelijk interessante stukken direct aan de gebruiker worden voorgelegd.

Abstract

Over the past couple of years a lot of studies have analyzed the affective level in video content. 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.

In this thesis we carry out a study on the automatic detection of peaks in emotional intensity in a conversation and a narrative setting. Both settings involve unilateral intent on the part of one participant (the evoker) to shift the affective state of another participant (the experiencer).

To detect peaks in emotional intensity a literature study is carried out how affective patterns emerge in video content and how these patterns can be detected. This study focuses on sport videos and movies. Based on this, assessment features (e.g. speech rate, pitch, etc.) are identified that can be used to detect peaks in emotional intensity.

After the feature set is defined we evaluate our models based on their peak detection performance. For this evaluation we used two publicly available datasets with affective annotations that encode information about change in affective state. For the narrative setting we use VideoCLEF 2009. Here, the evoker's intent is to maintain the interest in the video by providing moments where viewers feel an intensified sense of involvement. For the conversational setting we use the SEMAINE corpus of emotionally colored character interactions. Here, the evoker has a particular emotional agenda and a conversational goal of shifting the experiencer toward that state. Results of our evaluation of the classification experiments confirm the viability of the models and provide insight into useful features.

Finally, we prototype a number of user interfaces, which utilize the emotional intensity information detected by our models. A peak browser can help users navigate through the emotional moments in video content. Assisted by this browser, users should be able to navigate faster through the content, because possibly interesting parts of the video can be shown directly to the user.

Table of Contents

1 Introduction

"What if you could remember everything? Soon, if you choose, you will be able to conveniently and affordably record your whole life in minute detail. You would have Total Recall."

In September 2009, the book "Total Recall" was published which was written by two Microsoft researchers, who described an exciting future where people could record and play back every moment of their lives. Today, this future is quickly becoming more reality than fiction.

Now imagine if this was truly possible today, and you are looking for a particular memory captured on video 10 years ago during your university graduation ceremony. Perhaps the date is memorable and if not it can be easily looked up but what if it were also possible to search through almost a lifetime of recordings using emotional states such as joy, relief, excitement as intangible reference points. In fact, what if all the recorded video was annotated with emotional experiences? One thing is for sure: multimedia retrieval would not also be easier but would also incorporate new nuances beyond those existing today.

In this thesis we set the first step towards this long-term goal by developing algorithms to automatically detect emotional intensity peaks in video content.

1.1 Peak Detection

With peak detection we refer the automatic identification of heightened emotional intensity, specifically what Banse and Scherer [1] define as the magnitude of the overall emotional reaction. We carry out two studies on peak detection 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 peaks 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, either with the evoker or the experiencer.

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.

1.2 Narrative and Conversational Settings

In the literature on affective analysis of video, two types of content have received particular attention: sports games and movies [2]. These two cases differ with respect to the source of the emotional intensity. In the case of sports, emotional intensity peaks arise as a result of the unpredictable interactions of the players within the rules and physical constraints of the game. In the case of movies, emotional intensity is carefully controlled by a team including scriptwriters, performers, special effect experts, directors and producers. The difference between the two cases is the amount and nature of human intention – i.e., premeditation, planning, intervention – involved in the creation of the sequence of events that plays out over time (and space). We are interested in investigating a two other settings of video content, namely a narrative and a conversational setting, in which both evokers intent to influence the interlocutor's affective state.

Our work differs in an important respect from previous work in the domains of sports and movies. In the both the narrative and conversational setting the emotional intensity is never completely spontaneous; the evoker has a goal he must comply. However, the emotional moments are characteristically less tightly controlled than it would be in a movie. In a movie, the entire content is subordinated to the plot, whereas in the narrative and conversational setting, the evoker has some freedom and therefore may follow one or more story lines as long as it simultaneously pursues the goal of shifting the interlocutor's affective state. Because of these differences, we chose to dedicate separate and specific attention to the affective analysis in narrative and conversational settings; and in particular to the automatic detection of emotional intensity peaks.

For each of these settings we choose a publicly available corpus with affective annotations that encode information about change in emotional intensity. For the narrative setting we use the VideoCLEF 2009 *Beeldenstorm* dataset consisting of short-forum 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 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 valence and arousal levels. 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.

1.3 Research Questions

As stated, this research concerns automatic peak detection in narrative and conversational settings, with one-sided models. The following questions are to be answered:

1. *Is it possible with a one-sided model to capture peaks in emotional intensity in -*
   a. *a conversational setting*
   b. *a narrative setting*
2. *How do lexical and acoustic features contribute to peak detection in these settings?*
3. *Is there a correlation between the emotion the evoker tries to shift the experiencer to and the peak detection performance?*
4. *How can we present these peaks in a useful manner to the end user?*

2 State of Art Peak Detection

Peak detection approaches aim to detect emotional intensity in video content. Peaks generally stand for the most interesting parts of a video, although the definition of what is interesting may vary widely across video genres and for different applications. For example, in sport-videos the peaks usually show specific patterns related to ball or player movement while in movies the peaks are carefully crafted into the content by a team including scriptwriters, performers, special effects experts, directors and producers. What peaks have in common in these different domains is that the viewer perceives an increase of the level of emotional intensity within the narrative flow of a video.

Once these peaks are detected they can be used in several applications. Summaries can be automatically generated from these highlights, allowing for faster browsing through relevant sections. This will save valuable time for any viewer who merely wants to see an overview of the clip. Highlights can also help with the retrieval of video clips. Peak detection approaches make it possible to generate an index describing the video content, which can be used for browsing, searching and manipulating video documents. It forms the basis for multimedia retrieval in digital libraries storing multimedia data.

Video content is a very broad domain, containing all available video content. It ranges from surveillance cams and home videos to movies and television programs. Because peaks are different in each content type, we focus on television programs and sport videos. In both the sports and movies domain, affect modeling takes the form of highlight detection, the identification of points at which string excitement is evoked in the viewer [2, 3]. Highlight models resemble our models in that they aim to predict the intensity of the experiencer.

This chapter provides an overview of highlight detection in video content. The chapter starts with a deeper look at the affective level of video content. Then we define the two domains, television programs and sports video. In paragraph 2.2 the different modalities within video content are explained. Paragraph 2.3 presents an overview of previous work, categorized by modality. Finally this chapter concludes with an overview of all cited highlight detection approaches listed in a table.

2.1 Affective Level of Video Content

In video content there are two basic levels of perception: the cognitive level and the affective level. The cognitive level describes the facts, like the structure of the story, the composition of a scene, and the object and people captured by the camera [2]. The affective content of a video is defined as the amount and type of affect (feeling or emotion) that are contained in video and expected to arise in users while watching that video. Recall from the introduction that peak detection involves identifying sections that evoke increased levels of emotional intensity in viewers' perception of video clips [4]. Because the affective content of a video contains information what emotions we study this level into more detail.

The affective level has three basic underlying dimensions [5]

- Valence
- Arousal
- Control (Dominance)

Valence is typically characterized as a continuous range of affective responses or states extending from pleasant or "positive" to unpleasant or "negative" [6]. Arousal is characterized on a continuous scale from excited to calm. In other words, valence is the type of emotion while arousal stands for the intensity of the emotion. Note that this is not the same as emotional intensity, which is the magnitude of the overall emotional reaction. Finally, control ranges from "no control" to "full control" and is useful for distinguishing emotional states having similar arousal and valence, for example fear and anger. These three dimensions are the basis for the entire scope of human emotions. However, valence and arousal also account for most of the indecent variance unemotional response [7]. For this reason, the control dimension will be ignored and only the arousal and valence dimensions are considered. These two scales can be transformed to a two-dimensional emotion space depicted in Figure 1 [8].

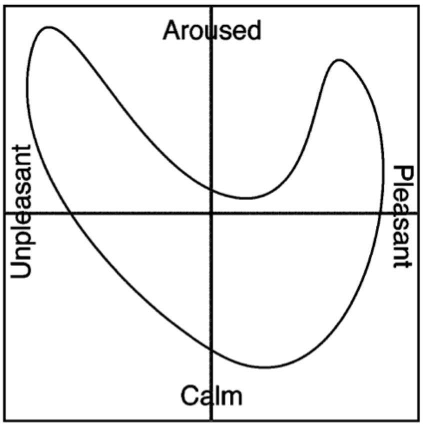


Figure 1: Illustration of the 2-D Emotion Space [8]

Figure 1 also visualizes the emotional intensity peaks, which are located at the upper left, in the aroused and unpleasant area, and in the upper right, in the aroused and pleasant area. Emotional intensity peaks are identified with a high arousal emotion combined with either a very pleasant or unpleasant emotion.

2.1.1 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 [9]. The body's reaction to emotional arousal results in loud, fast, high-energy speech [10]. Language style reflects social and psychological aspect of the world of the speaker [11]. The use of different word types, in particular, function words, emotion words and content words, has been identified as important for revealing psychological effects [11]. Pronouns are function words and one psychological aspect they reflect is speaker social engagement [11, 12].

Caffi and Janney [13] 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].

2.2 Domains

Within the television domain most research has focused on movies and sports. In these domains, affect modeling takes the form of highlight detection, the identification of points at which strong excitement is evoked in the viewer. Although, in sports, we cannot speak of unilateral intent on part of the evoker to shift the affective state of the experiencer, viewers show strong excitement when watching sport videos.

Most highlight detection approaches only focus on one of these sub domains. Each sub domain has it own triggers – or cues – and therefore it is hard to build a highlight detection system that works in all sub domains. In the following paragraphs an overview is given of the sub domains along with the main triggers used for highlight detection.

2.2.1 Sports

In the past few years the sports domain has gained the most attention. One of the reasons is the clear conventional triggers present in this domain - a broad spectrum of viewers will agree about the highlights in the sport domain. Think of goals scored in soccer games or crash incidents in a motor race. In sport events, the experiencer is often present in the video in the form of the audience. Most research in the sports domain focus only on one particular sport (basketball, baseball, etc.), since triggers are different for each kind of sport. Examples of more generic approaches for highlight detection [14] and [3] aim to detect highlights in sports that share common rules and/or characterizations.

Highlight detection approaches in the sports domain often isolate triggers such as applause, cheering, commentary speech, camera motion, colors and the scoreboard.

2.2.2 Movies

Highlight detection approaches in the movies domain can be broadly split into scene-based summaries and event-based summaries [15]. Scene-based summaries focus on obtaining an index of a movie by splitting it into key scenes. Examples of scene-based approaches are [16] and [17]. Event-based approaches aim to detect shots in the movie that belong to a certain event type. For example, [18] detects violent events in a movie by searching for visual cues such as flames or blood pixels, or audio cues such as explosions or screaming.

Another research focus is on automatic generation of movie trailers, or previews, which are film advertisements for feature films that will be exhibited in the future at a cinema, on whose screen they are shown. Trailers tend to feature the high points of the movie, which are edited together in such a way that they do not give away the storyline or conclusion, and yet act as a teaser to their audience. Trailers themselves can be quite cinematic with their own background music, sophisticated shot transition, and post-produced features such as overlaid text. Therefore, movie trailers have a very creative and artistic aspect and the highlight detection is more guidance for the director of the trailer then a real automatic trailer generator [15].

Just like the sports domain, movies contain conventional triggers such as a romantic kiss or a gunfight. To detect highlights in movies highlight detection approach looks for the following cues: boundary shots, camera movement, loudness and affect.

2.3 Modalities

The following three information channels or modalities are considered within a video document:

- *Auditory modality*; contains speech, music and environmental sounds that can be heard in a video document

- *Textual modality*; contains textual resources that describe the content of the video document (i.e., speech transcripts).

- *Visual modality*; contains everything, either naturally or artificially created, that can be seen in the video document.

Each modality has its own features for selecting highlights. Typical features for the auditory modality are: pitch, volume and intonation. For the textual modality the lexical items and meta-data are typical features. Finally, typical features for the visual modality include, shot boundaries, color and camera movement.

2.3.1 Modality Fusion

Highlight detection approaches are not bound to one modality. Approaches that extract peaks using only one of the auditory, textual or visual modalities are called unimodal highlight detectors while approaches that combine two or more modalities are called multimodal peak detectors.

Combining two or more modalities is challenging since it has to deal with indications obtained from different modalities, which might contradict each other. At present, there is enough experimental evidence to state that video content analysis yields the most effective index when a multimodal approach is adhered [19-21]. Additional modalities may serve as a verification method, a method compensating for inaccuracies, or as an additional information source [22]. As Cheng and Hsu [23] state: low-level visual features have their limitations to express high-level semantic meanings of scenes while audio signals can generally provide more semantic information, such as cheering of audience. On the other hand, the noise prevailing in audio signals is comparably high whereas motion information is more feasible against environmental noises. Thus, the combination of both features complements each other and improves the reliability of the highlight extraction.

In the literature two general strategies for the fusion of modalities in video content analysis have been identified, namely early fusion [24] and late fusion or decision-level fusion [20, 25]. These differ in the way results are integrated from feature extraction on the various modalities.

Early Fusion

Early fusion approaches first extract the features from the modalities, which are then combined into a single representation. Based on this representation highlight detection algorithms can assign scores to segments. Early fusion yields a truly multimedia feature representation, since the features are integrated from the start [26]. A schematic overview of the early fusion strategy is shown in Figure 2.

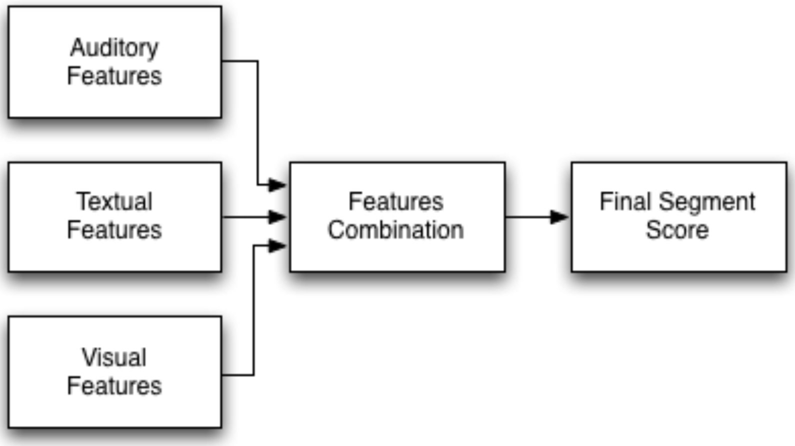


Figure 2: Early fusion, adapted from [26]

Late Fusion

Highlight detection approaches that rely on late fusion also start with extraction of features. In contrast to early fusion, where features are then combined into a multimodal representation, approaches for late fusion are first combined in a unimodal representation. After analyzing these unimodal representations the scores of this analysis are combined into a final segment score. A schematic overview of the early fusion strategy is shown in Figure 3.

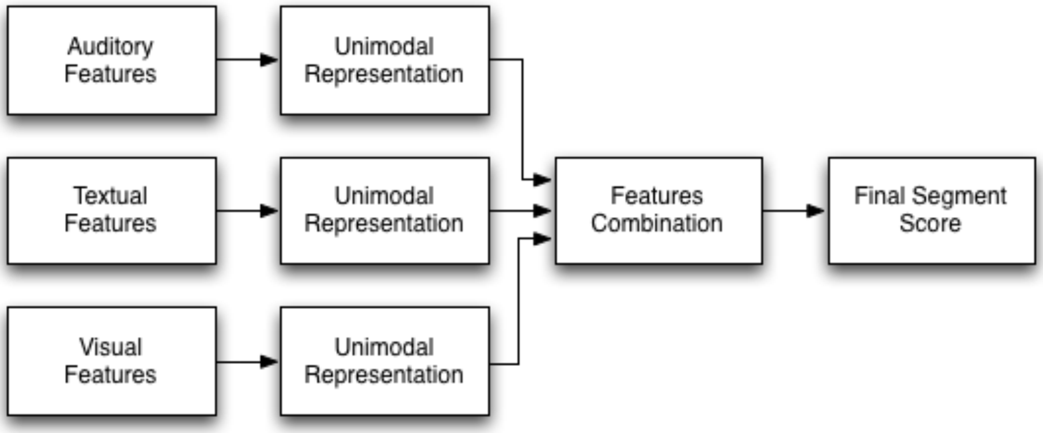


Figure 3: Late fusion, adapted from [26]

Usually, performing late fusion is chosen over performing early fusion for two primary reasons [27]. First, it is difficult to combine features into a common representation. The second reason is that late integration provides greater flexibility in modeling.

2.4 Previous Studies

In the following section an overview of previous studies is given. The studies are ordered by their modality and if they are using more than one modality they are listed in section 2.4.3, multimodal detectors.

2.4.1 Auditory Information

Sports

Rui et al. [28] use only announcers' speech and game specific sounds from the audio track to detect highlights in baseball programs. To detect these sounds in the audio track, the following audio features are used: short-time energy, phoneme-level analysis and prosodic features.

Xiong et al. [29] detect highlights in soccer, golf and baseball by using the MPEG-7 audio features and entropic prior hidden Markov Models to detect common audio events that are directly indicative of highlights. The audio features include frequency and intensity. The audio signal is divided into overlapping frames of 30 ms duration with 10ms overlapping for a pair of consecutive frames.

Movies

In the movie domain affect is often used as a feature, where events are detected by looking at the data from an affective point of view [30]. Arousal and valence are measured and these are combined in the affect curve. The arousal is modeled based on the energy of the audio signal, while the valence is modeled based on the pitch. Variations in pitch based on gender or audio source type are ignored. Users are presented with an affect-based retrieval system, showing graphs of the affective curve for the video content.

Cai et al. [31] locate highlights by considering the following audio cues: laughter, applause and cheer. These cues are detected by using short-time energy (amplitude variation), average zero-crossing rate (frequency) and sub-band energies (frequency spectrum). They propose a framework that can also be used in videos dealing with scenarios taken from sports, meetings and home settings. Hidden Markov models are chosen to model the sound effects.

2.4.2 Visual Information

Sports

In the sports domain there are not many approaches that use only information from the visual modality. Most research includes audio features because cheering and applause cues are a good indication for highlights. However, Assfalg et al. [32] have detected penalties, free kicks and corner kicks in soccer matches using only visual features by formulating a strong correlation between ball movement and camera action and therefore using camera movement as a main feature. A second feature is the position of the players in the field. To detect these positions three zones are defined for each of the sides of the field. Hidden Markov models are used for the detection and classification of the penalties, free kicks and corner kicks.

Lazarescu et al. [33] extract highlights from cricket matches using camera parameters (pan, tilt, zoom and roll). These parameters are converted from numeric to symbolic form and extract four main features for each cricket shot: the dominant motion in the shot, the average camera motion, the length of the shot and the angle of main camera movement in the play. To detect events an incremental learning algorithm is applied.

2.4.3 Multimodal Information

Most research combine features from two or more modalities in order to realize a better highlight detection performance. Approaches that combine two or more modalities are called multimodal detectors.

Sports

Nepal et al. [34] detect goals and other highlights from basketball games by relying on three features: crowd cheer, scoreboard display and change in direction. They use temporal models to classify the event. For example, after scoring the crowd cheers, the scoreboard display changes and the players move in the opposite direction.

There are also approaches that follow a more generic method for detecting highlights [14]. For example, looking for visual and audio features to detect highlights in all field sports, such as soccer, rugby, American football and hockey. Features included increased audio activity, cheering/applause detection, close-up detection and scoreboard activity. A support vector machine was used to combine these features.

Lui et al. [3] build a framework for detecting highlights in racquet sports, (e.g. tennis, table tennis etc), which again utilizes features from both the auditory and visual modality to detect events such as rallies, cheering, applauding and serving. A temporal voting strategy and highlight ranking was used to work properly on different racquet sports.

