Using Artistic Markers and Speaker Identification for Narrative-Theme Navigation of Seinfeld Episodes

Gerald Friedland, Luke Gottlieb, and Adam Janin
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704-1198
[fractor|luke|janin]@icsi.berkeley.edu

Abstract

This article describes a system to navigate Seinfeld episodes based on acoustic event detection and speaker identification of the audio track and subsequent inference of narrative themes based on genre-specific production rules. The system distinguishes laughter, music, and other noise as well as speech segments. Speech segments are then identified against pre-trained speaker models. Given this segmentation and the artistic production rules that underlie the "situation comedy" genre and Seinfeld in particular, the system enables a user to browse an episode by scene, punchline, and dialog segments. The themes can be filtered by the main actors, e.g. the user can choose to see only punchlines by Jerry and Kramer. Based on the length of the laughter, the top-5 punchlines are identified and presented to the user. The segmentation is then presented in an Applet-based graphical video browser that is intended to extend a typical YouTube videoplayer[1].

1 Introduction

More and more videos are watched on-demand from a cable-provider or from TV-stations' websites on the Internet. Even though a state-of-the-art cable-box is usually as powerful a computer as a commodity PC, the navigation elements presented to the user when retrieving a video from either of these media are still a legacy from the times of tape-based VCRs. The functionality is limited to play, pause, fast-forward, and fast-backward. More modern video players may also offer directed stream positioning. In other words, even though the computational capabilities are there, they are not used for multimedia content analysis, thus completely ignoring the rich information within the video/audio medium. Even portable devices such as smartphones are sufficiently powerful for more intelligent browsing than play/pause/rewind.

In the following article, we present a system that extends typical navigation features significantly. Based on the idea that TV shows are already conceptually segmented by their producers into narrative themes (such as scenes and dialog segments), we present a system that analyzes these "markers" to present an advanced "narrative-theme" navigation interface. We chose to stick to a particular example presented in the description of the ACM Multimedia Grand Challenge [1] ; namely, the segmentation of Seinfeld episodes [2].

Sitcoms, since their invention as a radio-format in the 1920s, have followed a strict set of rules encoded in the audio and video, e.g. every scene transition is marked by a piece of music and every punchline is labeled by laughter (either from a studio audience or artificially with a laugh track). Our system is able to extract these cues and use the production rules to infer narrative themes relevant to a sitcom: scenes, punch lines, and dialog elements. We use an existing speaker identification system to identify the main actors, as well as the female and male sets of supporting actors. The system is mainly based on existing components that have been evaluated for speed and accuracy in other articles [13, 15, 2]. Even though the system was tested on a conventional PC rather than on a cable box or portable viewer, we have no doubt that it could easily be ported to state-of-the-art set-top boxes or smartphones.

This article describes the motivation and the technical approach. Section 2 presents related work. Section 3 presents the use case that underlies the navigational interface. Section 4 presents the technical approach and under-

[1] A demonstration on the "Soup Nazi" episode of Seinfeld can be found at http://www.icsi.berkeley.edu/~ fractor/seinfeld/.

[2] The system presented here has been selected finalist in the ACM MM Grand Challenge [4] and this article constitutes the first full description of it.

lying methods. Section 5 then presents our evaluation of the system. Section 6 finally concludes and presents future work.

2 Related Work

There is a wealth of related work in multimedia content analysis, especially video analysis. A comprehensive description of the related work would easily exceed the page limits of this article. Therefore, we concentrate on surveying only parts of the most relevant work.

The TRECVid evaluation organized on a year-by-year basis by the US National Institute of Standards and Technologies (NIST) investigates visual event detection on broadcast videos [8]. The task is to detects event like "a person applauding" or "a person riding a bicycle". While many methods developed in the community participating in the evaluation are very interesting for the task presented here, the evaluation does not concentrate on navigation-specific events, but rather on the detection task. It's counterpart, the NIST Rich Transcription (RT) [7] evaluation, focuses on acoustic methods for transcribing multimedia content. The evaluation is currently focusing on meeting data, but previous evaluations included telephone conversations and, more relevantly, broadcast news from radio and television. The acoustic event detection and speaker recognition methods presented in this paper were used in our submission to the NIST RT evaluation 2009.

The Informedia project's [14] basic goal is to "achieve machine understanding of video and film media, including all aspects of search, retrieval, visualization and summarization in both contemporaneous and archival content collections". The main focus is retrieval of videos from a large database though, navigation interfaces are not explored on the level described in this article.

