Exploiting Speech Recognition Transcripts for Narrative Peak Detection in Short-Form Documentaries

Martha Larson[1] Bart Jochems[2] Ewine Smits[1] Roeland Ordelman[2]

[1]Mediamatics, Delft University of Technology, Netherlands
[2]Human Media Interaction, University of Twente, Netherlands
{m.a.larson,e.a.p.smits}@tudelft.nl, {b.e.h.jochems@student,ordelman@ewi}.utwente.nl

Abstract

Narrative peaks are points at which the viewer perceives a spike in the level of dramatic tension within the narrative flow of a video. This paper reports on four approaches to narrative peak detection in television documentaries that were developed by a joint team consisting of members from Delft University of Technology and the University of Twente within the framework of the VideoCLEF 2009 Affect Detection task. The approaches make use of speech recognition transcripts and seek to exploit various sources of evidence in order to automatically identify narrative peaks. These sources include speaker style (word choice), stylistic devices (use of repetitions), strategies strengthening viewers' feelings of involvement (direct audience address) and emotional speech. These approaches are compared to a challenging baseline that predicts the presence of narrative peaks at fixed points in the video, presumed to be dictated by natural narrative rhythm or production convention. Two approaches are tied in delivering top narrative peak detection results. One uses counts of first and second person pronouns to identify points in the video where viewers feel most directly involved. The other uses affective word ratings to calculate scores reflecting emotional language.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing

General Terms

Measurement, Performance, Experimentation

Keywords

Spoken content, Narrative, Dramatic tension, Speech recognition transcripts, Short-form documentaries, Dutch language

1 Introduction

While watching video content, viewers feel fluctuations in their emotional response that can be attributed to their perception of changes in the level of dramatic tension. 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 viewer-perceived dramatic tension. In the case of sports, tension spikes 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, dramatic tension is carefully crafted into the content by a team including scriptwriters, performers, special effects 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 refer to that sequence as a *narrative* and to high points in the dramatic tension within that narrative as *narrative peaks*. We are interested in investigating a third case of video content, namely television documentaries. We consider documentaries to be a form of "edu-tainment," whose purpose is both to inform and entertain the audience. The approaches described and tested here have been developed in order to detect narrative peaks within documentary videos.

Our work differs in an important respect from previous work in the domains of sports and movies. Dramatic tension in documentaries is never completely spontaneous – the narrative curve follows a previously laid out plan, for example a script or an outline, that is carried out during the process of production. However, dramatic tension is characteristically less tightly controlled in a documentary than it would be in a movie. In a movie, the entire content is subordinated to the plot, whereas a documentary may follow one or more story lines, but it simultaneously pursues the goal of providing the viewer with factual subject matter. Because of these differences, we chose to dedicate separate and specific attention to the affective analysis of documentaries and in particular to the automatic detection of narrative peaks.

This area of investigation is quite challenging since fluctuations in dramatic tension in television documentaries are not associated with conventionalized events. If an event is a conventional trigger, a broad spectrum of viewers will agree about its contribution to the drama of the video content – think of goals in the game of soccer or a kiss in a romantic comedy. The subtleness with which narrative peaks manifest themselves in video documentaries makes the task challenging with respect to the difficulty of both automatically detecting such peaks and also evaluating the detection algorithm. Our interest is contextualized within the broader goal of automatic prediction of topic-independent viewer preference. Given two videos with comparable informational content, viewers will often decide to choose to watch one over the other. Our ultimate research aim is to explore the contribution that analysis of affective aspects of video content can make to the automatic prediction of viewer preference.

This paper reports on joint work carried out by research groups at two universities in the Netherlands, Delft University of Technology[1] and the University of Twente, on the Affect Detection task of the VideoCLEF[2] track of the 2009 Cross-Language Evaluation Forum (CLEF)[3] benchmark evaluations. The Affect Detection task involves automatically identifying narrative peaks in short-form documentaries. In the rest of this paper, we first give a brief description of the data and the task. Then, we present the approach that we took to the task and give the details of the algorithms used in each of the five runs that we submitted. We report the results achieved by these runs and then conclude with a summary and outlook.