There have been similar approaches in other sports such as American football [35], Formula 1 car racing [36], tennis [37], baseball [38] and soccer [39].

Movies

Chen et al. [40] propose an action movie segmentation and summarization framework based on movie tempo, representing the delivery speed of important segments of a movie. In the tempo-based system, features from the auditory and visual modality are combined. Features include: shot change detection, motion activity analysis, and semantic context detection based on audio features to grasp the concept of tempo for story unit extraction.

Finally, Smeaton et al. [15] present an approach which automatically selects shots from action movies in order to assist in the creation of trailers. A set of audio and visual features are extracted that aim to model the characteristics of shots typically present in trailers, and a support vector machine is utilized in order to select the relevant shots.

An overview of the mentioned approaches is shown in Table 2.

Table 2: Overview of highlight detection approaches

	Modality			Domain
	Auditory	Textual	Visual	
Hanjalic and Xu [2]	✓			Video content in general
Lui, et al. [3]	✓		✓	Sports; Racquet
Sadlier and O'Conner [14]	✓		✓	Sports; field sports
Smeaton, et al. [15]	✓		✓	Movies
Rui, et al. [28]	✓			Sports; Baseball
Xiong et al. [29]	✓			Sports; Baseball, Gold, Soccer
Chan and Jones [30]	✓			Movies
Cai et al. [31]	✓			TV Shows
Assfalg et al. [32]			✓	Sports; Soccer
Lazarescu et al. [33]			✓	Sports; Cricket
Nepal et al. [34]	✓		✓	Sports; Basketball
Li and Sezan [35]	✓		✓	Sports; American Football
Petkovic et al. [36]	✓	✓	✓	Sports; formula 1
Kijak et al. [37]	✓		✓	Sports; Tennis
Gong et al. [38]	✓		✓	Sports; Baseball
Cabasson and Divakaran [39]	✓		✓	Sports; Soccer
Chen et al. [40]	✓		✓	Movies (sports)
Hsu [41]	✓			Sports: baseball, golf and soccer

3 Datasets Description

In this chapter both the VideoCLEF and SEMAINE dataset are described. For each dataset we give a short introduction, on why it was developed. We then describe the dataset and the ground truth. We conclude this chapter by listing the main differences between the two sets.

3.1 VideoCLEF

VideoCLEF is a track of the CLEF benchmark campaign dedicated to developing and evaluating tasks involving access to video content in a multilingual environment. In 2009, there were three tasks. The first task, called "Subject Classification", involved automatic tagging of videos with subject theme labels (e.g., 'Music', 'History'). The second task, called "Affect", involved detecting narrative peaks in short-form documentaries. A narrative peak is a point in the narrative flow of a video in which viewers perceive an increase in dramatic tension or a heightened sense of involvement. The affect task is what we focus on in this research. The final task, called "Finding Related Resources Across Languages", involved linking video to material on the same subject in a different language.

3.1.1 VideoCLEF Dataset

The VideoCLEF dataset consists of 45 episodes of the Dutch TV series *Beeldenstorm* (in English, 'Iconoclasm'). The series features topics in the visual arts, and integrates elements from history, culture and current events. *Beeldenstorm* is hosted by Prof. van Os, who is not only known for his art expertise, but also for his narrative ability. Prof. van Os is highly acclaimed and appreciated in the Netherlands, where he has established a reputation of appealing to a broad audience. All the *Beeldenstorm* episodes are in Dutch, and consist of video and audio. The length of the episodes varies between seven and nine minutes. Speech transcripts are available and are generated by SHoUT[1] (Spraak Herkennings onderzoek Universiteit Twente). The transcripts are aligned on word level and are not manually corrected.

Constraining the corpus to contain episodes from *Beeldenstorm* limits the spoken content to a single speaker speaking within the style of a single documentary series. This limitation is imposed in order to help control effects that could be introduced by variability in style or skill. Experimentation of the ability of algorithms to transfer performance to other domains is planned for future years.

[1] http://wwwhome.cs.utwente.nl/~huijbreg/shout/index.html

An additional advantage of using the *Beeldenstorm* series is that the episodes are relatively short, approximately eight minutes in length. Because they are short, the assessors who create the ground truth for the test collection are able to watch each video in its entirety. In short, the *Beeldenstorm* program provides a highly suitable corpus for developing and evaluating algorithms for narrative peak detection.

The dataset is divided into a training set and a test set. The training set contains five *Beeldenstorm* episodes in which a human assessor had identified example peaks. The test set contained 45 videos and was mutually exclusive with the training set. In the test set three human assessors had identified the top three peaks in each episode.

3.1.2 VideoCLEF Ground Truth

For the purposes of evaluation three Dutch speakers annotated the *Beeldenstorm* collection by identifying each of the three top peaks in each episode. Annotators were asked to mark the peaks where they felt the dramatic tension reached its highest level. They were not supplied with an explicit definition of a peak. Instead, all annotators needed to form independent opinions of where they perceived peaks. In order to make the task less abstract, they were supplied with the information that the *Beeldenstorm* series is associated with humorous and moving moments. They were told that they could use that information to formulate their notion of what constitutes a peak. Peaks were required to be a maximum of ten seconds.

In total the assessors identified 293 distinct narrative peaks in the 45 episodes. Peaks identified by different assessors were considered to be the same peak if they overlapped by at least two seconds. This value was set on the basis of observations by the assessor on characteristic distances between peaks. Overlapping peaks were merged by fitting the overlapped region with a ten second window. This process was applied so that merged peaks would never exceed the specified peak length of ten seconds. The start time of the merged peak is based upon the average start time of the overlapping peaks.

The average peak length is 9.4 seconds, based on the 405 peaks the assessors identified. Of all peaks, 316 peaks have a length longer than nine seconds. The length of only three is shorter than four seconds.

3.1.3 Remarks on the Ground Truth

General Narrative Peaks

It is difficult to define a truly narrative peak by all viewers since the dramatic tension of viewers is based on 1) personal experience, 2) cultural and background differences and 3) context and memory. As [42] state about affect:

people show their emotion according to a specific pattern defined by both their own experiences and their social environment. In addition, the evaluator, as a receiver of the emotion message, also perceives the expressed emotion according to his or her own background. Furthermore, people might be emotionally affected during an evolution. Therefore it is difficult to distinguish personal narrative peaks from general narrative peaks; peaks on which a large group of people agree on. In order to annotate the general dramatic tension, a large number of participants should annotate the *Beeldenstorm* episodes. Considering that the ground truth was created based on the annotations of only three assessors it is clear that it does not hold the general dramatic tension.

Inter-rater Agreement & Evaluation

The inter-rater agreement cannot be calculated due the ground way was created. Assessors were free to mark three peaks within an episode. Because peaks identified by different assessors are considered to be the same if they overlapped by at least two seconds, this can lead to the situation were A, B and B, C are the same peak but A and C are not the same peak. To illustrate this problem consider the following example: assessor A annotates a peak at 0:34 till 0:44 seconds, B annotates a peak at 0:40 till 0:50 and C annotates a peak at 0:48 till 0:58, which is shown in Figure 4. Since the peak of annotator A overlaps four seconds with the peak of annotator B this peak is considered to the same peak. Also, the peaks of annotator B and C overlap by at least two seconds, so this is too the same peak. However, the peak of annotator A and C are not the same since they do not overlap at all. This makes it impossible to calculate the inter-rater agreement and the kappa coefficient for participants of the VideoCLEF affect task.

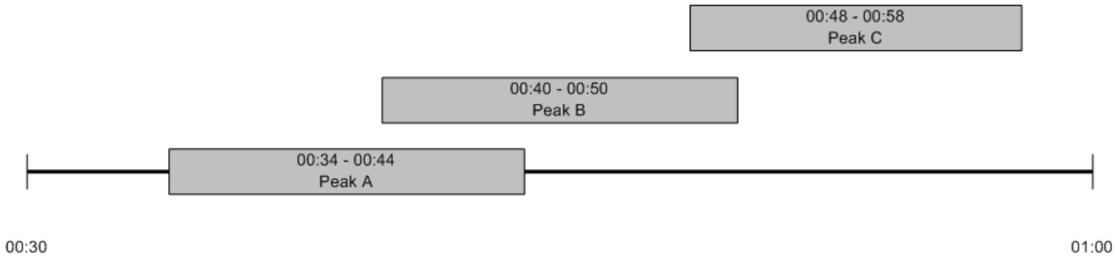


Figure 4: Illustration to the problem why inter-rate agreement cannot be calculated on the ground truth. Peak A overlaps with Peak B, Peak B overlaps with Peak C but Peak A does not overlap with Peak C.

In total, the assessors identified 293 different narrative peaks, of which 205 peaks are identified by only one assessor, 67 peaks that are identified by two assessors and 22 peaks that are identified by all three assessors.

Three Peaks in each Episode

Another issue is that annotators had to set three peaks per episode. For evaluation purposes annotators were asked to set three peaks per episode, no matter how they felt the dramatic tension was distributed in an episode. In episodes the annotators observed more than three peaks they were asked to annotate only the top three moments where dramatic tension reached its highest level. However, if there were only one or two or maybe even zero of these moments they still had to return three peaks. This of course introduces noise to the ground truth.

3.1.4 Peaks in VideoCLEF

While viewing the *Beeldenstorm* episodes, we identified three different types of behavior that increase the likelihood of a peak. First, we noticed that Prof. van Os makes a certain gesture with his hands when he is excited over what he is showing to the audience. Although we do not have the resources to build a detector for theses gestures it is still interesting to see how he tries to pull the audience into the show using these gestures. Other features we found to be good indicators of peaks are: distribution of peaks and pronoun usage, which is explained in the following subsections.

Distribution of Peaks

If we take a closer look at the distribution of peaks we see that most peaks are either set in the first or last minute of each episode. An overview of the peak distribution can be seen in Figure 5. This indicates that the TV program *Beeldenstorm* tries to grasp the attention of the viewer at the beginning and at the end of the episode. Peaks at the end of an episode are most times "summary peaks". Within one or two sentences Prof. van Os tries to summarize the episode, which most times the assessors annotate as a peak. A good example of such a peak is *...zonder al die andere engelen zou deze engel minder betekenis hebben gehad, en dat voel je, zo'n tocht op zoek naar engelen in musea is werkelijk lonend…* ('…without all the other angels this angel would be less significant, and you feel it, such a trip in search of angels in museums is truly worthwhile…').[1]

[1] From *Beeldenstorm* episode *Engelen op doek*, 'Angels on canvas'

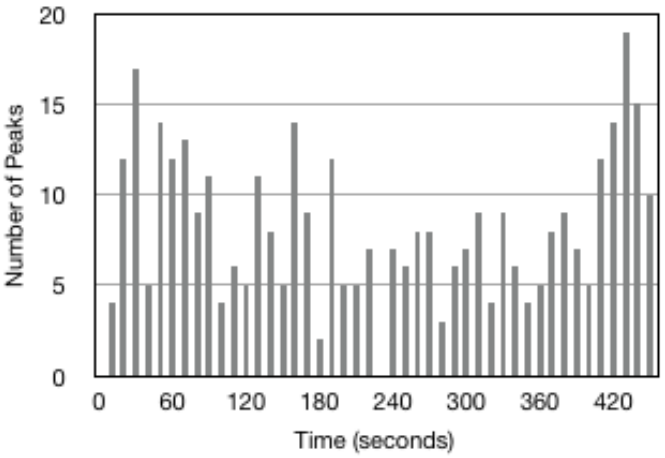


Figure 5: Distribution of peaks in VideoCLEF, although must episodes play for eight minutes the last minute consists of the credits and a commercial message. In this figure we therefore ignore the last minute, since it does not contain narrative peaks.

Peaks at the beginning of an episode mainly fall in two categories: “statement peaks” in which Prof. van Os explicitly states his view on a certain topic or painting or “humorous peaks” which contain humorous statements like: *…ik vind het zo erg dames en heren en dan zit er ook nog iets religieus bij, dat maakt het nog veel erger…* (‘…I'm so sorry ladies and gentlemen, and then there is something religious in that makes it even worse…’)[1].

This means we have “summary peaks”, “statement peaks” and “humorous peaks”. Of course, not all peaks can be categorized using these categories. For example, sometimes the dramatic tension of the viewer is increased by dramatic music to enforce the emotional intensity. However, after viewing the episodes with the peak annotations it becomes obvious that most peaks do fall in either of these categories.

First and Second Person Pronouns

Our second observation is that Prof. van Os sometimes addresses the audience directly. Although these events are not always annotated as a peak it might still prove a good indicator for a peak. We conjecture that dramatic tension rises along with the level to which the viewers feel that they are directly involved in the video content they are watching. From the observation we identified two possible conditions of heightened viewer involvement: when viewers feel that the speaker in the videos is addressing them directly or as individuals, or, second when viewers feel that the speaker is sharing something personal. Although we do not examine this aspect more closely here, it is possible that the importance of personal connection or personal revelation in documentary video is related to the fact that viewers perceive it to be a relatively rare event, which triggers them to sit up and take notice.

[1] From *Beeldenstorm* episode *Antioni Mancini*

Cases of peaks that support the viability of this approach occur in the example set, e.g., *...ziet u hoe diep de tulp in ons nationale volksziel is ingedrongen...* ('...you see how deeply the tulip has penetrated our national consciousness...').[1] In the case of *Beeldenstorm*, second person informal pronominal forms (e.g., *je*, 'you, your') should also be attributed this general role as well since they are used as impersonal pronouns to describe the thoughts and actions of a hypothetical person, rather than the viewer directly. This point is illustrated by the following peak from the example set *...en als je nou naar Amsterdam gaat, naar het Museum Willet-Holthuysen, kijk, daar heb je wat ik 'total design' zou willen noemen...* ('...and if you (informal) go to Amsterdam to the Willet-Holtuysen Museam, that's where you'll (informal) find what I call *total design*...').[2] Dutch usage conventions prevent Prof. van Os from addressing his audience using the informal, although it must also be kept in mind that his ability to stretch conventions is part of his narrative talent.

3.2 SEMAINE

The SEMAINE project is a European Commission Seventh Framework Program (EU-FP7)[3]. To this day conversations between humans and machines are substantially different from conversations between humans [43]. While humans can talk to one another for sustained periods, possibly hours, and may give limited importance to the actual content of the interaction, human-machine dialogue is often task oriented and finishes as soon as the task if fulfilled. The aim of the SEMAINE project is to build a Sensitive Artificial Listener (SAL), a multimodal dialogue system with the social interaction skills needed for sustained conversation with a human user. A very well-known SAL-like system is the Eliza chat bot [44], which was build to emulate a psychotherapist, using a text-interface. In SEMAINE the system will emphasize "soft" communication skills, like non-verbal, social and emotional perception, interaction and behavior capabilities. Therefore the systems contains only very limited verbal capabilities.

In the SEMAINE system users can interact with one of the four characters (known as SAL agents). Each agent has a particular emotional agenda and a conversational goal of shifting the user towards that state. They are Prudence, who is even-tempered and sensible; Poppy, who is happy and outgoing; Spike, who is angry and confrontational; and Obadiah, who is depressive. Users are free to choose which agent they will talk to at any given time.

[1] From *Beeldenstorm* episode *Tulpomanie*, 'Tulip mania'

[2] From *Beeldenstorm* episode *Leven met kunst*, 'Living with art'

[3] http://cordis.europa.eu/fp7/home_en.html

3.2.1 SEMAINE Dataset

The SEMAINE dataset [45] contains 13 recordings with each recording consisting of four sessions. In each recording a user has a conversation with each of the SAL agents (Poppy, Obadiah, Spike and Prudence), hence the four sessions. The user can switch to the next agent whenever they want and can also choose in which order to talk to them. Most conversations between the user and the agent last between two and ten minutes.

The conversations are annotated using the Feeltrace annotation toolkit. Feeltrace is an instrument developed to let observers track the emotional content of a stimulus, as they perceive it over time, which allows the capturing of valence and arousal levels of the sessions. More information about Feeltrace can be found in paragraph 3.2.2. Four annotators annotated the SEMAINE dataset. Each session is annotated by at least one annotator to the maximum of four annotators. Besides these Feeltrace annotations the following basic emotions are also annotated: happiness, surprise and anger. Other emotions were not annotated because they are likely to be either rare or absent. Finally, two of the 'epistemic / affective' states are annotated, namely agreement and interest [46].

The dataset also contains the audio recording of the sessions. Both user and operator (SAL agent) speech are available. Speech transcripts also exist, although unlike VideoCLEF, they are aligned on sentence level and not on word level. The transcripts also hold vocalizations such as laughing, yawning, audible breathing and coughing.

Finally, video is recorded at 49.979 frames per second and at a spatial resolution of 780 x 580 pixels. Both the user and the operator are recorded from the front by both a grayscale camera and a color camera. In addition, the user is recorded by a grayscale camera positioned on one side of the user to capture a profile view of their face.

Although in theory there should be 52 sessions that could be used for our peak detection approach we noticed when inspecting the data that only 23 sessions contain both the Feeltrace annotations and the sentence level aligned speech transcripts.

It should also be noted that due to the different research goal of SEMAINE that there is no ground truth about viewers' emotional intensity, as there was no need to. We discuss this in more dept in paragraph 3.2.3.

3.2.2 Feel Trace Annotations

The SEMAINE dataset is annotated using Feeltrace. Feeltrace is an instrument developed to let observers track the emotional content of a stimulus, as they perceive it over time, allowing the emotional dynamics of speech episodes to be examined [47]. Underlying Feeltrace is a representation called activation-evaluation spaces, which has a long history in psychology [48, 49]. The scales are alternately called arousal and valence, which are discussed in more depth in paragraph 2.1.

The essential idea behind Feeltrace is to present activation-evaluation space as a circle on a computer screen, and to have observers record their impressions of emotional state by moving a cursor to the appropriate position in the space using a mouse. Users have to move the mouse cursor to the state they feel. While they move the cursor it changes color to reflect the appropriate emotional state. To supplement the color-coding, verbal landmarks are added to the circle. An example Feeltrace display is shown in Figure 6.