In [9] the authors present a "content-adaptive audio texture based method to segment video into audio scenes". First models are trained for basic audio classes such as speech, music, etc. Then, "semantic audio textures" are defined based on those classes. The idea is to adaptively determine if special transition markers are present in the video. These are then used for audio scene based segmentation. A similar idea using a different approach was proposed in [3], where an unsupervised method for repeated sequence detection in TV broadcast streams was presented. The approaches would be very interesting to try on the data we used. However, the work mainly concentrates on scene transition markers. Actors, for examples, are not distinguished. The approaches are therefore not yet sufficient to create usable navigation interfaces based on the analysis.

The Name-It project [12] aims at recognizing actors from only the visual parts of broadcast shows. The work presented herein depends on the audio parts only. Future work (see Section 6) would certainly include integration of the two methods.

Audio analysis includes both speaker diarization and speaker recognition, both of which have been evaluated in many different domains. The goal of speaker diarization is to segment audio into speaker-homogeneous regions with the ultimate goal of answering the question "who spoke when?" [11]. The task is performed without prior training of specific speaker models. In fact, many systems work completely unsupervised, i.e. they do not require any a-priori knowledge. However, since no prior knowledge is used, speaker diarization is not able to output real names. The goal of speaker recognition, on the other hand, is to detect a person's identity and distinguish it from other speakers. In the classic speaker identification scenario, the test data is generally fairly long (up to three minutes in some cases). Five seconds, an impossibly large latency for a system presented here, is considered a very short utterance. For the work presented here, we use our low-latency, realtime speaker recognition system [5, 13] and train it to identify different actors and acoustic events.

The work presented in this article combines many of the ideas presented in this section to create a usable prototype navigation system for Seinfeld episodes. Since we entirely concentrate on audio analysis, an advantage of our system rarely discussed in related work is that the underlying analysis methods are efficient, fast, robust, and consume little memory and hard disk space, and are therefore easily implementable on a set-top box or portable viewers.

3 Browsing a Sitcom: Use Case

For the video navigation system presented in this article we assume the following use case: The first time a person watches a Seinfeld episode, he or she needs barely any navigation. Unlike other media such as recorded lectures, sitcoms are designed to be entertainment, and as such are intended to hold the attention of the viewer for the entire length of an episode. A play and a pause button should be sufficient. Any non-voluntary interruption of the flow of the episode would probably detract from the experience.

When a sitcom is watched a second or later time, however, a user might want to show a very funny scene to a friend, point out and post the sharpest punchline to his or her facebook network, or even create a home-made YouTube video composed of the most hilarious moments of his or her favorite actor. In order to do this quickly, a navigation interface should support random seek into a video. Although this feature makes it possible to search for a particular moment in the episode, it remains cumbersome, especially because most sitcoms don't follow a single thread of narration. Therefore, we present the user with the basic narrative elements of a sitcom such as the scenes, punchlines, and in-

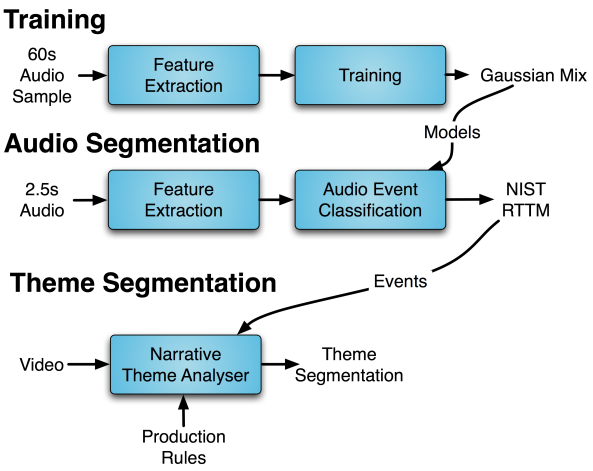

Figure 1. The technical approach for extracting the narrative themes. First, music, laughter, and speaker models for the actors are trained. The models are then used for a low-level acoustic segmentation. The extracted acoustic cues are then used for a high-level narrative theme segmentation.

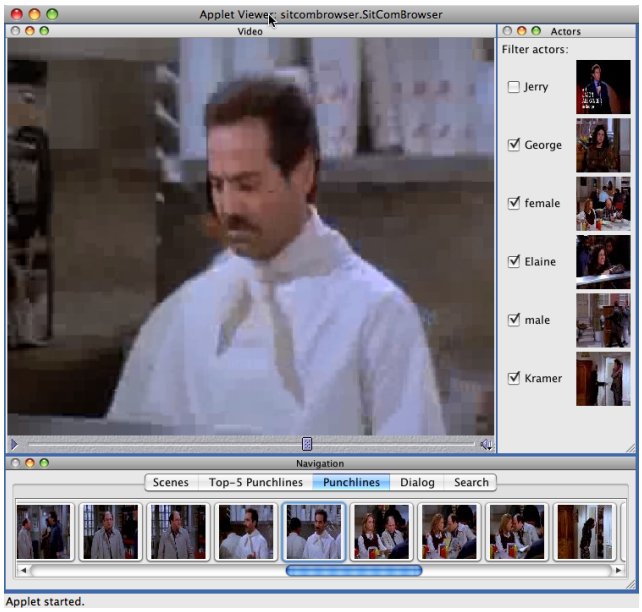

Figure 2. The Seinfeld navigation interface: Users can browse an episode by scene, punchline, and dialog line. The top-5 punchlines are shown in a separate panel. Actors can be selected and de-selected which filters the segments shown in the navigation. A demo is available at http://www.icsi.berkeley.edu/~fractor/seinfeld/.