2 Experimental Setup

2.1 Data Set and Task Definition

The data set for the VideoCLEF 2009 Affect Detection task consisted of 45 episodes from the Dutch-language short-form documentary series called *Beeldenstorm* (in English, 'Iconoclasm'). The series treats topics in the visual arts, integrating elements from history, culture and current events. *Beeldenstorm* is hosted by Prof. Henk van Os, known not only for his art expertise, but also for his narrative ability. Henk van Os is highly acclaimed and appreciated in the Netherlands, where he has established his ability to appeal to a broad audience.[4]

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. 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

[1] Delft University of Technology and Dublin City University are the coordinators of VideoCLEF

[2] http://www.cdvp.dcu.ie/VideoCLEF/

[3] http://www.clef-campaign.org/

[4] http://www.avro.nl/tv/programmas_az/beeldenstorm/

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.

Ground truth was created for the *Beeldenstorm* by a team of assessors who speak Dutch natively or at an advanced level. The assessors were told that the *Beeldenstorm* series is known to contain humorous and moving moments and told that they could use that information to formulate an opinion of what constitutes a narrative peak. They were asked to mark the three points in the video where their perception of the level of dramatic tension reached the highest peaks. Peaks were required to be a maximum of ten seconds in length.

For the Affect Detection task of VideoCLEF 2009, task participants were supplied with an *example set* containing five *Beeldenstorm* episodes in which example narrative peaks had been identified by a human assessor. On the basis of their observations and generalizations concerning the peaks marked in the example set, the task participants designed algorithms capable of automatically detecting similar peaks in the *test set*. The test set contained 45 videos and was mutually exclusive with the example set. Participants were required to identify the three highest peaks in each episode. Up to five different runs (i.e., system outputs created according to different experimental conditions) could be submitted. Further details about the data set and the Affect Detection task for VideoCLEF 2009 can be found in the track overview paper [4]. Participants were provided with additional resources accompanying the test data, including transcripts generated by an automatic speech recognition system [3]. Our approaches, described in the next section, focus on exploiting the contents of the speech transcripts for the purpose of automatically detecting narrative peaks.

2.2 Narrative Peak Detection Approaches

Our approaches consist of a sophisticated baseline and four other techniques for using speech recognition transcripts to automatically detect narrative peaks. We describe each algorithm in turn.

2.2.1 Fixing Time Points (duotu09fix)

Our baseline approach **duotu09fix**[5] hypothesizes *fixed time points* for three narrative peaks in each episode. These points are completely independent of episode content and are the same for every episode. This approach attempts to exploit regularities that exist in the narrative structure of every episode of a documentary series as a result of production conventions or of general documentary structure (i.e., a documentary consists or an opening, a body and a conclusion). We chose this approach in order to establish a challenging baseline against which our speech-transcript-based peak detection algorithms can be compared. In order to choose the three fixed time points we analyzed the peak positions in the example set. In the examples, the midpoint of the first peak occurred between 28 secs and 1 min 6 secs after the start of the video. The midpoint of the final peak occurred between 6 mins 42 secs and 7 mins 40 secs into the video. We fixed a peak at the average position of the initial peak (44 secs) and the final peak (7 mins 9 secs). We added a third located at the average midpoint of the episode: 3 mins 40 secs. The fact that four of the five example episodes have a peak within 10 seconds of this point confirmed that we had made a good choice for the third fixed point peak.

2.2.2 Counting Indicator Words (duotu09ind)

We viewed the example videos and examined the words that were spoken during the narrative peaks that the assessor had marked in these videos. We formulated the hypothesis that the speaker applies a narrow range of strategies for creating narrative peaks in the documentary. These strategies might be reflected in a relatively limited vocabulary of words that could be used as indicators in order to predict the position of narrative peaks.