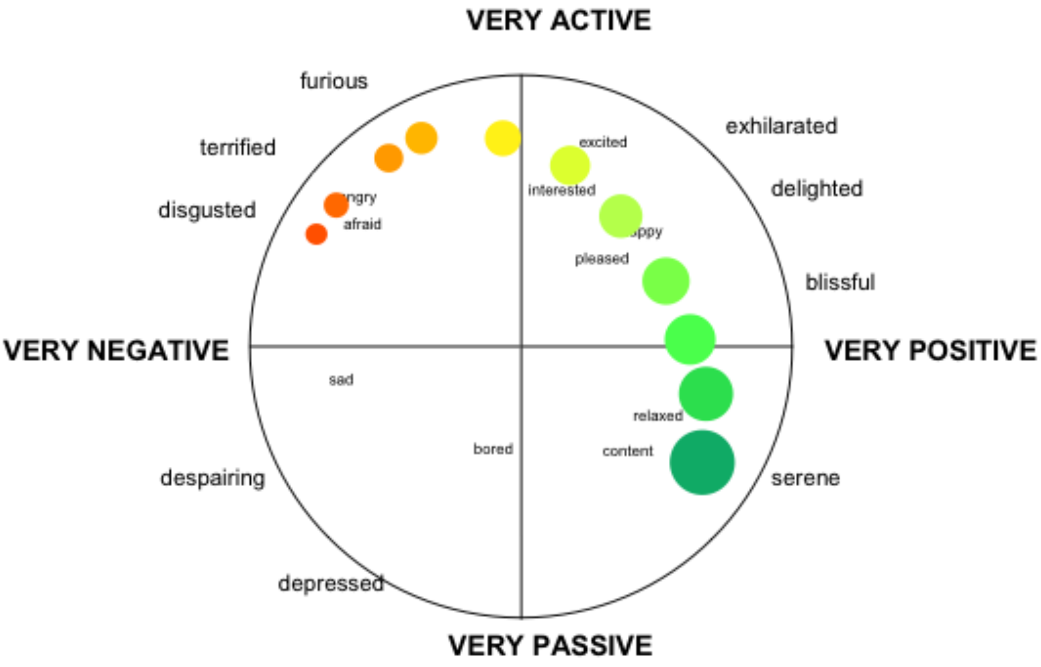
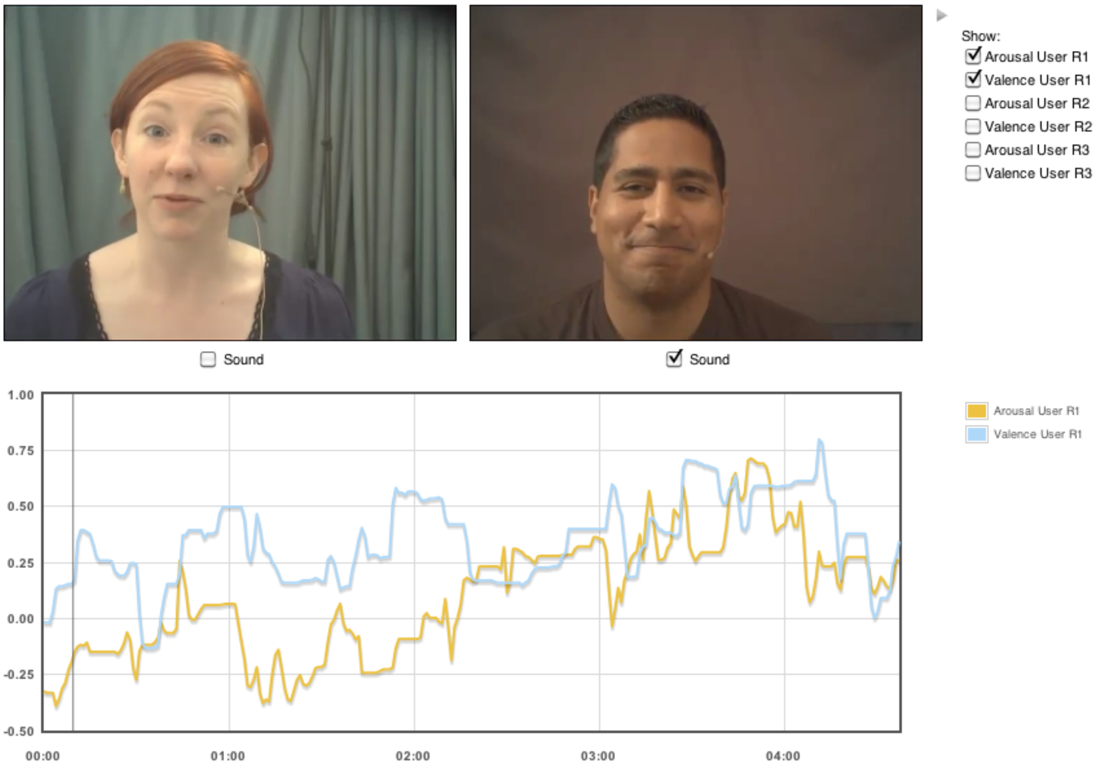


Figure 6: Example of a FEELTRACE display during a tracking session. Cursor color changes from red/orange at the left hand end of the arc, to yellow beside the active/passive axis, to bright green on the negative/positive axis, to blue-green at the right hand end of the arc. Image taken from [47].

It should be stressed that Feeltrace is not a perfect system. There are distinctions that failed to be captured, notably the distinction between fear and anger. This happens because the emotional space consists of the three dimensions (arousal, valence and dominance) and Feeltrace only has two (arousal and valence). Because the dominance dimension is left out is impossible to distinct fear and anger. For our research this is not a real issue, since we use the Feeltrace annotations to create a ground truth, which is based on valence and arousal (3.2.3).

Another issue with the Feeltrace annotation tool is that the annotations always have a delay. This is because users react to what they are seeing. This delay varies between user since some users realize quicker how they feel and experienced computer users are also able to operate the mouse cursor more proficiently. How we deal with this variation in reaction time is explained in 3.2.3. An example of these annotations is shown in Figure 7. The yellow line represents arousal and the blue line valence, both ranging from -1.00 to 1.00.



Show:
☑ Arousal User R1
☑ Valence User R1
□ Arousal User R2
□ Valence User R2
□ Arousal User R3
□ Valence User R3

□ Sound
☑ Sound

Figure 7: Example of the Feeltrace annotations, over time (x axis; minutes:seconds). The yellow line represents the arousal values and the blue line valence, both ranging from -1.00 to 1.00.

3.2.3 SEMAINE Ground Truth

Unlike VideoCLEF, the SEMAINE dataset has no annotated emotional intensity peaks. However, the dataset is annotated with continuous valence and arousal levels by up to four raters using Feeltrace. From these annotations a ground truth is extracted containing the emotional intensity peaks of the experiencer using the following steps.

First, 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 [50]. Besides the potential risk of sacrificing important individual information there is a second risk when averaging the traces; traces can flatten each other out, resulting in a flat trace. However, averaged annotations can also outperform models trained on solely on individual-specific annotations as [51] have shown.

Then the changes in valence and arousal are calculated for every 0.5 seconds. An increase in valence or arousal is assigned with a positive change in arousal, a decrease negative. The reason we choose to use the changes in valence and arousal and not the valence and arousal values itself is because we want to detect spikes and not a heightening in emotional intensity. By calculating the changes in valence and arousal only the spikes in emotional intensity are detected. Unlike VideoCLEF annotators were asked to annotate valence and arousal and not peaks. From our observations using the feature browser we noticed that most spikes in valence and arousal happen within five seconds of the video content, while in VideoCLEF peaks would last for 9.4 seconds on average. As we already noted earlier, one of the issues with Feeltrace is the reaction time of the users. Some users annotated slower than others. Therefore, we extend the windows of five seconds to seven seconds.

In order to reflect intensity, we take changes in arousal only in account when those cases are active, i.e., positive. A negative arousal corresponds to passive affect – a lack of involvement or engagement and should not contribute to intensity. We calculated the *intensity* using Equation 1 for every 0.5 seconds using the average arousal and valence values of the continuous annotation. Equation 1 is based on Figure 1 and returns the distance between from the origin to a point in the 2D arousal-valence space. In order to reflect intensity, we take arousal into account only in those cases that it is active, i.e., positive. 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. This ground truth is created in order to compare the results directly with the VideoCLEF results. The resulting total is 69 ground truth intensity peaks in the 23 interaction sessions.

Equation 1

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

A Second Ground Truth

In paragraph 3.1.3 we had some remarks on the VideoCLEF ground truth, especially because the task stated that in each episodes three peaks must be identified. Therefore we created a second ground truth (@ALL), containing all of the peaks that are spikes in the emotional intensity. To create a ground truth based on those peaks we used the Java application Peak Pick[1]. With Peak Pick it is possible to select peaks from a data collection. It works by calculating a baseline for the intensity scores, and then selecting peaks based on their height compared to the baseline. Peak Pick determines the optimal value for the height/baseline ratio automatically. Based on this data it turned out that peaks are considered to be a valid peak of the height of a peak is 3 times greater than the baseline.

[1] http://redpoll.pharmacy.ualberta.ca/lab_talks/PeakPick.pdf

An example of the difference between the @3 (ground truth containing the top 3 peaks) and @ALL ground truth is shown in Figure 8 below:

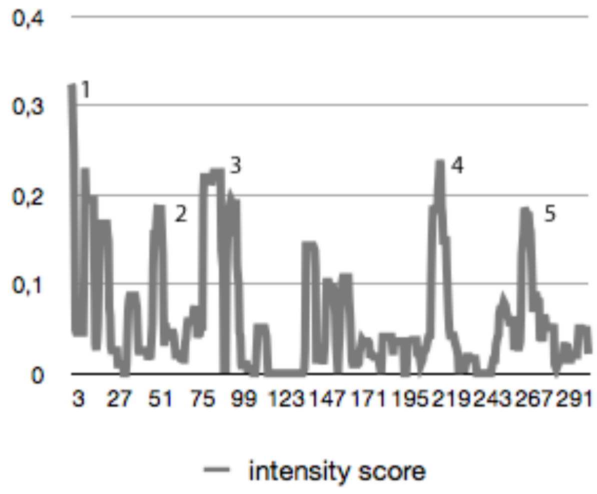


Figure 8: Example of the two ground truths. The @3 ground truth contains peak 1, 3 and 4 while the @ALL contains all peaks 1-5

In total Peak Pick identified 51 peaks in the 23 sessions. The number of peaks per episode ranges from zero to five.

3.3 Main Differences between the Datasets

Both settings involve unilateral intent on the evokers' part to shift the affective state of the experiencer. In this section the three main differences between the VideoCLEF and SEMAINE dataset are presented.

SEMAINE sessions are interactive

In VideoCLEF the episodes are not interactive, while the SEMAINE sessions are. Prof. van Os presents the series but the experiencer cannot react on his statements. In SEMAINE the experiencer can react to the evoker at any given time.

Communication

The conversations in SEMAINE are one-on-one, while in VideoCLEF there is one-to-many. In SEMAINE the evoker can use a more personal approach to influence the experiencers' state towards a particular emotion. Prof. van Os talks to a larger audience and therefore has to utilize a more general approach.

Dimensions of annotations

Finally, as we already mentioned earlier in this chapter the dimensions of annotations are different. Annotations of VideoCLEF consist of viewer-reported narrative peaks, while SEMAINE is annotated with continuous affective ratings.

4 Experimental Framework

Chapter two described the state of art peak detection approaches and the features they used. In this chapter we formulate which features were selected for the automatic intensity peak detection. We build and test two models: the evoker model in a narrative setting, predicting peaks of viewer response and the experiencer model in a conversational setting, which detects peaks in interlocutor's speech. 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.

4.1 Domain-specific Features

The following features are selected based on the viewing of the video clips in each dataset. However, since the SEMAINE dataset was not available in time we could only select features that indicate a peak from VideoCLEF. Although the selected features are domain-specific features we still use the same features in both models, so the results of the models can be compared.

4.1.1 Features Related to Dramatic Pauses

In paragraph 3.1.4 we observed that in the *Beeldenstorm* episodes peaks could be categorized in "summary peaks", "statement peaks" and "humorous peaks". The problem with these peaks is that high-level features determine these peaks; it is the meaning that increases the dramatic tension of the viewer and not some low-level features. Luckily, the producer of *Beeldenstorm* helps us here a little. To increase the dramatic effect, often these peaks are followed by a pause in speech to increase the dramatic tension even more. Pause is a low-level audio feature, but because most of the times music begins to play in a pause, we use the speech transcript to detect these pauses. Features related to these pauses are shown in Table 3.

4.1.2 Features Related to Pronoun Usage

From our observations in paragraph 3.1.4 we found that first and second pronoun usage are good indicators for a peak and also in our literature study we found this (cf. paragraph 2.1.1). To detect social engagement we used second personal pronominal forms (e.g., *u*, 'you'; *uw,* 'your') to identify audience directed speech and first person pronominal forms (e.g., *ik*, 'I') to identify personal revelation of the speaker. Notice that first person plural forms (e.g., *wij*, 'we') might actually be correlated with either case, serving generally to draw the audience into the story. Table 3 summarizes the domain-specific features we selected for VideoCLEF.

Table 3: Domain-specific features in VideoCLEF

Feature	Description
next-pause	Number of seconds to the next pause
previous-pause	Number of seconds to the previous pause
pronouns	Number of pronouns in a give segment

4.2 Acoustic Features

To choose our acoustic feature set we look at indicators from the literature and select a set that have been shown to perform well and also can be straightforwardly extracted from the speech. We use acoustic features corresponding to speech characteristics triggered by the physiology of emotion (cf. paragraph 2.1.1). In particular, popular features related to pitch, energy and speech rate [52] are chosen. In the following the word *segment* is used to which we refer as a subpart of a video clip. In order for our peak detection approach to detect peaks we extract features from these segments to determine if that segment is a peak or not. In section 4.4 we describe this peak detection approach in more detail.

Intensity

The intensity is based on the energy of the audio signal. To calculate the intensity for each segment, the sampled audio is first divided into non-overlapping frames of 0.5 seconds. Then the energy for each frame is calculated program *Praat*, which is developed by [53]. The *Praat* script used to extract the intensity can be seen in Appendix A. A segment consists of several frames, depending on the segment length. From each segment the features of Table 4 are extracted. In previous studies [2, 30] the acoustic intensity is used to model arousal, where an increased intensity is associated with an increased arousal.

Table 4: Intensity (acoustic) features based on the audio signal

Feature	Description
intensity	The average intensity of all frames within a given segment
min-intensity	The minimum level of intensity
max-intensity	The maximum level of intensity

range-intensity	The range of the intensity, calculated by max-intensity – min-intensity
d-intensity	Standard deviation of the intensity, (square root of the variance)
d-next-intensity	Absolute difference between the average intensity of the current frame and the average intensity of the next frame
d-previous-intensity	Absolute difference between the average intensity of the current frame and the average intensity of the previous frame

Pitch

The second acoustic feature used to detect peaks is pitch. Pitch represents the perceived fundamental frequency of a sound. Pitch is used by [2, 30] for affective modeling in video content, where they use the pitch to measure valence levels. Valence is much more complex to model than arousal [54], and the exact relation between pitch and valence is still unknown. However, since both [2, 30] successfully apply pitch features in their affective models, we also choose to use these features. Again we use *Praat* to extract the pitch from the audio signal, see Appendix A for the *Praat* script. The pitch values are calculated for the same 0.5-second non-overlapping frames we used to calculate the intensity. Table 5 shows the features based on the pitch.

Table 5: Pitch features based on the audio signal

Feature	Description
pitch	The average pitch of all frames within a given segment
min-pitch	Minimum value of the pitch
max-pitch	Maximum value of the pitch
range-pitch	The range of the pitch, calculated by max-pitch – min-pitch
d-pitch	Standard deviation of the pitch, (square root of the variance)
d-next-pitch	Absolute difference between the average pitch of the current frame and the average pitch of the next frame
d-previous-pitch	Absolute difference between the average pitch of the current frame and the pitch of the previous frame

Speech Rate

We conjecture that emotional intensity rises along with the speech rate. As recent studies have shown, increased speech rate is associated with high arousal ratings [55] and indicates emotional intensity [56]. A slower speech rate is generally associated with passive emotions [57]. Therefore we believe that an increased speech rate as a feature contributes to the identification of a peak. The speech rate is calculated from the speech transcripts.

Speech rate can be defined in many different ways [58], depending on whether the focus is on information transfer (normally expressed in terms of the number of words per second) or on the number of events per unit (typically expressed in terms of the number of syllables or phonemes per second. Other variables determining the definition of speech rate are the inclusion or exclusion of silent pauses and the representation (orthographic or phonetic transcriptions) of the event under investigation. Because the segments have a relatively short length (between three and ten seconds see section 4.4), we choose to define the speech rate as phonemes per second, as the information transfer in such a short time is limited.

Table 6: Features based on speech rate

Feature	Description
speech-rate	Speech rate within a given segment (phonemes/second), excluding pauses
speech-rate-pause	Speech rate within a given segment (phonemes/second), including pauses

4.3 Lexical Features

Finally, the last features that are selected for peak detection are features based on the speech transcripts. We choose to focus on functional words, in particular on pronouns (cf. paragraph 2.1.1). We leave content words out of consideration due to issues related to their topic dependence, mentioned in [11]. The following features are extracted from the speech transcripts: affective word rating, parts of speech and stop words. In the following subsection these features are described in more detail.

Affective Word Scores

Our final feature based on the affective level of video content is based on the hypothesis that dramatic tension rises when the speaker in the video uses speech made vivid by the use of certain emotional words. Although emotion can be conveyed by prosodic variation, including changes in loudness, pitch and speed, emotion is also conveyed by the choice of lexical items. People tend to use specific words to express their emotions because there is a conventionalized relationship between certain word forms and certain emotions. In the field of psychology, one way of establishing the connection between word forms and emotions is to ask subjects to list the English words that describe specific emotions [48].

Each segment is assigned an affective rating score that is calculated in a straightforward manner using these basic affective levels in order to identify emotional intensity. The approach makes use of Whissell's Dictionary of Affect in Language as deployed in the implementation of [59], which is available online[1]. This dictionary of words and scores focuses on the scales of valence and arousal levels. The scales are alternately called evaluation and activation and range from -1.00 to 1.00.

Under our approach, emotional intensity peaks are identified with a high arousal emotion combined with either a very pleasant or unpleasant emotion. In order to score words, we combine the evaluation and the activation score into an overall affective word score using Equation 1. From each segment the average affective word score is extracted using Equation 2.

Equation 2

$$\textit{affective-word-score} = \frac{\sum_N wordscore}{N}$$

Here, *N* is the number of words within a segment that are included in Whissell's Dictionary. In order to apply the dictionary, we first translate the Dutch-language speech recognition transcripts into English using the Google Language API[2]

Table 7: Features based on speech rate

Feature	Description
affective-word-score	The affective score as calculated with Formula 1 and 2.

Part of Speech Information

We conjecture that the part-of-speech tags of words contain information that helps identifying peaks. As we saw from the observations in paragraph 3.1.4 the use of first and second person pronouns are a good indicators for a peak. One good example is the use of adjectives: *...op een hele specifieke manier gaat beleven, want dit is een hele mooie ruimte in een groot paveljoen...* ('...experience

[1] http://technology.calumet.purdue.edu/met/gneff/Publications/ica02/affectdictionary.html

[2] http://code.google.com/intl/nl/apis/ajaxlanguage/

in a very specific way, because this is a beautiful space in a large pavilion…').[1] Although we do not see any direct correlation between the number of adjectives and the annotations we still believe this information might prove helpful in combination with other features.

The part of speech information is extracted by using the speech transcripts. For the *Beeldenstorm* episodes *Tadpole is used*. *Tadpole*, stands for *Tagger, Dependency Parser, and Other Linguistic Engines*, is an integration of memory-based language processing modules developed for Dutch [60]. The following parts-of-speech are extracted from each segment: adjectives, nouns and verbs (and of course pronouns, paragraph 3.1.4).

Table 8: Part of speech features

Feature	Description
adjectives	The total number of adjectives
nouns	The total number of nouns
verbs	The total number of verbs
pronouns	The total number of pronouns

Stop Words

Our final hypothesis is that peaks contain an assortment of words that are not used on a frequent basis (e.g. a richer vocabulary) than non-peak segments. In order to measure the words in a segment we use a stop word list. Stop words are words that are filtered out prior to, or after, processing of natural language data. Words that are filtered out are words that would make poor index terms [61]. Stop lists contain the most frequently occurring words. We used a stop list that was provided by TNO (Netherlands Organization for Applied Scientific Research)[2], which contains 1358 stop words. The majority of this list consists of functional words. The features related to stop words are shown in Table 9.