dividual dialog segments on top of a standard video player interface. A per-actor filter helps to search only for elements that contain a certain protagonist. The user is now able to selectively skip certain parts and to directly navigate into elements he or she remembers from the past.

While keyword search based on speech recognition would certainly improve the search capabilities, people often do not know what they want to search for and would rather *surf* the episode.

4 Technical Approach

The system consists of two main elements: First, a preprocessing and analysis step and second, the online video browser. The preprocessing step consists of an acoustic event detection and speaker identification step, and of a narrative element segmenting step (see Figure 1). The models are trained in a prior training step. The online video browser then uses the output of the narrative element analysis step to present a video navigation interface to the user.

4.1 Acoustic Event and Speaker Identification

The goal of the acoustic event detection component is to segment the audio track into event regions. We distinguish the following types of events: Jerry, Kramer, Elaine, George, male supporting actor, female supporting actor, laughter, music, non-speech (e.g. other noises). The speakers were trained with both pure speech and laughter and music-overlapped speech.

For both acoustic event detection and speaker identification, we use a derivative of the ICSI speaker diarization engine [6] that we presented as a live speaker identification system for meetings in [5] and [13].

For training, we use a one-minute audio sample of the event. From the audio, we compute 19-dimensional MFCCs (a standard feature used in acoustic processing). The features are then used to train a Gaussian Mixture Model (GMM) with 20 Gaussians. The number of Gaussians and iterations has been determined empirically, as described in [13].

In the actual recognition mode, the system records and processes chunks of audio as follows. In a first step of feature extraction, the sampled audio data is converted into 19th-order MFCCs. Cepstral Mean Subtraction (CMS) is implemented to address stationary channel effects. Although in subtracting the mean, some speaker-dependent information is lost [10], according to the experiments performed, the major part of the discriminant information remains in the temporal varying signal.

In the classification step, the likelihood for each set of features is computed against each set of Gaussian Mixtures obtained in the training step. As determined by the experiments on large meeting corpora (see [13]), we use 2.5-second chunks of audio and a frame-length of 10 ms. This means that a total of 250 frames are examined to determine if an audio segment belongs to a certain speaker, to laughter, to music, or to the non-speech model. The decision is reached using majority vote on the likelihoods. The result is saved in NIST-standardized RTTM-format (see [6]).

Similar to our experiments with meetings, the speech in sitcoms is typically single-channel. However, the speech data contained in sitcoms has slightly different properties compared to speech recorded in meetings. TV shows are usually recorded using a boom microphone; therefore, the signal to noise ratio is much better than for speech recordings using a single desktop microphone. More importantly, unlike real-world meetings, there is very little overlapped speech in sitcoms. So in general, detecting acoustic events and speakers in a TV show should be easier than in a real-world meeting recording.

One disadvantage over meeting recordings is that speech is very often overlapped with either music or laughter. While it is often possible to separate the music/laughter track by implementing a 4-on-2 surround decoder (because laughter/music and speech are usually on a different channels), we found that there is little benefit in doing so as the methods presented in [13] seem to be quite robust against overlapping laughter and music.

4.2 Narrative Theme Analysis

The narrative theme analyzer is a rule-based system that transforms the segmentation generated by acoustic event and speaker detection into segments that reflect narrative themes. We expect having to adjust these rules for different TV series, e.g. "I Love Lucy" might have similar but not identical rules. The rules for the Seinfeld themes are as follows:

- A dialog element is a single contiguous speech segment by one speaker.
- A punchline is a dialog element that is followed by pure laughter. Punchlines are prioritized by the length of this laughter segment. The longer the laughter, the more important is the punchline.
- The top-5 punchlines are the 5 punchlines followed by longest laughter.
- A scene is a segment of at least 10 seconds between two music events or a music event bracketing the beginning or end an episode.

The narrative theme analyzer also creates icons for use in the graphical interface. A representative image is needed for each actor as well as for each narrative theme (scene, punchline, dialog element). Again, the approach follows the artistic production rules for sitcoms. TV production rules prescribe that, other than for artistic exceptions, the actor has to be shown once a certain speaking time is exceeded. So for the actor image, we take the median frame of the longest speech segment of the particular actor. Of course, this does not preclude multiple actors appearing within this frame. Thus, a visual recognition approach could be used to crop the picture and only generate an icon from the face of the actor. For scenes, punchlines, and dialog elements, we take the median frame of the corresponding segment.