We compiled a list of narrative peak indicators by analyzing the words spoken during each of the example peaks and compiled a list of words and word-stems that seemed relatively independent of the topic at the point in the video and which could be plausibly characteristic of the general word use of the speaker during peaks. The indicator words selected are listed in Table 1. It is noteworthy that most of these words are adjectives or adverbs and that they have a basic positive or negative meaning, or they serve as

[5] textbfduotu is an acronym indicating the combined efforts of Delft University of Technology and the University of Twente

Table 1: List of narrative peak indicators consisting of words and word stems (marked with *) selected from example episodes for use in **duotu09ind**

Indicator word or stem	**English form**
helemaal	entirely
eigenlijk*	actually
verkeerd*	wrong
werkelijk*	actually
gelukkig*	happy
mooi*	beautiful
grappig*	funny
ontzettend*	horribly
absolu(u)t*	absolute
fout*	wrong
genial*	ingenious
echt	true
wonder*	extraordinary
belangrijk*	important
maar	but
goed	good
(na)tuurlijk	naturally

an intensifier. The word *maar*, 'but', appears to be an exception to this generalization. We included this word because it occurred in 20% of the peaks in the example set. 'But' is a lexical item frequently used to indicate a contrast with previously established state of knowledge or expectations. We hypothesize that its importance is related to the suspense introduced by statements that contrast with established knowledge or viewer expectations.

The **duotu09ind** algorithm detects narrative peaks using the following sequence of steps. First, a set of all possible peak candidates was established by moving a 10-second sliding window over the speech recognition transcripts, advancing the window by one word at each step. Each peak candidate is maximally 10 seconds in length, but can be shorter if the speech in the window lasts for less than the 10-second duration of the window. Peak candidates of less than three seconds in length are discarded. Then, the peak candidates are ranked with respect to the raw count of the indicator words (cf. Table 1) that they contain. The size limitation of the sliding window already introduces a normalizing effect and for this reason we do not undertake further normalization of the raw counts. Finally, peak candidates are chosen from the ranked list, starting at the top, until a total of three peaks has been selected. If a candidate has a midpoint that falls within eight seconds of the midpoint of a previously selected candidate occurring in the list, that candidate is discarded and the next candidate from the list is considered instead.

2.2.3 Counting Word Repetitions (duotu09rep)

Analysis of the word distributions in the example set suggested that repetition may be a stylistic device that is deployed to create peaks. Particular examples of the use of repetition during narrative peaks in the example episodes include, *...kunsttulpen mooie kunsttulpen...* ('...artificial-tulips, beautiful artificial-tulips...'),[6] *...het is werkelijk een ervaring, een ervaring van stilte...* ('...it is really an experience, an experience of tranquility...'),[7] and *...wordt belangrijk, is altijd belangrijk geweest...* ('...will be important, has always been important...').[8] We do not attempt to measure repetition of phrases or of morphologically related

[6] from *Beeldenstorm* episode *Tulpomanie*, 'Tulip mania'
[7] from *Beeldenstorm* episode *Rust bij Rothko*, 'Peace with Rothko'
[8] from *Beeldenstorm* episode *Maria Magdalena*, 'Mary Magdalene'

words, but rather assume that counting repeated word forms will yield an adequate indicator or places in the documentary where repetition is being applied as a stylistic device.

The **duotu09rep** algorithm uses the same list of peak candidates described in the previous section in the explanation of **duotu09ind**. The peak candidates are ranked by the number of occurrences they contain of words that occur multiple times. In order to eliminate the impact of function words, stop word removal is performed before the peak candidates are scored. Three peaks are selected starting from the top of the ranked list of peak candidates, using the same procedure as was described above.

2.2.4 Counting First and Second Person Pronouns (duotu09pro)

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. The **duotu09pro** approach identifies 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.

In the **duotu09pro** approach we use second person 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 narrative. Cases of narrative 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...').[9] 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 narrative 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-Holthuysen Museum, that's where you'll (informal) find what I call *total design*.')[10] 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.

The **duotu09pro** algorithm uses the same list of peak candidates and the same method of choosing from the ranked candidate lists that was used in **duotu09ind** and **duotu09rep**. For **duotu09pro**, the candidates are ranked according to the raw count of first and second person pronominal forms that they contain. Again, no normalization was applied to the raw count. It should also be noted that in this case no stop word removal was applied since first and second person pronouns are themselves function words and are included in standard formulations of stop word lists.