Table 9: Stop word features

Feature	Description
total-words	The total number of words
stop-words	The total number of stop words

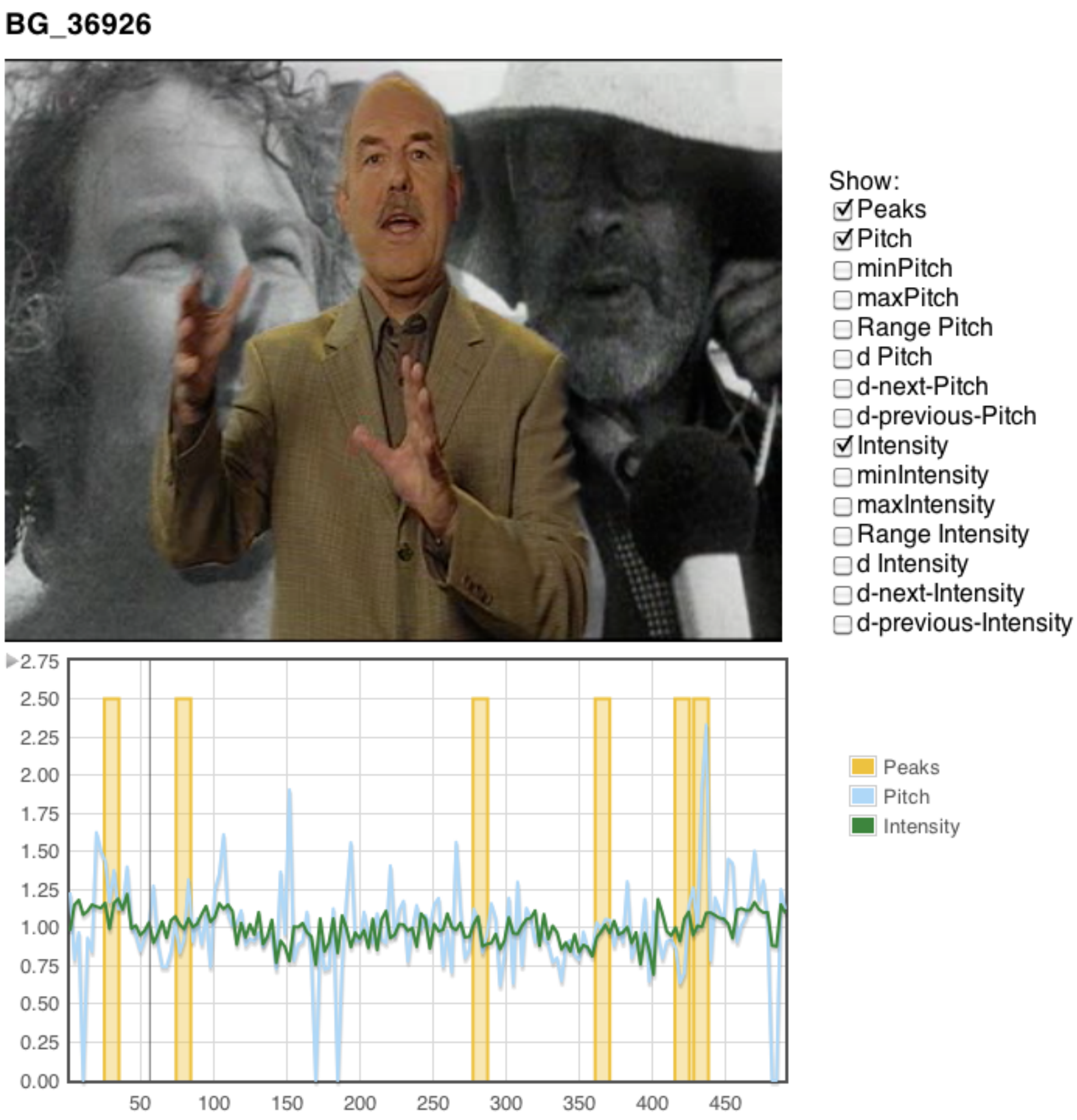
[1] *Beeldenstorm* episode *Museum Insel Hombroich*

[2] http://www.tno.nl/

non-stop-words	The total number of non stop words

4.4 Peak Detection Approach

Our initial thought was to make a rule based peak detection system. Based on a list of rules the peaks would be detected for each episode. To develop a successful rule-based peak detection system a set of rules must be defined. From our observations in paragraph 3.1.4 we already mentioned that we failed to see any clear indicators as to which specific audiovisual features could be used to identify peaks, even when looking at the annotations that were provided. To get a better insight in the features and what combination of features form a peak we build a feature browser. The feature browser shows all the features from the previous sections on a timeline, together with the annotations. The video is shown above the browser. Features can be turned on and off to see which combinations are useful for detecting peaks. The feature browser is shown in Figure 9.



BG_36926

Show:
☑ Peaks
☑ Pitch
□ minPitch
□ maxPitch
□ Range Pitch
□ d Pitch
□ d-next-Pitch
□ d-previous-Pitch
☑ Intensity
□ minIntensity
□ maxIntensity
□ Range Intensity
□ d Intensity
□ d-next-Intensity
□ d-previous-Intensity

Figure 9: Feature browser (showing only audio features)

However, even with the feature browser we were not able to find any obvious combinations of features that indicate peaks. In total there are 27 different features, which makes it hard to get a good understanding of the relations between the different features. Another problem is that a combination of features that lead to good peak detection results in one episode, do not necessarily lead to good peak detection results in other episodes. Furthermore, with 45 episodes it is almost impossible for humans to find a list of rules based on this data.

Therefore, the idea of using a rule-based peak detection system was rejected. Instead we apply learning algorithms to identify which combination of features offers the best peak detection performance. Learning algorithms "learn" from observations and have generally a good performance. This performance depends on the quality and quantity of training data. If the quality of the training data is high (less noise) and sufficient data is available performance will be good. On the other hand, if there is insufficient training data available performance of a classifier will decrease. Both datasets do not have a training set available. Therefore we use all data as training data and use a leave-one-out cross-validation to evaluate the models.

In a leave-one-out cross-validation, the original dataset is randomly partitioned into *k* sub sets, where *k* is equal to the number of video clips in the dataset. Of the *k* sub sets, a single set is retained as the validation data for testing the model, and the remaining *k* – 1 sub sets are used as training data. The cross-validation process is then repeated *k* times, with each of the *k* sub sets used exactly once as the validation data. The *k* results from the folds then can be average to produce a single estimation.

4.4.1 Applying Learning Algorithms on the Dataset

To apply machine learning techniques on the peak detection task we first need to convert the dataset. Each episode must be divided into segments of the same length. From each segment, the features as described in the previous sections are extracted, so that each segment contains 27 attributes (pitch, min-pitch, range-pitch etc). An extra attribute called "peak" is also added. This peak attribute indicates whether the segment is a peak or not and the values are *yes* or *no*. We set this value to *yes* if more than 70% of the segment overlaps with a reference peak from the ground truth.

Because the length for all segments must be the same we need to set it to a certain value. From the VideoCLEF definition of peaks we know that ten seconds is the maximum peak length, so using segments of ten seconds would make sense. However, using a length of ten seconds could average the pitch and intensity features as small increases or decreases are faded out. The shortest peak in the ground truth is 3.1 seconds, indicating that a peak needs some time to manifest. By using segments of three seconds, small increases/decreases in pitch and/or intensity are not lost with feature extraction. It is impossible to determine the optimal value for the segment length and thus we choose to vary the length between three and ten seconds and obtain the optimal value from the evaluation.

4.4.2 WEKA Toolkit

To apply learning algorithms on the dataset the Weka toolkit is used [62]. Weka is a collection of machine learning algorithms for data mining tasks. Before we can use our data in Weka we first need to convert it to the ARFF (Attribute-Relation File Format) file format used in Weka. An ARFF file is an ASCII text file that describes a list of segments sharing a set of attributes.

```
@attribute id string
@attribute pitch numeric
@attribute intensity numeric
@attribute range-pitch numeric
@attribute range-intensity numeric
@attribute affect numeric
@attribute speech-rate numeric
@attribute speech-rate-pause numeric
@attribute peak yes or no

@data
'BG_36926 0.0 5.0',147.67783,70.415708,188.42131,59.30469,0,5.034325,4.4,no
'BG_36926 5.0 10.0',120.848728,77.358411,43.13556,14.07578,0.076039,5.839416,3.2,no
'BG_36926 10.0 15.0',122.178043,72.681101,58.28801,15.33588,0,11.409396,3.4,no
'BG_36926 15.0 20.0',192.227918,78.056019,148.62988,20.67351,0.082297,11.201629,11,no
'BG_36926 20.0 25.0',191.334702,73.52165,117.35513,23.56143,0.006992,8,8,yes
'BG_36926 25.0 30.0',179.467202,71.052695,75.83876,23.29088,0.057265,7.569721,7.6,yes
'BG_36926 30.0 35.0',165.265538,77.129202,80.99775,21.70483,0,10.973085,10.6,no
'BG_36926 35.0 40.0',161.946066,77.224823,129.19201,22.81541,0.073195,8.130081,8,no
```

Figure 10: Example ARFF file

As can be seen from Figure 10 the ARFF file contains all the segments and each are assigned a set of features (attributes). The last attribute indicates whether this segment is a peak or not. We added an extra attributed called ID. This attribute is used to identify each segment so that it later can be used for the VideoCLEF evaluation.

4.4.3 Classification Algorithms

What we are doing is a typical example of supervised learning, where the training data is labeled. Supervised learning involves learning a function from examples of its inputs and desired outputs [63]. Learning is done by so called classifiers, which maps sets of input attributes to tagged classes (in this case *yes* or *no*). In Weka a wide range of classifiers is available, each with its own strengths and weaknesses. We choose the Naïve Bayes and J48 classifiers; first, because they are relatively simple compared to other classifiers and therefore generate a more generic model. A second reason for choosing these classifiers is that they both are able to output the predictions per segment. Using these predictions it is possible to select the top three peaks per episode needed for the VideoCLEF evaluation. The final reason is that the Naïve Bayes classifier is a generative model while the J48 is a discriminative model. Generative models contrast with discriminative models, in that a generative model is a full probability model of all variables, whereas a discriminative model provides a model only of the target variable(s) conditional on the observed variables. By using two different models we can test whether the performance depends on the model or on the features. The following subsections give a bit more detail about the classifiers and why we choose them.

Naïve Bayes

A Naïve Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with naïve independence assumptions. A Naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Although, we expect that features in our datasets are dependent of each other - a peak is not formed by only one feature, but a combination – Naïve Bayes classifiers have worked quite well in situations with dependent features. The main reason for choosing the Naïve Bayes classifier is that it requires only a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

J48

The J48 is a decision-tree classifier based on the C4.5 algorithm developed by Ross Quinlan [64]. Although in general decision-tree classifiers cannot output probability distributions, C4.5 has some built-in algorithms to calculate these distributions. A decision tree takes as input an object described by a set of attributes and returns a "decision", the predicted output value for the input. Decision trees reach their decision by performing a sequence of tests. Each internal node in the tree corresponds to a test of the value of one of the properties, and the branches from the node are labeled with the possible values of the test. Each leaf node in the tree specifies the value to be returned if that leaf is reached. The main reason for choosing the J48 classifier is that the decision tree returned by WEKA can be visualized and expressed in a set of rules; hence a rule based peak detection system. An example of such a decision tree is shown in Appendix B. The J48 implementation provided by Weka automatically applies pruning to the generated decision-trees, this avoids overfitting and over-complex trees.

4.4.4 Non-overlapping and Overlapping Segments

When creating the segments we created both non-overlapping and overlapping segments. The non-overlapping segments move along the timeline of the video by increasing the start time of the segments by the window length. For example if a window length of five seconds is used the start and end times of the first three segments look like this: [0 – 5] [5 – 10] [10 – 15]. Overlapping segments are created by moving an *x*-second sliding window over the videos, advancing the window by one second at a step, where *x* is the window length. The first three segments for the overlapping segments look like this: [0 – 5] [1 – 6] [2 – 7].

Between the overlapping and non-overlapping segments there are two major differences. First, when using the overlapping segments the amount of data increases by a factor 3 to 10 compared to the non-overlapping segments, depending on the window length. More data means in general better classification performance. The second difference is that the times of peak segments are more accurate when using the overlapping segments. For example, if in the ground truth a peak is identified from 25 till 35 seconds, this peak is ignored when using the non-overlapping segments, since it does not overlap by 70% with one of the segments ([20 – 30], [30 - 40]). When using the overlapping segments this peak will not be ignored.

The reason to use both non-overlapping and overlapping segments is that the overlapping segments in combination with a leave-one-out cross-validation is very expensive from a computational point of view. Because there much less data when using the non-overlapping segments, training does not take long compared to training on overlapping segments. Also since the Naïve Bayes classifiers does not need a lot of training data compared to the J48 classifier it is interesting to see how well this combination will perform. For the J48 classifier we expect the best performance when using overlapping segments. Because there are more segments in the training set this decision tree should have better thresholds for the leafs when using these overlapping segments, which results in a better classification.

As stated earlier we use a leave-one-out cross-validation to evaluate the models. To make sure no properties of the data are shared between the training and test set, the folds of leave-one-out cross-validation are across the episodes.

4.4.5 From Weka Results to VideoCLEF/SEMAINE Evaluation

The final step in our peak detection system is to select the peaks from the probability distribution provided by the classifiers. Participants of the VideoCLEF affect task were required to identify the three highest peaks in each episode. Our approach detects narrative peaks using the following sequence of steps. First, to convert the Weka results into the top three peaks for each episode we used the “output predictions” option from Weka. When this option is selected Weka outputs the probability distribution for each segment. The ID attributed was used to select all peak candidates from one episode. Then, the peak candidates are ranked with respect to their predictions for being a peak. If predictions are the same across several candidates, we rank the candidates according to how many surrounding neighbors candidates they have with the same prediction. Candidates that have more neighbors are ranked higher. Finally, peak candidates are chosen from this ranked list, starting at the top, until a total of three peaks have been selected.

4.5 Peak Detection Approach SEMAINE

For SEMAINE the same peak detection approach is used as described in paragraph 4.4. In short: the sessions are divided in small segments varying from three to ten seconds. Next we extract features we selected in this chapter to each of the segments. After the features are extracted from all segments, machine-learning algorithms are applied in order to predict peaks. Although our peak detector should be as generic as possible we still had to makes some changes due to differences between the VideoCLEF and SEMAINE dataset. In the following subsections we describe these changes.

4.5.1 Sentence aligned transcripts

Speech transcripts from the datasets differ with respect to their alignment. In VideoCLEF the speech transcripts are aligned on word level while in SEMAINE the transcripts are aligned on sentence level. When the start and end time for each word is known it was no problem calculating the pronouns for a certain segment. With sentence level alignments only the start and end times of the sentences are known. When counting for example the pronouns this becomes a problem, as it is impossible to detect to which segment the word belongs to. In our approach we worked around this problem by dividing the total words in a sentence by the length of the sentence. This way we can estimate when a certain word is being said. Of course this is not as accurate as the word level alignments and therefore the textual features are not as precise as in VideoCLEF.

Because of these sentence-aligned transcripts we could not use the same implementation for the next-pause and previous-pause features as we used for VideoCLEF. The implementation relies on the exact time markers in the speech transcripts. Although it would still be possible to build a pause detector based on the audio signal, we decided not to implement these two features for SEMAINE, since these features are domain specific features (paragraph 3.1.4). Based on our observations we noticed that there are no dramatic pauses in SEMAINE - the datasets consists of continuous dialogs – making the pause detector redundant.

4.5.2 Part of Speech Information

In SEMAINE the spoken language is English, while in VideoCLEF the spoken language is Dutch. *Tadpole*, the part of speech tagger we used for the VideoCLEF dataset only works for Dutch language. We decided to use the Stanford part of speech tagger[1], originally developed by [65]. The main advantage of this tagger is that it is written in Java making it possible to implement it directly in the peak detection workflow and therefore eliminating manual intervention that was needed with *Tadpole*.

4.5.3 Speech Rate

As we already mentioned in paragraph 4.5.1 the speech transcripts are aligned on sentence level. Because they are not aligned on word level it is impossible to

[1] http://www-nlp.stanford.edu/software/tagger.shtml

calculated an accurate speech rate from the speech transcripts. In order to still calculate the speech rate we use a *Praat* script to detect the syllable based on audio only. This script is developed by [66] and allows us to calculate the speech rate from the syllable nuclei.

4.5.4 More Speakers

In SEMAINE there are nine users that have conversations with the agents, both females and males, while in VideoCLEF there is only one narrator. Because females have a higher average pitch than males, and the intensity between users vary, the pitch and intensity features must first be normalized. We normalized the features with a z-normalization based on each session.

5 Evaluation VideoCLEF

This chapter describes how the evaluation on the VideoCLEF dataset was performed. First, we explain the two different scoring methods that were used for the evaluation. Then we present the results of our models for the narrative peak detection task and do a failure analysis, to get a better understanding where our models correctly predict peaks and where it fails. Finally, we conclude this chapter with a discussion on the results. But before we explain the scoring methods, we present a short summary of the VideoCLEF dataset and our peak detection approach.

The VideoCLEF dataset consists of 45 episodes of the Dutch TV series *Beeldenstorm*. *Beeldenstorm* limits the spoken content to a single speaker. Three Dutch annotators created the ground truth by marking three points per episode, where they felt the dramatic tension reached its highest level. We used a Naïve Bayes classifier and a decision tree for our experiments and trained models based on the 45 episodes. Evaluation was done by leave-one-out cross-validation. In each fold, one episode/session was left out. Peaks are selected based on the probability distribution from the models. The three segments with the highest probability are returned as a peak.

5.1 Scoring Methods

Evaluation is done with two scoring methods, a point-based scoring method and a peak-based scoring method, both provided by VideoCLEF 2009. The point-based scoring method gives a point for each overlapping peak with the ground truth. Peaks are considered overlapping if that peak is within eight seconds of the midpoint of a reference peak. The total number of points is the evaluation score. A single episode can return a score between three points (assessors chose completely different peaks) and nine points (assessors all chose the same peaks). However, there are no episodes were all assessors picked the same peak, or completely different peaks. A perfect run returns 246 points with the point-based scoring method.

Under peak-based scoring, the total number of correct peaks is reported as the run score. The peak-based scoring is subdivided into personal peaks (peaks identified by only one assessor), pair peaks (peaks that are identified by at least two assessors) and general peaks (peaks that all three assessors agreed on). In total, there were 205 personal peaks, 89 pair peaks and 22 general peaks. The scores for a perfect run under the peak-based scoring method are 203 for the personal peaks, 89 for the pair peaks and 22 for the general peaks.

Via simulation we calculated that an approach that randomly picks points at which to hypothesize three peaks in a file will automatically score, on average, approximately 40 points under the point-based scoring method. Under the peak-based method it would score on average 28 correct "personal peaks", nine correct "pair peaks" and two correct "general peaks". We use these scores as a baseline.

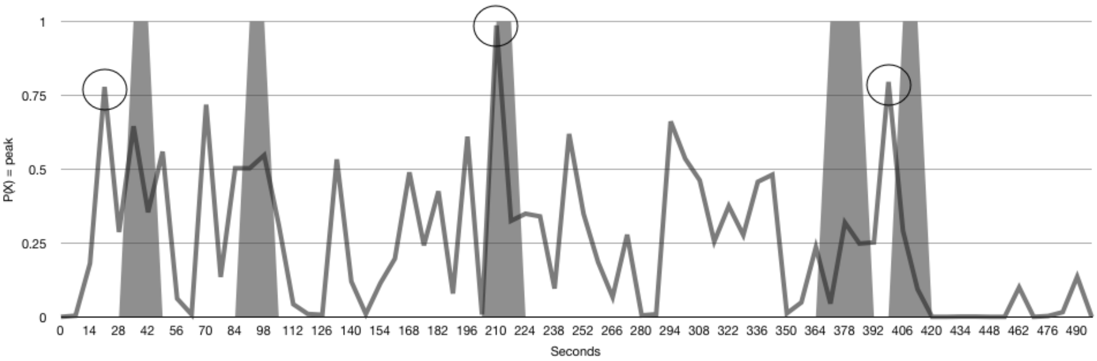


Figure 11: Peak-based scoring example, showing the probability for a peak over time. The top three peaks are circled. The grey areas are the peaks identified by the annotators.

Figure 11 shows the output predications for the *Beeldenstorm* episode *"Leven met kunst"* (Living with art). The peak detector selects the top three peaks with the highest probability for a peak, in Figure 7 these probabilities are circled. The grey areas are the peaks identified by the assessors. If we apply the peak-based scoring method on this episode, we get a result of two, since the second and third peaks are overlapping. The best possible score for this episode is four: two points for detecting the fourth peak - because two assessors marked this peak - and two points for detecting two other peaks.

5.2 Evaluation Results

The results are subdivided by classifier, Naïve Bayes and J48. For each classifier we ran six different configurations. These configurations are based upon the non-overlapping and overlapping segments and the kind of features added to the segments, audio features, text features or a combination of both features. The following subsections outline our results. Note that the results are identified by the run id, which is made of the classifier (nb for Naïve Bayes), the window length (range 3-10) and finally if the classifier was trained on the overlapping or non-overlapping segments, for example the run id nb-7-n represents a run from the Naïve Bayes classifier trained on a window length of seven seconds with non-overlapping segments.

Table 10 shows the perfect and baseline score for the different scoring methods.

Table 10: Overview of the perfect and baseline scores

Run	Point-based	Peak-based "Personal Peaks"	Peak-based "Pair Peaks"	Peak-based "General Peaks"
perfect	246	135	89	22
random baseline	40	28	9	2

5.2.1 Naïve Bayes

Table 11 lists the top three results for the Naïve Bayes classifier based on both the non-overlapping and overlapping segments. The complete results are shown in Appendix C, Table 25 and Table 26. The segments with a window length of nine seconds shows the best performance followed by seven seconds.

Table 11: Naïve Bayes results

Run	Point-based	Peak-based "Personal Peaks"	Peak-based "Pair Peaks"	Peak-based "General Peaks"
nb-7-o	64 (26.0%)	42 (20.7%)	19 (21.3%)	4 (13.6%)
nb-7-n	63 (25.6%)	41 (20.2%)	17 (19.1%)	3 (22.7%)
nb-5-n	62 (25.2%)	47 (23.2%)	15 (16.9%)	3 (13.6%)

Audio

In this section the same evaluation is done except this time only audio features are assigned to the segments. All textual features are ignored. As can be seen from Table 12 the performance decreases using only audio features, indicating that audio features are not a good indicator for a peak in this setting.

Table 12: Naïve Bayes results using only audio features

Run	Point-based	Peak-based "Personal Peaks"	Peak-based "Pair Peaks"	Peak-based "General Peaks"
nb-10-n	52 (21.1 %)	38 (18.7%)	12 (13.5%)	3 (13.6%)
nb-9-n	51 (20.7%)	38 (18.7%)	10 (11.2%)	2 (9.1%)
nb-9-o	46 (18.7%)	32 (15.8%)	11 (12.4%)	2 (9.1%)

Text

Using only text features the performance is almost as good as using both audio and text features, indicating that text features form the main contribution to the peak detection. Again we see almost the same performance when using overlapping segments compared to the non-overlapping segments.

Table 13: Naïve Bayes results using only text features

Run	Point-based	Peak-based “Personal Peaks”	Peak-based “Pair Peaks”	Peak-based “General Peaks”
nb-9-n	65 (26.4%)	47 (23.2%)	17 (19.1%)	3 (13.6%)
nb-10-n	62 (25.2%)	43 (21.2%)	13 (14.6%)	5 (22.7%)
nb-6-o	59 (24.0%)	38 (18.7%)	17 (19.1%)	4 (18.2%)

5.2.2 J48

Table 14 shows the top three results for the J48 classifier using both audio and text features. Here, we see clearly that models trained on the overlapping segments outperform models trained on the non-overlapping segments. Using non-overlapping segments the performance drops to the level of the random baseline detector. Again models trained on segments with a longer window length perform better than segments with a shorter window length. Overall the models trained with the J48 classifier perform better then models trained with the Naïve Bayes classifier.

When we take a deeper look at the results of the best scoring run, j48-10-o, we see that the trees generated for this model are very complex; the total number of leaves for this tree is 564 and the size of the tree is 1117. These numbers are based upon the pruned tree, provided by Weka.

Table 14: J48 results

Window Length	Point-based	Peak-based “Personal Peaks”	Peak-based “Pair Peaks”	Peak-based “General Peaks”
j48-10-o	75 (30.5%)	49 (24.1%)	22 (24.7%)	8 (36.4%)
j48-8-o	74 (30.1%)	48 (23.6%)	22 (24.7%)	7 (31.8%)
j48-7-o	74 (30.1%)	46 (22.7%)	21 (23.6%)	7 (31.8%)

Audio

Just like the models trained on both audio and text features, the models based on the non-overlapping segments perform similar to the random baseline detector. The top three results for the overlapping segments are shown in Table 15. Just like the results of the Naïve Bayes classifier the performance decreases when using only audio features compared to models trained on both audio and text features, showing that audio features alone are not able to identify the narrative peaks created by Prof. van Os with our approach.

Table 15: J48 results using only audio features

Window Length	Point-based	Peak-based "Personal Peaks"	Peak-based "Pair Peaks"	Peak-based "General Peaks"
j48-9-o	50 (20.3%)	36 (17.7%)	11 (12.4%)	3 (13.6%)
j48-7-o	48 (19.5%)	35 (17.2%)	10 (11.2%)	2 (9.1%)
j48-10-o	47 (19.1%)	34 (16.7%)	11 (12.4%)	2 (9.1%)

Text

Table 16 shows the results of models trained with only text features. Again, the overlapping segments clearly outperform the models trained on the non-overlapping segments. The performance using only text features is almost equal to the performance of models trained on both audio and text features, indicating that, text features from the main contribution to the peak detection performance, similar to the results based on the Naïve Bayes classifier.

Table 16: J48 results using only text features

Window Length	Point-based	Peak-based "Personal Peaks"	Peak-based "Pair Peaks"	Peak-based "General Peaks"
j48-10-o	77 (31.3%)	51 (25.1%)	21 (23.6%)	9 (40.9%)
j48-9-o	73 (29.7%)	52 (25.6%)	18 (20.0%)	5 (22.7%)
j48-8-o	73 (29.7%)	49 (24.1%)	19 (21.3%)	8 (36.4%)

5.3 Failure Analysis

After the experiments we analyzed a selection of cases manually in order to better understand the results. We focus on the 22 general peaks since all the annotators agreed upon these peaks. 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. In our failure analysis we use the results based upon the overlapping segments with a window length of ten seconds because in general this setting showed the best performance.

First, we checked if the performance of the two classifiers, the Naïve Bayes and the J48 classifier, is related. For text, the Naïve Bayes classifier finds two of the same correct peaks that the J48 decision tree classifiers finds. For audio, there is one. Then we looked at the three peaks that both the audio and the text classifier correctly identified. These are dramatic pauses, he gives his opinion directly using "I", he tells the audience what the show will be about and he discusses the deep emotions of one of the painters. The features that we selected seem to be straightforwardly picking up characteristic properties of points that annotators agree are peaks.

For the analysis of where our algorithm fails, we looked at ten points in ten different episodes at which both audio and text classifiers predicted to be peaks, but which turned out to be false positives (i.e. annotators did not select this part as their top three peaks). We used the better performing classifier, J48, for this test. All of these predicted peaks seem to be plausible as narrative peaks. In two cases, the classifier picked up the final sentence of the episode, (which we defined as "summary peaks" in 3.1.4). Since closing remarks were often agreed upon as peaks by all three annotators, these seem to be indeed very plausible peaks. Indeed, both of these points had been chosen by one of the three annotators to be a peak.

In four of the cases, the peak did have characteristic properties, but within the narrative it was functioning as a transition from one topic to the next. Apparently, such transitions do not have a peak-like affective impact on viewers. Another two peaks fell right after a reference peak, which still gave the impression of being a heightening of narrative tension, but the real spike in narrative tension were just before these peaks.

5.4 Conclusions

Before we draw any conclusions based upon the results, the best performing runs are listed in the overview Table 17.

Table 17: Results overview

Run	Point-based	Peak-based "Personal Peaks"	Peak-based "Pair Peaks"	Peak-based "General Peaks"
perfect	246	135	89	22
random baseline	40	28	9	2
All Features				
nb-7-n	63 (25.6%)	41 (20.2%)	17 (19.1%)	3 (22.7%)
nb-7-o	64 (26.0%)	42 (20.7%)	19 (21.3%)	4 (18.2%)
j48-8-n	51 (20.7%)	35 (17.2%)	14 (15.7%)	4 (18.2%)
j48-10-o	75 (30.5%)	49 (24.1%)	22 (24.7%)	8 (36.4%)
Audio Features				
nb-10-n	52 (21.1%)	38 (18.7%)	12 (13.5%)	3 (13.6%)
nb-9-o	46 (18.7%)	32 (15.8%)	11 (12.4%)	2 (9.1%)

j48-8-n	38 (15.4%)	26 (12.8%)	8 (9.0%)	2 (9.1%)
j48-9-o	50 (20.3%)	36 (17.7%)	11 (12.4%)	3 (13.6%)
		Text Features		
nb-9-n	65 (26.4%)	47 (23.2%)	17 (19.1%)	3 (13.6%)
nb-6-o	59 (24.0%)	38 (18.7%)	17 (19.1%)	4 (18.2%)
j48-9-n	44 (17.9%)	26 (12.8%)	13 (14.6%)	5 (22.7%)
j48-10-o	77 (31.3%)	51 (25.1%)	21 (23.6%)	9 (40.9%)

As can be seen from Table 17 the best performing runs are those that were trained using the J48 classifier with overlapping segments. Under the point-based evaluation a score of 75 was achieved using both audio and textual features and a score of 77 using only textual features.

Our initial assumption that peaks could be detected by a set of (simple) set of rules (paragraph 4.4) appears to be wrong. Although a decision tree algorithm achieves the best performance, this tree is far from simple. The decision tree contains 564 leaves and the total size of the tree is 1117, which suggests that this model is very specific and it would probably also perform poorly in other domains or even other short-form documentaries. The results of the Naïve Bayes classifier are also interesting. It outperforms the random peak detector easily. As noted in the failure analysis we found the properties of narrative peaks in the corpus to be highly variable, reflecting a broad palette of creative narrative strategies. 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.

Both classifiers show, in general, an increase in performance when using segments with a longer window length. The best results are achieved when a window length between seven and ten seconds is used, indicating that a peak needs at least seven seconds to manifest itself.

When we compare the two classifiers based on the non- and overlapping segments we see that the Naïve Bayes classifiers accomplishes the best results based on the non-overlapping segments while with the J48 classifier the opposite is shown; best results are achieved using overlapping segments. The fact that Naïve Bayes performs better on less training data with less accurate peak times can be explained by the scoring methods used by VideoCLEF, which give a point if a peak is within eight seconds of a reference peak. Because of this relatively large overlap (peaks last ten seconds), models with less accurate peak times can still achieve a high performance. Also, the Naïve Bayes only needs a small amount of data for a relatively good performance compared to other classifiers.

A deeper look into the features shows that the best performing models trained with the Naïve Bayes classifier combine both the auditory and textual features. The best performance of the J48 classifier is achieved when using only textual features or a combination of auditory and textual features. When comparing the auditory features against the textual features based on performance we see that both classifiers achieve the best results when using the text features. However, if we look at each of the features individually, we see that the range-intensity, d-intensity and min-intensity are the top three features based on the information gain algorithm. This algorithm measures how much information is gained by doing a split on the dataset based on a particular feature. Other audio features did not improve the performance, as their information gain is zero. The text-based features that have the highest information gain are: speech-rate-pause, total-word, stop-words, pronouns and verbs.

Finally, since VideoCLEF is a multimedia benchmark evaluation we compare the results with the other participants of VideoCLEF 2009. In total three teams participated and their results are shown in Table 18. It should be noted that they did not train models based on the ground truth as we did, and therefore it is not possible to compare them directly. However it still gives a good indication about our performance.

Table 18: VideoCLEF 2009 results

Run	Point-based	Peak-based "Personal Peaks"	Peak-based "Pair Peaks"	Peak-based "General Peaks"
duotu09fix [67]	47 (19.1%)	28 (13.8%)	8 (8.9%)	4 (18.2%)
duotu09ind [67]	55 (22.4%)	38 (18.7%)	12 (13.3%)	2 (9.1%)
duotu09rep [67]	30 (12.2%)	21 (10.3%)	7 (7.8%)	0 (0.0%)
duotu09pro [67]	63 (25.6%)	44 (21.7%)	17 (18.9%)	4 (18.2%)
duotu09rat [67]	63 (25.6%)	37 (18.2%)	20 (22.2%)	5 (22.7%)
unige-cvml1 [68]	39 (15.9%)	32 (15.8%)	6 (6.7%)	0 (0.0%)
unige-cvml2 [68]	41 (16.7%)	30 (14.8%)	11 (12.2%)	2 (9.1%)
unige-cvml3 [68]	42 (17.1%)	31 (15.3%)	8 (8.9%)	0 (0.0%)
unige-cvml4 [68]	43 (17.5%)	31 (15.3%)	9 (10.0%)	0 (0.0%)
unige-cvml5 [68]	43 (17.5%)	32 (15.8%)	8 (8.9%)	3 (13.6%)
uaic-run1 [69]	33 (13.4%)	26 (12.8%)	7 (7.8%)	2 (9.1%)
uaic-run2 [69]	41 (16.7%)	29 (14.3%)	10 (11.1%)	3 (13.6%)

uaic-run3 [69]	33 (13.4%)	24 (11.8%)	7 (7.8%)	2 (9.1%)
nb-7-n	63 (25.6%)	41 (20.2%)	17 (19.1%)	3 (22.7%)
nb-7-o	64 (26.0%)	42 (20.7%)	19 (21.3%)	4 (18.2%)
j48-8-n	51 (20.7%)	35 (17.2%)	14 (15.7%)	4 (18.2%)
j48-10-o	75 (30.5%)	49 (24.1%)	22 (24.7%)	8 (36.4%)

Most runs failed to yield significantly better than the random peak detector. The two best scoring approaches exploited the speech recognition transcripts, in particular, the occurrence of pronouns reflecting user directed speech (duotu09pro) and the use of words with high affective rating (duotu09rat). Both of these features are included in our approach (total number of pronouns and the affective score). Models trained with the J48 classifier clearly outperform these approaches. However, it is quite surprising that models trained with the Naïve Bayes classifier fail to achieve a better performance considering all other features we included. It is hard to say why the Naïve Bayes classifier does not perform better, since models based on this classifier are so called black boxes, only the input and output can be viewed, not the internal workings. A possible explanation is that the duotu09pro and duotu09rat are very basic implementations; they both select the peaks based on the highest values of these features. It could be the case that all these values are close to each other. An algorithm that tries to learn from these features will not find any useful information since they are all so close together, while a basic approach will simply select the top three highest values.

6 Evaluation SEMAINE

This chapter presents the results achieved on the SEMAINE dataset. Two evaluations are performed, based on the two ground truths we created in chapter 4. We also evaluated performance based on agent character. The chapter concludes with a discussion. But before we start with the evaluation, a short summary is presented of the SEMAINE dataset and our peak detection approach.

The SEMAINE dataset consists of 23 sessions of emotionally colored character interactions between a human interacting fully naturally (the experiencer), and a human playing an agent with a particular emotional style. In total there are six different speakers in the 23 session, three males and three females. The corpus is annotated with continuous affective ratings from which we extracted two ground truths for emotional intensity peaks. We used a Naïve Bayes classifier and a decision tree for our experiments and trained models based on the 23 sessions. Evaluation was done by leave-one-out cross-validation. In each fold, one episode/session was left out. Peaks are selected based on the probability distribution from the models. The three segments with the highest probability are returned as a peak.

6.1 Evaluation Ground Truth @3

The first evaluation is based on the @3 ground truth, which is similar to the VideoCLEF ground truth. We apply the same point-based scoring method from VideoCLEF (paragraph 5.1), which gives a point for each overlapping peak with the ground truth. Since there are 69 peaks in the ground truth this is also the perfect score. Again, the Naïve Bayes and the J48 classifier are evaluated in the same way as the previous chapter. For each run (based on window length) we ran three different feature sets: one set with all features, one set with only audio features and one set with only text features. Only the best scoring runs are presented, the complete results are listed in Appendix D.

6.1.1 Naïve Bayes

Table 19 shows the results of the Naïve Bayes classifier with both the overlapping and non-overlapping segments. Results suggest that audio features are the more appropriate choice than text features, scoring the same as the combination of text and audio features.

Table 19: Results of the Naïve Bayes classifier

Window Length	All Features	Audio Features	Text Features
nb-7-o	20 (28.9%)	19 (27.5%)	15 (21.7%)
nb-10-o	19 (27.5%)	20 (28.9%)	14 (20.3%)
nb-9-o	19 (27.5%)	19 (27.5%)	14 (20.3%)
nb-10-n	19 (27.5%)	16(23.2%)	14 (20.3%)

6.1.2 J48

In Table 20 the results of the J48 classifier are shown. Models trained on overlapping segments outperform models trained on non-overlapping as we expected. Here, also the audio features show a better performance over the text features. In general, the combination of both text and audio features performs best.

Table 20: Results of the J48 classifier

Window Length	All Features	Audio Features	Text Features
j48-10-o	12 (17.4%)	12 (17.4%)	9 (13.0%)
j48-8-o	11 (15.9%)	10 (14.5%)	5 (7.2%)
j48-9-o	10 (14.5%)	13 (18.8%)	9 (13.0%)
j48-6-n	8 (11.6%)	6 (8.7%)	5 (7.2%)

6.2 Evaluation Ground Truth @ALL

The second evaluation is based on the ground truth @ALL, containing all peaks as indicated by the Peak Pick tool. In this evaluation we only consider models trained on the ten second overlapping segments. Based on this window length the precision and recall are calculated, both based on the correctly identified peaks. Table 21 shows the results. Recall in this context is also referred to as the true positive rate. In total there are 353 peaks in the training data. A random baseline is added to the table. This random baseline sets peaks based on a fixed probability obtained from the training set.

Table 21: Precision and recall, based on 10 second overlapping segments

Run	Precision	Recall	# Predicted Peaks
Random baseline	0.05	0.05	355
NB all features	0.18	0.21	339
NB audio features	0.20	0.19	348
NB text features	0.09	0.01	11
J48 all features	0.13	0.12	311
J48 audio features	0.11	0.08	314
J48 text features	0.09	0.07	90

The Naïve Bayes classifier clearly outperforms the J48 classifier. Highest precision and recall is achieved by using only audio features. However a combination of both text and audio features also shows high precision and recall.

6.3 Evaluation SAL Agents

In our final evaluation we analyze how well the models perform based on the different agents. To analyze the models the dataset is first divided into four smaller sets, each containing only episodes from that agent. In total there are six sessions of Obadiah, Prudence and Poppy and five of Spike. Again, the Naïve Bayes classifier is used with a window length of ten seconds. Table 22 lists the results showing the correct peaks, precision and recall of each of the SAL agents.