2.2.5 Calculating Affective Ratings (duotu09rat)

Our final approach to narrative peak detection is based on the hypothesis that dramatic tension rises when the speaker in the video uses speech made vivid by emotion. We conjecture that narrative peaks contain more emotion than other parts of the narrative. Human speech is an important conduit for the communication of emotions. 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 words 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 [6].

The **duotu09rat** approach uses an affective rating score that is calculated in a straightfoward manner using known affective levels of words in order to identify narrative peaks. The approach makes use of Whissell's Dictionary of Affect in Language as deployed in the implementation of [5], which is available

[9] from *Beeldenstorm* episode *Tulpomanie*, 'Tulip mania'

[10] from *Beeldenstorm* episode *Leven met kunst*, 'Living with art'

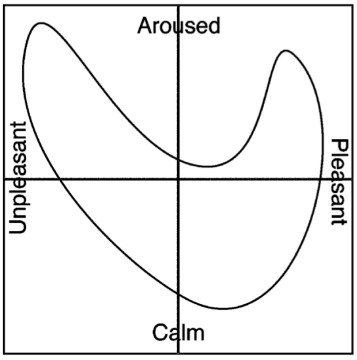

Figure 1: Illustration of the 2D Emotion Space, from [1]

online.[11] This dictionary of words and scores focuses on the scales of pleasantness and arousal levels. The scales are alternately called evaluation and activation. Dietz and Lang [1] transformed these two scales to the two-dimensional emotion space depicted in Figure 1. Under our approach, narrative 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 scores into an overall affective word score. In order to apply the dictionary, we first translate the Dutch-language speech recognition transcripts into English using the Google Language API.[12]

The **duotu09rat** algorithm uses the same list of peak candidates used in **duotu09ind**, **duotu09rep** and **duotu09pro**. Candidates are ranked according to the average affective word score of the words that they contain. Words that are not contained in the dictionary are excluded from the calculation. Selection of peaks proceeds as in the other approaches with the exception of the fact the peak proximity condition was set to be more stringent. Edges of peaks are required to be 4 secs apart from each other. The imposition of the more stringent condition reflects an incidental difference in the experimental set up and does not represent an optimized value.

3 Experimental Results

We tested our five experimental approaches on the 45 videos in the test set. Evaluation of results was carried out by comparing the peak positions hypothesized by each experimental system with peak positions that were set by human assessors. In total, three assessors viewed each of the test videos and set peaks at the three points where he or she felt most highly affected by narrative tension created by the video content. In total the assessors identified 293 distinct narrative peaks in the 45 test 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 could never exceed the specified peak length of ten seconds.

Two methods of scoring the experiments were applied, the *point-based approach* and the *peak-based approach*. Under point-based scoring, a peak hypothesis scores a point for each assessor who selected a reference peak that is within eight seconds of that hypothesis peak. The total number of points returned by the run is the reported run score. A single episode can earn a run between three points (assessors chose completely different peaks) and nine points (assessors all chose the same peaks). In reality, no episode however, falls at either of these extremes. The distribution of the peaks in the files is such that a perfect run would earn 246 points. Under peak-based scoring, the total number of correct peaks is reported as the run score. Three different types of reference peaks are defined for peak-based scoring. The difference is related to the number of assessors required to agree for a point in the video to be counted as a peak. Of these 293

[11] http://technology.calumet.purdue.edu/met/gneff/Publications/ica02/affectdictionary.html
[12] http://code.google.com/intl/nl/apis/ajaxlanguage/

Table 2: Narrative peak detection results

measure	duotu09fix	duotu09ind	duotu09rep	duotu09pro	duotu09rat
point-based	47	55	30	**63**	**63**
peak-based $>$ 1 assessor ("personal peaks")	28	38	21	**44**	37
peak-based $>$ 2 assessors ("pair peaks")	8	12	7	**17**	20
peak-based $>$ 3 assessors ("general peaks")	4	2	0	4	**5**

total peaks identified, 203 peaks are "personal peaks" (peaks identified by only one assessor), 90 are "pair peaks" (peaks that are identified by at least two assessors) and 22 are "general peaks" (peaks upon which all three assessors agreed). Peak-based scores are reported separately for each of these types of peaks. A summary of the results of the evaluation of our five approaches is given in Table2.