Table 22: Results per character

Character	Emotional Color	Precision	Recall	Correct Peaks
Obadiah	Depressive	0.06	0.07	2 (11.0%)
Prudence	Sensible	0.11	0.13	4 (22.0%)
Poppy	Happy	0.18	0.16	7 (39.3%)
Spike	Angry	0.07	0.18	3 (20.0%)

Table 23: Precision and recall for text and audio features

Character	Precision	Recall
	Audio Features	
Obadiah	0.06	0.12
Prudence	0.14	0.15
Poppy	0.21	0.17
Spike	0.06	0.09
	Text Features	
Obadiah	0.06	0.01
Prudence	0.0	0.0
Poppy	0.11	0.07
Spike	0.05	0.07

Peak detection performance is highest in the Poppy-sessions. Precision and recall on both features is also relatively high for the Poppy-sessions compared to the others. Performance of the models based on the Spike and Obadiah sessions is lowest. Noteworthy is the Prudence performance, which is in between Poppy and Spike using all features but with only textual features not one peak is identified correctly.

6.4 Failure Analysis

Just like we did in the VideoCLEF evaluation we analyzed a selection of cases manually in order to better understand our results. The failure analysis is based on models trained on the 10 second overlapping segments. First, we checked if performance between the two classifiers is related. For audio, the Naïve Bayes classifier finds five of the same correct peaks as the J48 decision tree classifiers finds. For text, there are two, which is inline with our results from the previous sections where audio features showed better scores.

We then checked the first four sessions, twelve peaks in total. Of these twelve peaks there is one peak that all of the classifiers identified correctly. This peak is a moment where the user tries to cheer up Obadiah who is depressed. Both audio and text features are picked up nicely as the pitch is higher and he says works like “happy” and “pleasant”. Other peaks that are picked up correctly are points where the agent makes an excessive statement (e.g. “you are a doormat” or “only fools believe that”) and the user responds to this by laughing and telling them why they are not. Two peaks are where the agents leave their role for a moment and have a laugh with the user. Both audio classifiers pick up these peaks.

For the analysis of where our algorithm fails, we looked again at ten points that both the Naïve Bayes text and audio classifiers predicted peaks. Of these ten peaks, seven of them seem to be plausible peaks. Four of these peaks are moments the user reacts to excessive statements by Spike and Obadiah, by either laughing or trying to convince the agent to see it otherwise. Two other plausible peaks are when the user tells about how he renovated the house, which seem like a good memory to him, however these peaks are not present in the ground truth. The last peak is where the user concludes the session and thanks the agent for a good talk and moves on to the next agent.

6.5 Conclusions

We conclude this chapter with an overview of the best performing runs, which are presented in Table 24. A random peak detector is added for the @3 evaluation. This peak detector sets three peaks in each session randomly.

Table 24: SEMAINE results overview

Run ID	All Features	Audio Features	Text Features
Random	10 (14.5%)	10 (14.5%)	10 (14.5%)
nb-7-o	20 (28.9%)	19 (27.5%)	15 (21.7%)
nb-10-o	19 (27.5%)	16 (23.2%)	14 (20.3%)
j48-10-o	12 (17.4%)	12 (17.4%)	9 (13.0%)
j48-6-n	8 (11.6%)	6 (8.7%)	5 (7.2%)

From Table 24 we can see that the highest performance is achieved using a window length of ten seven with the Naïve Bayes classifier. The performance is equal to the performance of the Naïve Bayes classifier in VideoCLEF; both identify 28% correct peaks. In the previous chapter J48 had the best performance. The better performance of the Naïve Bayes on the SEMAINE dataset is presumably due to the fact that there is less training data available compared to the VideoCLEF dataset. In VideoCLEF there is about six hours of training data available, while in SEMAINE there is only about two hours of data. And as we already saw in the previous chapter, the Naïve Bayes classifier performs better on smaller datasets then J48.

Another interesting difference with VideoCLEF is that here, acoustic features outperform lexical features. However, again we see no clear benefit in the combination of both features. We attribute the 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 markers codes are available.

Another possible explanation for the better performance of the lexical items, are the different kind of models used. In VideoCLEF an evoker models was used to detect peaks in emotional intensity, while in SEMAINE an experiencer model was used. We formulate the hypothesis that evokers depend more on lexical features to create peaks in emotional intensity, while reactions of experiencers are based on acoustic features. We cannot test this hypothesis since VideoCLEF lacks experiencer data and SEMAINE has no evoker annotations.

The better performance of the acoustic features can clearly be seen when using the same information gain algorithm as used in the previous chapter. Here, max-pitch, max-intensity and range-pitch are the top three features to split the dataset on. Other features that are selected based on this algorithm are min-pitch, min-intensity and range-intensity. Only one lexical feature is selected: the number of nouns in a segment.

In an agent character breakdown of the performance of the Naïve Bayes classifier we see that models based on the Poppy-sessions perform best. Strategies used by the agents to shift the user to an emotional state are exaggerations and encourage the user 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 Obadiah and Spike-sessions, who use the exaggeration strategy more than the other two agents. However, upbeat stories had characteristic word usage (e.g., "pleasant", "magical" and "wonderful"), reflected in the relatively good performance achieved for the Poppy-sessions. Additional data is required in order to draw conclusions with stronger validity since the results above are based on a small amount of training data per character (approximately 30 minutes).

In our failure analysis we found peaks to be characterized by conflicting views (e.g., agent says, "You are a doormat" and the interlocutor contracts) and by laughter. Our models were able to find such peaks. Of the ten false alarms we examined, seven gave the impression of plausible peaks, e.g., the interlocutor is contradicting/correcting or telling a happy story.

7 Prototype of a Peak Browser

In this chapter we prototype a user interface that utilizes peaks in emotional intensity, detected by our approach. This interface should help users navigate through a list of video clips. We call this user interface a peak browser, to browse through all detected peaks. Characteristic for the browsing model is that there is no explicit specification of information need [70].

Blanken et al. [70] state that because of the complexity of multimedia objects, there are two levels of browsing multimedia databases:

1. Browsing within a multimedia object (e.g., when looking for a frame within a movie);
2. Browsing through a collection of multimedia objects (e.g., when looking for a movie).

We also treat these levels separately; in paragraph 7.2 we propose a number of prototypes for browsing within episodes/session and in paragraph 7.3 a prototype for browsing through a collection is also proposed. This chapter starts by presenting a short overview of previous work.

7.1 Previous Studies

Foote et al. [71] have prototyped a browser that displays information extracted from the multimedia stream, such as speaker identity and shot boundaries. Because automatically derived information is not wholly reliable, they transfer these features into a confidence score. This confidence score is visualized and presented to the user in what they call and intelligent media browser. Figure 12 shows the user interface of their browser. To the top left are the usual playback window and controls. On the middle right are menu controls that select which confidence scores to display on the bottom time bar. Confidence scores are displayed time-synchronously with the video slider bar.

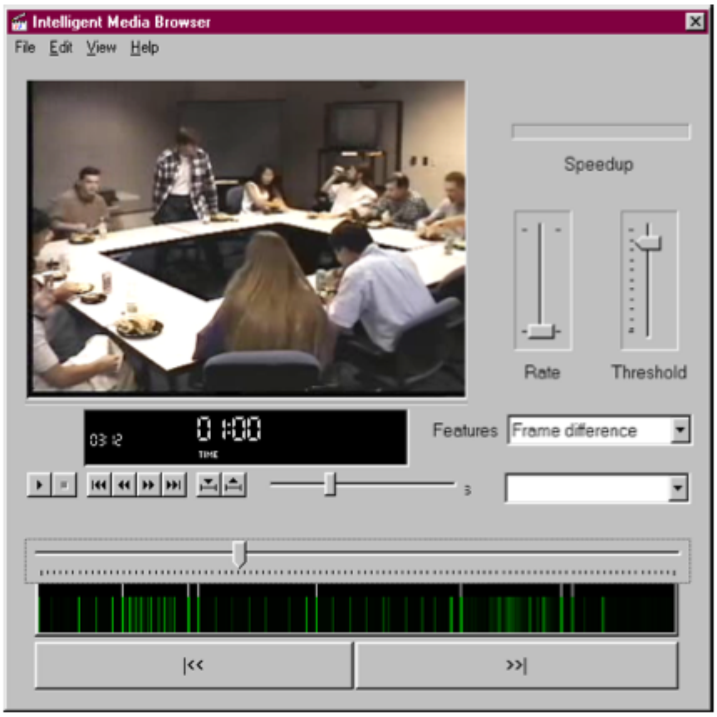
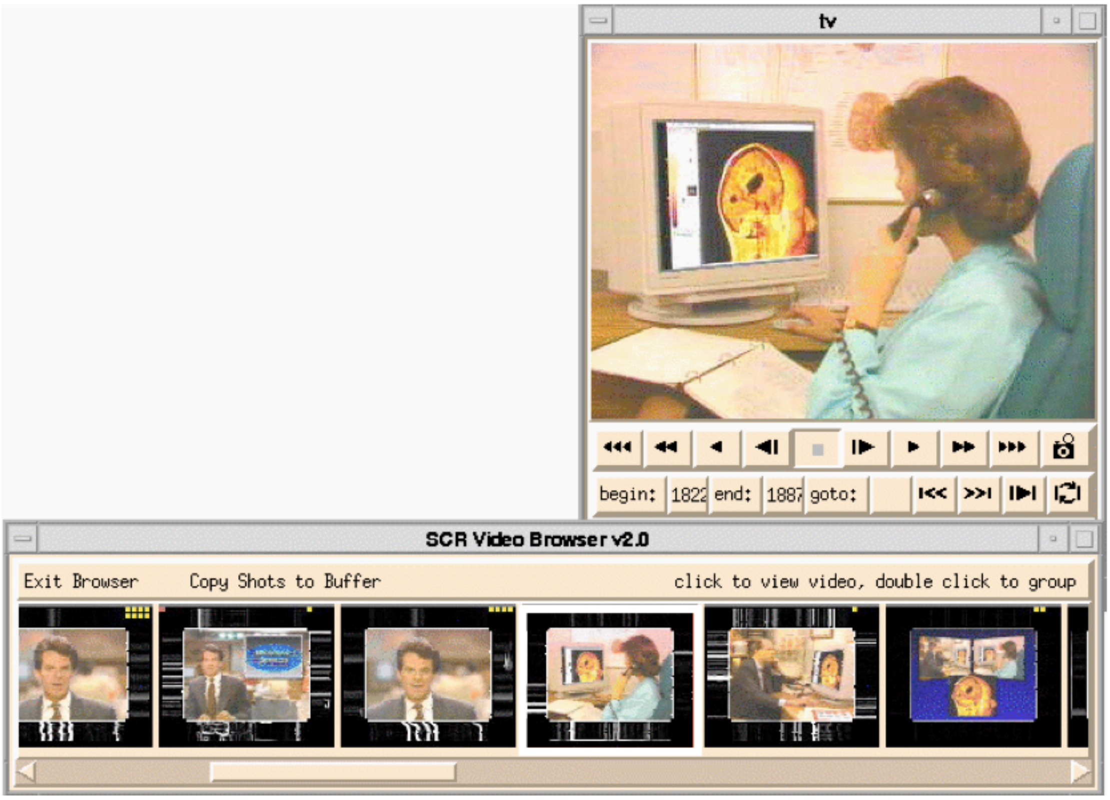


Figure 12: Intelligent browser prototype by [71]

Arman et al. [72] propose a content-based browser for video sequences based on representative frames. These frames are detected using shot boundaries detections, shape and color analyses and a very simple motion analysis. Their browser has advantages over fast forward and rewind while remaining as convenient to use. Using fast forward and rewind the user must view every frame at rapid speeds, missing shots that last a short period, while being forced to watch long lasting and irrelevant shots. With the representative frames they overcome these problems. Figure 13 show the content-based browser, with the row of representative frames on the bottom, the sequence at the point chosen by the user is displayed on top.



begin: 1822 end: 1887 goto:

Figure 13: Content-based browsing based on representative frames by [72]

Haubold and Kender [73] prototype a much more complex user interface. They apply visual segmentation techniques to determine likely changes of topics. Speaker segmentation methods are employed to determine individual user appearances, which are linked to extracted headshots to create a visual speaker index. Videos are augmented with time-aligned filtered keywords and phrases from speech transcripts. Their user interface combines streaming videos, visual, and textual indices for browsing and searching. Figure 14 shows the user interface, where the video summary is displayed as a collection of horizontal tracks, each representing a different modality: thumbnail images, time line, speaker segmentation, visual segmentation, search phases, topic phrases, and content phrases. Based on an evaluation, users find the multimedia retrieval using this user interface effective.

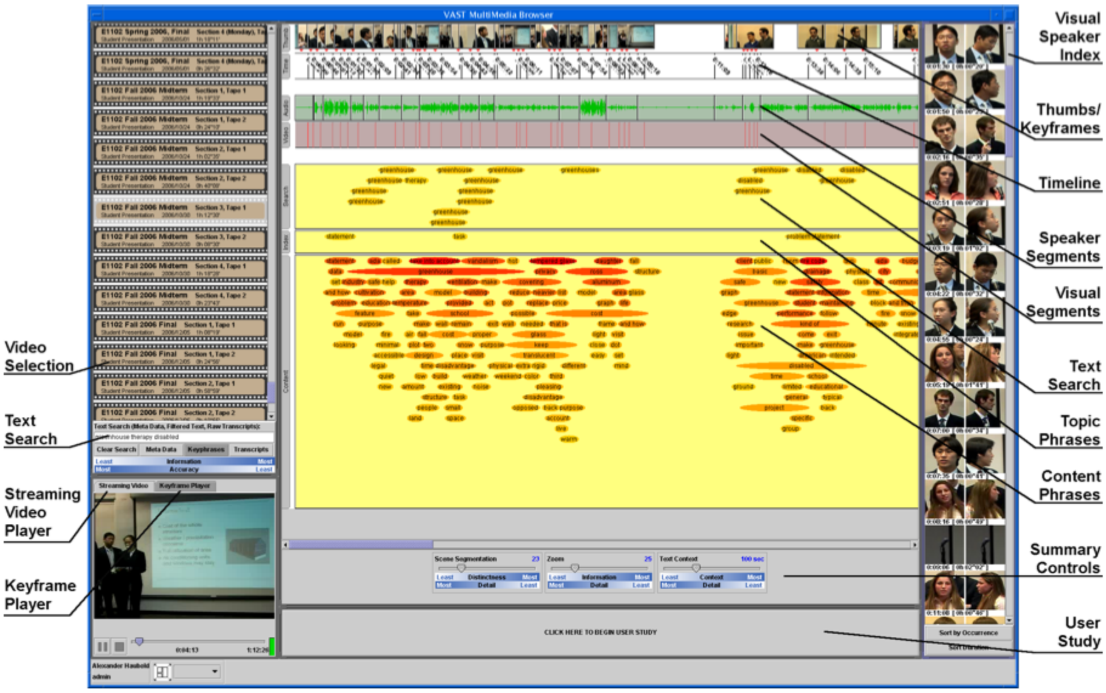


Figure 14: User interface developed by [73]

Heasen et al. [74] developed a user interface based on a user-centered software engineering approach, by involving end users from the beginning of the development process. Their prototype combines an advanced time slider, with a timeline video visualization, shown in Figure 15. A time slider is employed to manipulate the current time of the played video fragment (Figure 15, part A) and to specify an area of interest around this time (Figure 15, part B). The timeline (Figure 15, part C) gives a detailed view on the content in this area of interest.

Their video browser contains several mechanisms to keep an overview on the large amount of information that is visualized in the timeline. Each layer can be maximized/minimized (Figure 15, part D). Content filters (Figure 15, part E) are provided to filter content from the timeline.

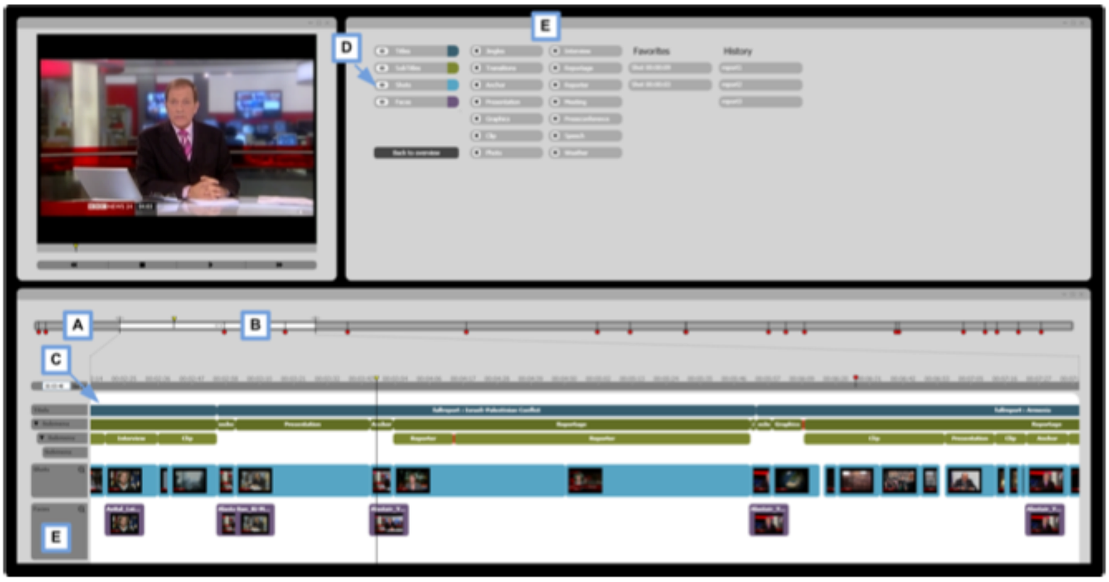


Figure 15: Video browser by [74]

Cheng et al. [75] propose a video interaction they call adaptive fast-forwarding. This user interface helps people quickly browsing the videos with predefined semantic rules. They designed the model around the metaphor of “scenic car driving” in which the driver slows down near areas of interest and speech through unexciting areas. Figure 16 shows the SmartPlayer user interface. In addition to the basic control button, the playback speed is shown at the center of the control panel. A seeker bar is shown near the bottom of the SmartPlayer. In this bar they use scented widgets, which use embedded visualization to enhance the graphical user interface controls. Their visual scent on the video seeker bar is encoded by the amount of saturation on the red color. If a video segment has a relatively high amount of motion, its red color saturation value on the seeker bar will be higher than those of other segments. This information is used to inform the user that the browser will likely slow down when playing this motion-rich video segment.

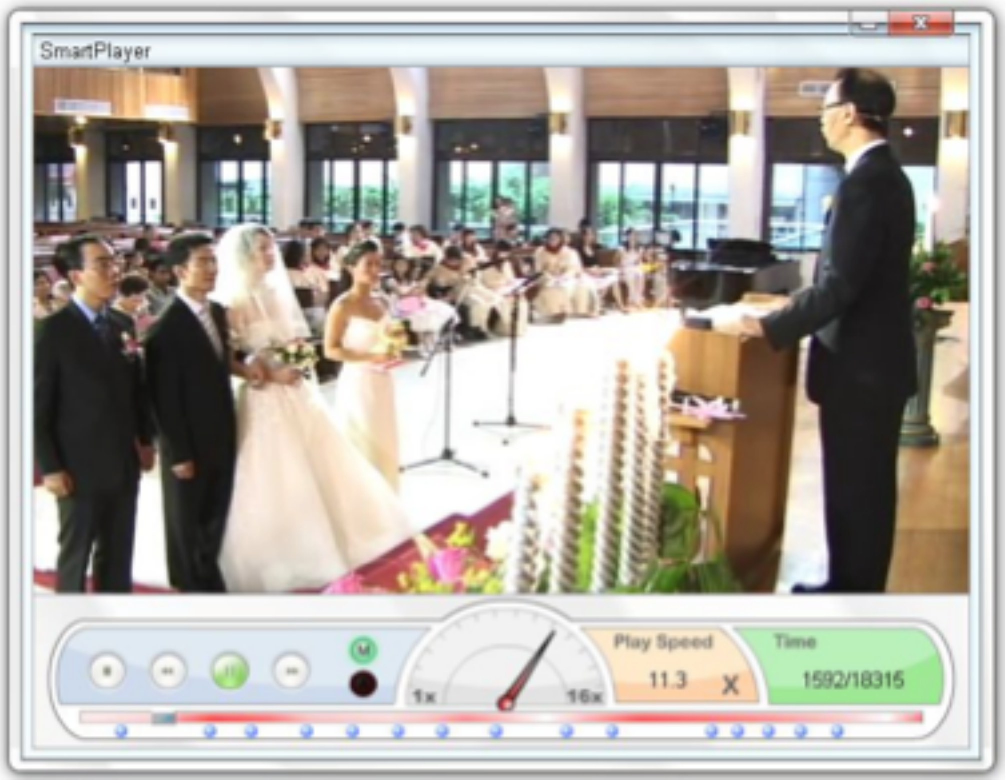


Figure 16: User interface developed by [75]

7.2 Episode Browser

In this section a prototype for an episode browser is proposed. The episode browser is built with HTML5, CSS and JavaScript and can be viewed in supporting web browsers. In this prototype we only used clips from VideoCLEF, as it contains more visual information than the clips of SEMAINE, which show only the face of the experiencer. But the episode browser can be applied to both datasets.

In Figure 17 the episode browser is shown. Traditional playback controls are placed right under the episode. Controls that should help the user navigate the emotional intensity peaks are shown below the traditional playback controls. Users can play a summary based on the emotional intensity peaks. When users play this summary each detected peak is played for 10 seconds until all peaks have been played. Users can also manually navigate through the peaks by using the controls at the bottom middle. Using these controls they can jump from one peak to the next or the previous one. The last interface element we added is shown at the bottom right (“Show Details”). When a user clicks this element more advanced control elements are shown that are described in the next sections.

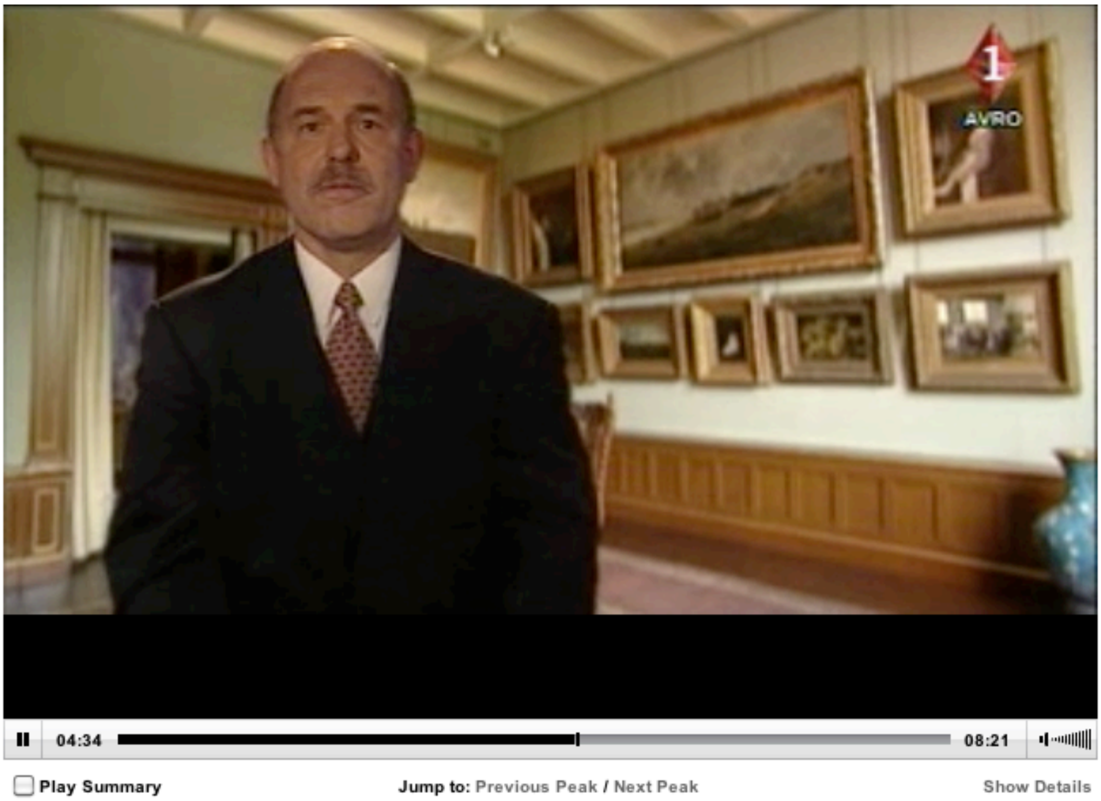


Figure 17: Episode browser, with controls on the bottom

7.2.1 Graph Element

The graph representation is inspired by the work of [72] and can be seen in Figure 18. The probabilities (y axis) for each segment are plotted against the time, (x axis). The graph is displayed time-synchronously with the time slider. Peaks in the graph visualize peaks in emotional intensity. Users can click these peaks to play such a peak.

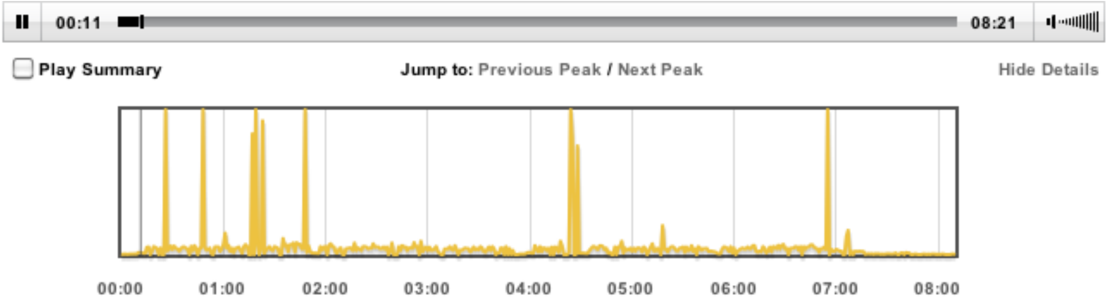


Figure 18: Browsing by graph

7.2.2 Bullets Element

The bullets representation is based on the work of [75], where interesting events are shown in red, while less interesting events have no color. This representation can be seen in Figure 19. The colors and the size of the bullets are based on the probability distribution of the classifier. Bigger bullets reflect higher probabilities, as does the color that changes to red. Segments with a low probability of being a peak are colored green and are also smaller than segments with high probabilities. Just like the graph representation, the bullets representation is displayed time-synchronously with the time slider.

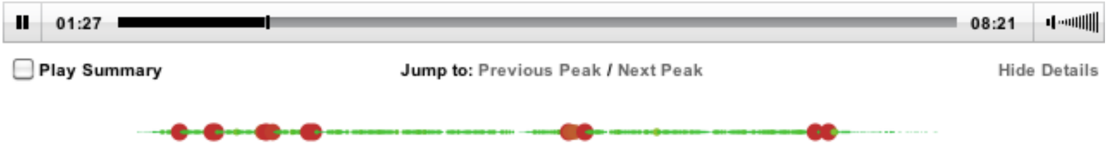


Figure 19: Browsing by bullets

7.2.3 Representative Frames

Our final peak representation is shown in Figure 20, which is based on the work of [73], where representative frames were used to indicate interesting parts of the video. This representation can only be applied on the VideoCLEF dataset since in SEMAINE they recorded only the faces of people, which in this case would not give enough information to the user. Representative frames are based on the peaks in emotional intensity where the first frame of a peak is selected. When a user clicks on a representative frame the browser jumps to the selected peak.

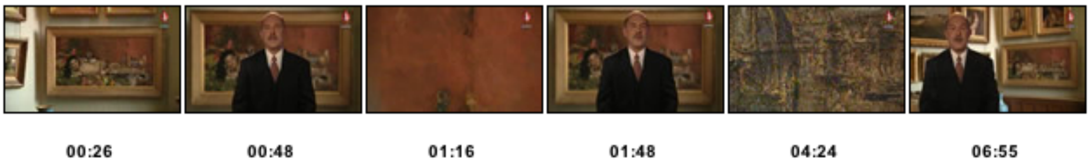
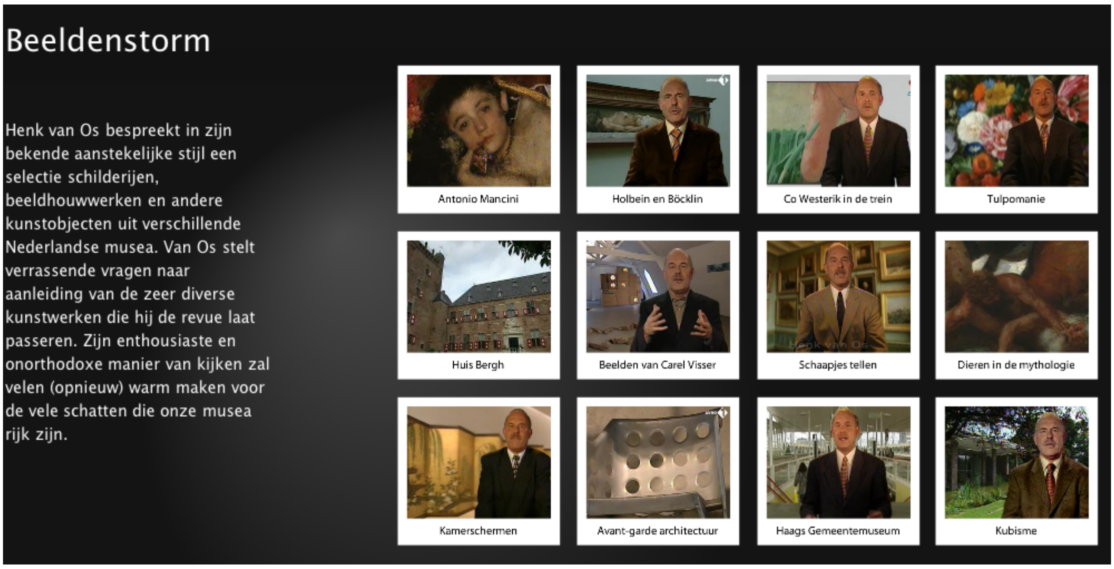


Figure 20: Browsing by key frame

7.3 Collection Browser

Besides the episode browser, a collection browser is also developed. For the collection browser we focus again only on the VideoCLEF dataset. The collection browser is inspired by the “movies rental store” metaphor. Users can look at the front of a movie, reading the backs of movie boxes, going to the next box, etc.

Figure 21 shows the collection browser. On the left the user is provided with some basic information about *Beeldenstorm*. On the right side the episodes of *Beeldenstorm* are shown. Here, we use representative frames like [73] to give users an impression of the episode. These frames are based on the first peak in emotional intensity detected in that episode.

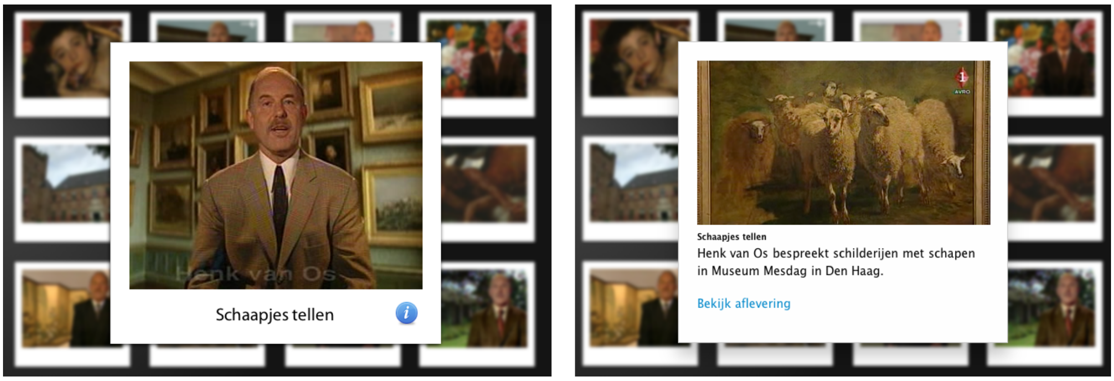


Beeldenstorm

Henk van Os bespreekt in zijn bekende aanstekelijke stijl een selectie schilderijen, beeldhouwwerken en andere kunstobjecten uit verschillende Nederlandse musea. Van Os stelt verrassende vragen naar aanleiding van de zeer diverse kunstwerken die hij de revue laat passeren. Zijn enthousiaste en onorthodoxe manier van kijken zal velen (opnieuw) warm maken voor de vele schatten die onze musea rijk zijn.

Figure 21: Collection browser for the *Beeldenstorm* series

In Figure 22 the metaphor of the "movies rental store" is comes more to life. When users pick an episode this episode moves in front of other episodes and users are presented with the "front-side of the movie box" containing again the representative frame and the title of the episode. When users click the blue *i* the "back-side of the movie box" is shown, containing a summary of the episode of 30 seconds and a short description of the episode, collected from the *Beeldenstorm* website[1]. The summary is based on the three emotional intensity peaks detected by our peak detection approach.



Schaapjes tellen

Henk van Os bespreekt schilderijen met schapen in Museum Mesdag in Den Haag.

Bekijk aflevering

Figure 22: The user interface after a user selects an episode. *Left*: shows what happens when a user selects and episode. *Right*: is shown to the user after he clicked the blue *i* on the left image, the user is presented with a summary based on the emotional intensity peaks plus a little description.

[1] http://www.avro.nl/tv/programmas_a-z/beeldenstorm/

8 Conclusions and Future Work

First of all, the original research questions are stated once more, and the conclusions are added per sub question. After that, some general conclusions about this research are added. Finally some recommendations for future work are presented.

8.1 Conclusions

1. *Is it possible with a one-sided model to capture peaks in emotional intensity in a narrative setting and a conversational setting?*

   Our experimental results confirm that models that use the speech signal only can be trained that easily outperform the random baseline. 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 conversational setting, our models captured strategies intended to shift emotional state, including exaggeration and humor. Although precision and recall values are low, our failure analysis showed that in the narrative setting none of ten false alarms we choose for further examination are implausible as emotional peaks. In the conversational setting, seven out of ten false alarms were also plausible as an emotional peak, showing the potential of the automated emotional peak intensity detector in both settings.

2. *How do lexical and acoustic features contribute to peak detection in these settings?*

   In both settings our experimental results show that both lexical and acoustic features make contributions. In the narrative setting, lexical features outperform acoustic features. The combination of both acoustic and lexical features does not, yield a clear advantage over lexical features alone. In the conversational setting, acoustic features outperform lexical features and also in this setting there is no clear benefit in the combination. The lower performance of the lexical features can be attributed to the conversational style of the sessions and that the word-level time markers were estimated.

   Based on our results we also formulated the hypothesis that evokers depend more on lexical features to create peaks in emotional intensity, while reactions of experiencers are based acoustic features. We were not able to test this hypothesis since our corpora lack either data or annotations to build these models.

3. *Is there a correlation between the emotional state the evoker tries to shift the experiencer to and the peak detection performance?*

   To test if there is a correlation between the emotional state and the peak detection performance we build four models based on the four characters. From our results we see that the highest performance is achieved when users talk to Poppy, who is happy and outgoing. Word usage is characteristic when experiencers talk to Poppy (e.g., "pleasant", "magical" and "wonderful"). The strategies used by the evoker are exaggeration and encouraging the experiencer to tell stories that are reminiscent to certain emotional experiences. Reactions to exaggeration are difficult to detect on a lexical level because of topical variation, which serve, in part, to the account for low peak detection performance in Spike, who is angry and confrontational, and Obadiah, who is depressive. Models based on the Prudence, who is even-tempered and sensible perform better than models based on Spike or Obadiah but not as good as models based on Poppy. However, additional data is required in order to draw conclusions with stronger validity since these conclusions are based on a limited amount of training data per character (approximately 30 minutes).

4. *How can we present these peaks in a useful manner to the end user?*

   To present the peaks in emotional intensity, a number of user interfaces have been developed, which utilize the emotional intensity information detected by our models. These peaks can be helpful in two situations: when browsing within a multimedia object and when browsing through a collection of multimedia objects. For both situations a different user interface was developed, that utilize the emotional information detected by the models to help users find interesting moments in video clips, as well interesting video clips in general. Because of the limited available time we were not able to tests these interfaces with users. Therefore, the answer to this last question remains open.

In this thesis we have presented two studies on emotional intensity peak detection, 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. Emotional intensity peak detection is a challenging task, but for a first attempt the results are encouraging.

8.2 Future Work

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 shifting it for use in a different domain.

During this thesis there was not enough time left to add features from the visual channel. Adding features from this channel could greatly improve peak detection performance. Especially when facial expressions and gestures can be recognized.

In this thesis emotional intensity peaks are detected by using one-sided models. We had to leave out two-sided models since our corpora lack either data or annotations to build these models. It would be interesting to see how well two-sided models perform since they contain knowledge of the “other” in one-sided models.

Finally, since the current implementation of the peak browser has not been tested with users, it should be evaluated with users to test its interface. At the same time, it would also be interesting to evaluate how useful peaks are for users when browsing through video collections.

References

[1] R. Banse and K. Scherer, "Acoustic profiles in vocal emotion expression," *Journal of personality and social psychology,* vol. 70, pp. 614-636, 1996.

[2] A. Hanjalic and L. Xu, "Affective video content representation and modeling," *IEEE Transactions on Multimedia,* vol. 7, pp. 143-154, 2005.

[3] C. Liu*, et al.*, "A framework for flexible summarization of racquet sports video using multiple modalities," *Computer Vision and Image Understanding,* vol. 113, pp. 415-424, 2009.

[4] M. Larson*, et al.*, "Overview of VideoCLEF 2009: New perspectives on speech-based multimedia content enrichment," *Working Notes of CLEF,* 2009.

[5] M. Bradley, "Emotional memory: A dimensional analysis," *Emotions: Essays on emotion theory,* pp. 97-134, 1994.

[6] B. Detenber*, et al.*, "Roll'Em!: The Effects of Picture Motion on Emotional Responses," *Journal of Broadcasting & Electronic Media,* vol. 42, 1998.

[7] M. Greenwald*, et al.*, "Affective judgment and psychophysiological response: Dimensional covariation in the evaluation of pictorial stimuli," *Journal of Psychophysiology,* vol. 3, pp. 51-64, 1989.

[8] R. Dietz and A. Lang, "Affective agents: Effects of agent affect on arousal, attention, liking and learning," 1999.

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

[10] P. Oudeyer, "The production and recognition of emotions in speech: features and algorithms," *International Journal of Human-Computer Studies,* vol. 59, pp. 157-183, 2003.

[11] J. Pennebaker*, et al.*, "Psychological aspects of natural language use: Our words, our selves," *Annual review of psychology,* vol. 54, pp. 547-577, 2003.

[12] R. Campbell and J. Pennebaker, "The secret life of pronouns," *Psychological Science,* vol. 14, p. 60, 2003.

[13] C. Caffi and R. Janney, "Toward a pragmatics of emotive communication* 1," *Journal of Pragmatics,* vol. 22, pp. 325-373, 1994.

[14] D. Sadlier and N. O'Connor, "Event detection in field sports video using audio-visual features and a support vector machine," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 15, pp. 1225–1233, 2005.

[15] A. Smeaton*, et al.*, "Automatically selecting shots for action movie trailers," 2006, p. 238.

[16] J. Kender and B. Yeo, "Video scene segmentation via continuous video coherence," 1998, pp. 367-373.