From these results it can be seen that **duotu09pro**, the approach that counted first and second person pronouns, and **duotu09rat**, the approach that made use of affective word scores are the best performing approaches. The approach relying on a list of peak indicator words, i.e., **duotu09ind**, performed surprisingly well considering that the list was formulated on the basis of a very limited number of examples.

It should be kept in mind, that the performance of a random classifier on the narrative peak detection task reaches a relatively high level since the videos are relatively short. 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." In light of these statistics, the approach **duotu09rep**, which counted use of repeated words, deserves further comment. This approach failed to achieve the performance level of the random baseline detector, which indicates that repetitions, as they are counted by our implementation of the algorithm, actually are a negative indicator for the existence of a peak. We believe that this result may be due to the fact that assessors tend not to set peaks at places where there might be disfluencies or unintentional repetitions. The speech recognition transcripts contain a high level of noise and it is conceivable that this noise contributes to creating word repetitions where none existed in the original speech. Such an effect could further prevent stylistic repetition from being effectively exploited for the purpose of peak detection.

4 Conclusion and Outlook

We have proposed five approaches to the automatic detection of narrative peaks in short-form documentaries and have evaluated these approaches within the framework of the VideoCLEF 2009 Affect Detection task, which uses a test set consisting of episodes from the Dutch language documentary on the visual arts called *Beeldenstorm*. Our proposed approaches exploit speech recognition transcripts. The two most successful algorithms are based on the idea that narrative peaks are perceived where particularly emotional speech is being used (**duotu09rat**) or when the viewer feels specifically addressed by or involved in the video (**duotu09pro**). These two approaches easily beat both the random baseline and also a challenging baseline approach hypothesizing narrative peaks at set positions in the video. Approaches based on capturing speaking style, either by using a set of indicator words typical for the speaker, or by trying to determine where repetition is being used as a stylistic device, proved less helpful. However, the experiments reported here are not extensive enough to exclude the possibility that they would perform well given a different implementation.

Future work will involve returning to many of the questions opened here, for example, while selecting peak-indicator words, we noticed that contrasts introduced by the word 'but' appear to often be associated with narrative peaks. Stylistic devices in addition to repetition, for example, use of questions, could also prove to be helpful. Under our approach, peak candidates are represented by their spoken content. We

would also like to investigate the enrichment of the representations of peak candidates using words derived from surrounding regions in the speech transcripts or from an appropriate external text collection. Finally, we intend to develop peak detection methods based on the combination of information sources, in particular, exploring whether using pronoun occurrence based information can provide enhancement to affect based rating.

5 Acknowledgements

The research leading to these results was carried out to a substantial degree within the PetaMedia Network of Excellence and has received funding from the European Commission's 7th Framework Program under grant agreement no. 216444.

References

[1] Richard Dietz and Annie Lang. Aefective agents: Effects of agent affect on arousal, attention, liking and learning. In *Proceedings of the Third Annual Cognitive Technology Conference*, 1999.

[2] Alan Hanjalic and Li-Qun Xu. Affective video content representation and modeling. *Multimedia, IEEE Transactions on*, 7(1):143–154, Feb. 2005.

[3] Marijn Huijbregts, Roeland Ordelman, and Franciska de Jong. Annotation of heterogeneous multimedia content using automatic speech recognition. In *Proceedings of SAMT*, 2007.

[4] Martha Larson, Eamonn Newman, and Gareth Jones. Overview of VideoCLEF 2009: New perspectives on speech-based multimedia content enrichment. In Francesca Borri, Alessandro Nardi, and Carol Peters, editors, *Working Notes of CLEF 2009*, September 2009.

[5] Gregory Neff, Bonita Neff, and Paul Crandon. Assessing the affective aspect of languaging:the development of software for public relations. In *The 52nd Annual Conference of the International Communication Association*, July 2002.

[6] Robert Plutchik. *The Psychology and Biology of Emotion*. New York: HarperCollins, 1994.