[17] H. Sundaram and S. Chang, "Determining computable scenes in films and their structures using audio-visual memory models," New York, NY, USA, 2000, pp. 95-104.

[18] J. Nam*, et al.*, "Audio-visual content-based violent scene characterization," in *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on*, 1998, pp. 353-357 vol.1.

[19] G. Iyengar and H. Nock, "Discriminative model fusion for semantic concept detection and annotation in video," 2003, p. 258.

[20] T. Westerveld*, et al.*, "A probabilistic multimedia retrieval model and its evaluation," *EURASIP Journal on Applied Signal Processing,* vol. 2, pp. 186-198, 2003.

[21] A. Amir*, et al.*, "IBM research TRECVID-2003 video retrieval system," 2003.

[22] C. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools and Applications,* vol. 25, pp. 5-35, 2005.

[23] C. Cheng and C. Hsu, "Fusion of audio and motion information on HMM-based highlight extraction for baseball games," *IEEE Transactions on Multimedia,* vol. 8, p. 585, 2006.

[24] C. Snoek*, et al.*, "The MediaMill TRECVID 2004 semantic video search engine," 2004.

[25] Y. Wu*, et al.*, "Optimal multimodal fusion for multimedia data analysis," presented at the Proceedings of the 12th annual ACM international conference on Multimedia, New York, NY, USA, 2004.

[26] C. G. M. Snoek*, et al.*, "Early versus late fusion in semantic video analysis," presented at the Proceedings of the 13th annual ACM international conference on Multimedia, Hilton, Singapore, 2005.

[27] H. Gunes and M. Piccardi, "Affect recognition from face and body: early fusion vs. late fusion," 2005.

[28] Y. Rui*, et al.*, "Automatically extracting highlights for TV Baseball programs," presented at the Proceedings of the eighth ACM international conference on Multimedia, Marina del Rey, California, United States, 2000.

[29] Z. Xiong*, et al.*, "Audio-based highlights extraction from baseball, golf and soccer games in a unified framework," in *Proceedings of the ICASSP Conference*, Hong Kong, China, 2003, pp. 401-404.

[30] C. Chan and G. Jones, "Affect-based indexing and retrieval of films," 2005, p. 430.

[31] R. Cai*, et al.*, "Highlight sound effects detection in audio stream," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, Baltimore, NY, 2003, pp. 37-40.

[32] J. Assfalg*, et al.*, "Soccer highlights detection and recognition using HMMs," in *Proc. of Int'l Conf. on Multimedia and Expo*, Lausanne, Switzerland, 2002, pp. 825-828.

[33] M. Lazarescu*, et al.*, "On the automatic indexing of cricket using camera motion parameters," 2002, pp. 809-813.

[34] S. Nepal*, et al.*, "Automatic detection of'Goal'segments in basketball videos," 2001, pp. 261-269.

[35] B. Li and M. Sezan, "Event detection and summarization in American football broadcast video," 2001, p. 202.

[36] M. Petkovic*, et al.*, "Multi-modal extraction of highlights from TV Formula 1 programs," 2002, pp. 817-820.

[37] E. Kijak*, et al.*, "HMM based structuring of tennis videos using visual and audio cues," 2003.

[38] Y. Gong*, et al.*, "Maximum entropy model-based baseball highlight detection and classification," *Computer Vision and Image Understanding,* vol. 96, pp. 181-199, 2004.

[39] R. Cabasson and A. Divakaran, "Automatic extraction of soccer video highlights using a combination of motion and audio features," in *SPIE Conference on Storage and Retrieval for Media Databases*, Santa Clara, CA, USA, 2003, p. 272.

[40] H. Chen*, et al.*, "Action movies segmentation and summarization based on tempo analysis," New York, NY, USA, 2004, pp. 251-258.

[41] W. Hsu, "Speech audio project report," *Class Project Report,* 2000.

[42] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *Proceedings IEEE Automatic Speech Recognition and Understanding Workshop*, San Juan, Peurto Rico, 2005, pp. 381-385.

[43] M. Schröder*, et al.*, "Towards responsive sensitive artificial listeners," 2008.

[44] J. Weizenbaum, "ELIZA - a computer program for the study of natural language communication between man and machine," *Communications of the ACM,* vol. 9, pp. 36-45, 1966.

[45] M. Valstar*, et al.*, "The SEMAINE corpus of emotionally colored character interactions," *Proceedings International Conference Multimedia & Expo 2010,* to appear.

[46] S. Baron-Cohen*, et al.*, "Mind reading: The interactive guide to emotions," *London: Jessica Kingsley Limited,* 2004.

[47] R. Cowie*, et al.*, "'FEELTRACE': An instrument for recording perceived emotion in real time," 2000.

[48] R. Plutchik, *The psychology and biology of emotion*: Harper Collins College Publishers, 1994.

[49] H. Schlosberg, "A scale for the judgment of facial expressions," *Journal of Experimental Psychology,* vol. 29, pp. 497-510, 1941.

[50] M. Grimm*, et al.*, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication,* vol. 49, pp. 787-800, 2007.

[51] E. Mower*, et al.*, "Evaluating evaluators: A case study in understanding the benefits and pitfalls of multievaluator modeling," 2009.

[52] Z. Zeng*, et al.*, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence,* pp. 39-58, 2008.

[53] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot international,* vol. 5, pp. 341-345, 2001.

[54] R. Picard, *Affective computing*: MIT Press, 2000.

[55] C. Breitenstein*, et al.*, "The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample," *Cognition and Emotion,* vol. 15, pp. 57-79, 2001.

[56] K. Tusing and J. Dillard, "The sounds of dominance," *Human Communication Research,* vol. 26, pp. 148-171.

[57] W. Johnson*, et al.*, "Recognition of emotion from vocal cues," *Archives of General Psychiatry,* vol. 43, pp. 280-283, 1986.

[58] D. Binnenpoorte*, et al.*, "Gender in everyday speech and language: a corpus-based study," 2005.

[59] G. Neff*, et al.*, "Assessing the affective aspect of languaging: the development of software for public relations," *The 52nd Annual Conference of the International Communication Association,* 2002.

[60] A. Van den Bosch*, et al.*, "An efficient memory-based morphosyntactic tagger and parser for Dutch," *Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting,* vol. Leuven, Belgium, pp. 99-114, 2007.

[61] C. Fox, "A stop list for general text," 1989, p. 21.

[62] I. Witten and E. Frank, "Data mining: practical machine learning tools and techniques with Java implementations," *ACM SIGMOD Record,* vol. 31, pp. 76-77, 2002.

[63] S. Russell and P. Norvig, "Artificial intelligence: a modern approach," *New Jersey,* 1995.

[64] J. Quinlan, *C4. 5: programs for machine learning*: Morgan Kaufmann, 2003.

[65] C. Manning and K. Toutanova, "Enriching the knowledge sources used in a maximum entropy part-of-speech tagger," 2000, pp. 63-70.

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

[67] M. Larson*, et al.*, "Exploiting Speech Recognition Transcripts for Narrative Peak Detection in Short-Form Documentaries," in *Proceedings of CLEF*, 2009.

[68] J. Kierkels*, et al.*, "Identification of Narrative Peaks in Clips: Text Features Perform Best."

[69] T. Dobril *, et al.*, "UAIC: Participation in VideoCLEF Task."

[70] H. Blanken*, et al.*, "Multimedia retrieval," *Journal of Electronic Imaging,* vol. 17, pp. 300-302, 2008.

[71] J. Foote*, et al.*, "An intelligent media browser using automatic multimodal analysis," 1998, p. 380.

[72] F. Arman*, et al.*, "Content-based browsing of video sequences," 1994, p. 103.

[73] A. Haubold and J. Kender, "VAST MM: Multimedia browser for presentation video," 2007, p. 48.

[74] M. Heasen*, et al.*, "Visualising Digital Video Libraries for TV Broadcasting Industry: A User-Centred Approach," 2009.

[75] K. Cheng*, et al.*, "SmartPlayer: user-centric video fast-forwarding," 2009, pp. 789-798.

Appendix A – Praat Script

```
form Get arguments
      sentence File
endform

Read from file... 'file$'
Rename... my_wav
dur = Get total duration
print 'To Pitch...''newline$'
To Pitch... 0 75 500
select Sound my_wav
print 'To Intensity...''newline$'
To Intensity... 100 0 yes

t=0
while t < dur
      select Pitch my_wav
      p = Get value at time... 't' Hertz Linear
      select Intensity my_wav
      i = Get value at time... 't' Cubic
      print 't''tab$''p:5''tab$''i:5''newline$'
      t =t + 0.5
endwhile
```

Appendix B - J48 Pruned Tree

```
speech-rate-pause <= 2
|   range-intensity <= 22.88489: no (521.0/4.0)
|   range-intensity > 22.88489
|   |   pitch <= 151.43756
|   |   |   pitch <= 136.116453
|   |   |   |   speech-rate <= 13.240418
|   |   |   |   |   range-intensity <= 23.42649: yes (2.0)
|   |   |   |   |   range-intensity > 23.42649: no (33.0/3.0)
|   |   |   |   speech-rate > 13.240418: yes (2.0)
|   |   |   pitch > 136.116453: yes (4.0)
|   |   pitch > 151.43756: no (23.0)
speech-rate-pause > 2
|   pronouns <= 1
|   |   next-pause <= 19.77
|   |   |   next-pause <= 0.38: no (484.0/30.0)
|   |   |   next-pause > 0.38
|   |   |   |   min-intensity <= 64.81239
|   |   |   |   |   total-words <= 9
|   |   |   |   |   |   pronouns <= 0: no (515.0/131.0)
|   |   |   |   |   |   pronouns > 0
|   |   |   |   |   |   |   intensity <= 67.245117
|   |   |   |   |   |   |   |   previous-pause <= -162.5
|   |   |   |   |   |   |   |   |   stop-words <= 6
|   |   |   |   |   |   |   |   |   |   d-next-pitch <= -32.440142
|   |   |   |   |   |   |   |   |   |   |   nouns <= 0: yes (2.0)
|   |   |   |   |   |   |   |   |   |   |   nouns > 0: no (3.0)
|   |   |   |   |   |   |   |   |   |   d-next-pitch > -32.440142: yes (8.0)
|   |   |   |   |   |   |   |   |   stop-words > 6: no (3.0)
|   |   |   |   |   |   |   |   previous-pause > -162.5
|   |   |   |   |   |   |   |   |   total-words <= 7: no (27.0)
|   |   |   |   |   |   |   |   |   total-words > 7
|   |   |   |   |   |   |   |   |   |   speech-rate-pause <= 10.666667
|   |   |   |   |   |   |   |   |   |   |   next-pause <= 14.75
|   |   |   |   |   |   |   |   |   |   |   |   verbs <= 0: no (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   verbs > 0
|   |   |   |   |   |   |   |   |   |   |   |   |   speech-rate-pause <= 8.666667: no (2.0)
|   |   |   |   |   |   |   |   |   |   |   |   |   speech-rate-pause > 8.666667: yes (9.0/1.0)
|   |   |   |   |   |   |   |   |   |   |   next-pause > 14.75: no (6.0)
|   |   |   |   |   |   |   |   |   |   speech-rate-pause > 10.666667: no (11.0)
|   |   |   |   |   |   |   intensity > 67.245117
|   |   |   |   |   |   |   |   d-next-pitch <= 18.453969: yes (9.0)
|   |   |   |   |   |   |   |   d-next-pitch > 18.453969
|   |   |   |   |   |   |   |   |   speech-rate <= 9.961686: yes (2.0)
|   |   |   |   |   |   |   |   |   speech-rate > 9.961686: no (4.0)
|   |   |   |   |   total-words > 9: no (60.0/6.0)
|   |   |   |   min-intensity > 64.81239
|   |   |   |   |   previous-pause <= -25.84
|   |   |   |   |   |   d-intensity <= 4.795832
|   |   |   |   |   |   |   nouns <= 1
|   |   |   |   |   |   |   |   total-words <= 4: no (2.0)
|   |   |   |   |   |   |   |   total-words > 4: yes (5.0)
|   |   |   |   |   |   |   nouns > 1: no (7.0/1.0)
|   |   |   |   |   |   d-intensity > 4.795832: no (7.0)
|   |   |   |   |   previous-pause > -25.84: no (85.0/1.0)
|   |   next-pause > 19.77
|   |   |   adjectives <= 2: no (5207.0/557.0)
|   |   |   adjectives > 2
|   |   |   |   d-next-intensity <= 11.103635
|   |   |   |   |   total-words <= 10
|   |   |   |   |   |   non-stop-words <= 2
|   |   |   |   |   |   |   d-pitch <= 12.845233
|   |   |   |   |   |   |   |   range-pitch <= 18.55075
|   |   |   |   |   |   |   |   |   speech-rate-pause <= 11: no (13.0/1.0)
|   |   |   |   |   |   |   |   |   speech-rate-pause > 11: yes (3.0)
|   |   |   |   |   |   |   |   range-pitch > 18.55075: yes (5.0/1.0)
|   |   |   |   |   |   |   d-pitch > 12.845233: no (45.0/2.0)
|   |   |   |   |   |   non-stop-words > 2: no (43.0/1.0)
|   |   |   |   |   total-words > 10
|   |   |   |   |   |   speech-rate-pause <= 13.666667: no (3.0)
|   |   |   |   |   |   speech-rate-pause > 13.666667: yes (5.0/1.0)
|   |   |   |   d-next-intensity > 11.103635
|   |   |   |   |   non-stop-words <= 1: no (4.0)
|   |   |   |   |   non-stop-words > 1: yes (10.0/1.0)
|   pronouns > 1
|   |   total-words <= 6
|   |   |   nouns <= 0
|   |   |   |   verbs <= 1
|   |   |   |   |   speech-rate <= 6.646526: yes (2.0)
|   |   |   |   |   speech-rate > 6.646526: no (3.0)
|   |   |   |   verbs > 1: yes (4.0)
|   |   |   nouns > 0: no (3.0)
|   |   total-words > 6: no (164.0/27.01
```

Appendix C – VideoCLEF Results

Naïve Bayes

Table 25: Naïve Bayes results (non-overlapping segments)

Window Length	Point-based	Peak-based "Personal Peaks"	Peak-based "Pair Peaks"	Peak-based "General Peaks"
3	54	40	11	3
4	52	33	15	5
5	62	47	15	3
6	53	40	13	3
7	63	41	17	3
8	53	38	11	2
9	56	43	13	2
10	58	44	13	3

Table 26: Naïve Bayes results (overlapping segments)

Window Length	Point-based	Peak-based "Personal Peaks"	Peak-based "Pair Peaks"	Peak-based "General Peaks"
3	49	36	10	3
4	57	40	12	3
5	61	40	18	3
6	59	36	16	3
7	64	42	19	4
8	61	41	16	3
9	57	39	15	2
10	58	39	15	3

Table 27: Naïve Bayes results audio features (non overlapping segments)

Window Length	Point-based	Peak-based "Personal Peaks"	Peak-based "Pair Peaks"	Peak-based "General Peaks"
3	46	31	12	3
4	45	35	13	1
5	48	36	10	4
6	38	34	5	0
7	50	35	13	3
8	49	37	10	2
9	51	38	10	2
10	52	38	12	3

Table 28: Naïve Bayes results audio features (overlapping segments)

Window Length	Point-based	Peak-based "Personal Peaks"	Peak-based "Pair Peaks"	Peak-based "General Peaks"
3	46	34	9	1
4	19	17	2	0
5	41	28	11	1
6	38	22	11	1
7	33	25	9	1
8	41	30	7	2
9	46	32	11	2
10	36	29	6	1

Table 29: Naïve Bayes results text features (non-overlapping segments)

Window Length	Point-based	Peak-based “Personal Peaks”	Peak-based “Pair Peaks”	Peak-based “General Peaks”
3	46	34	11	3
4	48	37	11	2
5	59	44	13	3
6	48	37	11	2
7	61	41	15	4
8	57	38	14	3
9	65	47	17	3
10	62	43	13	5

Table 30: Naïve Bayes results text features (overlapping segments)

Window Length	Point-based	Peak-based “Personal Peaks”	Peak-based “Pair Peaks”	Peak-based “General Peaks”
3	57	41	13	3
4	53	38	13	3
5	50	34	13	2
6	59	38	17	4
7	55	39	14	2
8	54	38	12	2
9	54	39	14	4
10	51	34	12	5

J48

Table 31: J48 results (non-overlapping segments)

Window Length	Point-based	Peak-based "Personal Peaks"	Peak-based "Pair Peaks"	Peak-based "General Peaks"
3	35	26	7	3
4	34	24	5	1
5	45	35	13	1
6	38	34	5	0
7	48	36	10	4
8	51	35	14	4
9	49	37	10	2
10	50	35	13	3

Table 32: J48 results (overlapping segments)

Window Length	Point-based	Peak-based "Personal Peaks"	Peak-based "Pair Peaks"	Peak-based "General Peaks"
3	69	39	16	5
4	63	44	15	4
5	69	45	20	6
6	71	42	19	6
7	74	46	21	7
8	74	48	22	7
9	73	44	20	7
10	75	49	22	8

Table 33: J48 results audio features (non-overlapping segments)

Window Length	Point-based	Peak-based “Personal Peaks”	Peak-based “Pair Peaks”	Peak-based “General Peaks”
3	25	16	6	2
4	26	16	6	2
5	36	23	9	2
6	36	25	9	2
7	37	22	9	2
8	38	26	8	2
9	36	25	7	1
10	36	25	7	1

Table 34: J48 results audio features (overlapping segments)

Window Length	Point-based	Peak-based “Personal Peaks”	Peak-based “Pair Peaks”	Peak-based “General Peaks”
3	23	13	5	1
4	34	24	7	2
5	32	22	8	2
6	34	24	7	1
7	48	35	10	2
8	43	32	11	3
9	50	36	11	3
10	47	34	11	2

Table 35: J48 results text features (non-overlapping segments)

Window Length	Point-based	Peak-based "Personal Peaks"	Peak-based "Pair Peaks"	Peak-based "General Peaks"
3	23	15	5	2
4	36	23	10	2
5	34	24	8	2
6	32	21	7	1
7	35	26	7	3
8	37	24	8	2
9	44	26	13	5
10	38	26	8	2

Table 36: J48 results text features (overlapping segments)

Window Length	Point-based	Peak-based "Personal Peaks"	Peak-based "Pair Peaks"	Peak-based "General Peaks"
3	62	40	16	6
4	58	37	13	6
5	71	45	18	7
6	71	50	18	5
7	69	45	19	6
8	73	49	19	8
9	73	52	18	5
10	77	52	21	9

Appendix D – SEMAINE Results

Naïve Bayes

Table 37: Results Naïve Bayes non-overlapping

Window Length	All Features	Audio Features	Text Features
3	14	13	9
4	16	16	11
5	16	14	15
6	17	14	14
7	18	17	13
8	16	17	11
9	17	18	12
10	19	16	14

Table 38: Results Naïve Bayes classifier overlapping segments

Window Length	All Features	Audio Features	Text Features
3	13	13	10
4	16	16	11
5	15	14	12
6	17	15	14
7	20	19	15
8	18	19	14
9	19	19	14
10	19	20	14

J48

Table 39: Results J48 classifier non-overlapping segments

Window Length	All Features	Audio Features	Text Features
3	5	5	4
4	3	3	3
5	4	6	4
6	8	6	5
7	5	5	3
8	3	5	7
9	3	6	7
10	7	4	3

Table 40: Results J48 classifier overlapping segments

Window Length	All Features	Audio Features	Text Features
3	8	9	10
4	10	10	11
5	9	9	11
6	10	9	11
7	6	13	9
8	11	10	5
9	10	13	9
10	12	12	9