4.3 Video Browser

We envision the thematic browser (see Figure 2) replacing a typical YouTube video player. The browser shows the video and allows play and pause, as well as seeking to random positions. The navigation panel on the bottom shows iconized frames of the video, as described previously. The navigation panel allows the user to directly jump to the beginning time of either the scene, punchline, top-5 punchline, or dialog element. Also, the current narrative element is highlighted while the show is playing. In order to make navigation more selective, the user can deselect one of the main actors or the male/female supporting actors. In this case, scenes, punchlines, or dialogs that only contain deselected actors are no longer visible in the navigation bar. The actor icons are grabbed from the center of their longest dialog segment — we imagine using the localization approach presented in [6] to obtain better results in the future.

5 Evaluation

In order to evaluate the usefulness of our approach, we performed several experiments. First we measured the performance of the presented algorithm on a current Mac Book Pro laptop (2.8 GHz dual core, 64 bit, 4 GB RAM). The training of the models for the acoustic event detection and speaker identification takes about $0.3\times$ realtime. Since we only need to train 9 classes of 60-second duration, the total training time is roughly 3 minutes. The actual classification step then takes about 10% realtime. The narrative theme analysis including icon extraction takes another $10\,\%$ realtime. Therefore, given the models, a new 25-minute episode of Seinfeld can be analyzed in about 6–7 minutes. Note that this needs only be performed at most once per episode. The results are stored in a small text file, and can be loaded and browsed exceedingly quickly.

In order to test the accuracy of the acoustic event detection, we hand-annotated the Seinfeld episode "The Soup

NAZI" completely. We did not use forced alignment and we only used one annotator. Therefore, direct performance comparison to e.g. [5] or [6] is difficult. For measuring the accuracy, we used NIST's "md-eval" tool which uses a dynamic programming procedure to find the optimal one-to-one mapping between the hypothesis and the ground truth segments such that the total overlap between the reference event and the corresponding mapped hypothesized event cluster is maximized. The difference is expressed as Diarization Error Rate (DER) which is defined by NIST[3]. We obtained a Diarization Error Rate of 46 %, which is an error of about 5 % per class and therefore roughly consistent with [13]. We expect the system to perform much better with more training data and the exact annotation of all supporting actors might improve the system. When analyzing different Seinfeld episodes across different seasons, we found that the music changes slightly, and therefore music models must be retrained.

DER does not measure the subjective performance of the system when tested by the user. Therefore, we also performed a comparative anecdotal user study with 10 colleagues. The colleagues were either students or professionals working in machine learning, both in academia and industry. We created two versions of the browser available for viewing. One used the manually annotated acoustic segments as a basis for the narrative theme analysis, and the other used the automatic processing chain described above. Two statements seemed to hold — the narrative theme analysis on the ground truth data was de-facto perfect and, even though the generated icons for ground-truth vs. automatic segmentation were different, there was no perceived usability difference between the two. Some subjects even claimed the automatic version was "better". We interpret this as a good sign that the automatic method is basically indistinguishable from the hand-annotated version.

Finally, the demo of the system as presented here was submitted to the ACM Multimedia Grand Challenge, where it was reviewed by five jury members and invited to the finals. At the time of writing of this article, the winner has not been selected yet.

6 Conclusion and Future Work

This article presented a system that enables enhanced navigation in Seinfeld episodes. Users can navigate directly to a punchline, a top-5 punchline, a scene, or a dialog element, and can explicitly include or exclude actors in the navigation. The method for producing the segmentation leverages the artistic production rules of the genre, which specify how narrative themes should be presented to the audience. The article describes the use of our low-latency speaker identification and acoustic event detection system in combination with a rule-based narrative theme analyzer. An evaluation of the approach shows the system to be very efficient and decently robust, since most of the techniques are derived from existing and thoroughly evaluated machine learning components. We believe the automatic segmentation could still be improved by training models more accurately and with more data.

Since sitcoms were invented as a radio format, narrative theme markers are mainly observed in the audio track. It seems that even very modern sitcoms still obey the production rules invented in the 1920s. However, segmenting with only audio cues may be more difficult for some modern sitcoms such as "Hannah Montana", as they contain a large amount of music that does not indicate scene transitions. Also, punchlines might be purely visual and thus contain laughter but no preceding speech segment. For this and other reasons, combining the methods presented in this article with visual analysis will undoubtedly improve the user experience. Therefore a multimodal approach should be the next step.

Future work will include sitcoms other than Seinfeld. Other genres could also be explored, since many follow similarly strict patterns of narrative themes (e.g. medical dramas, soap operas, game shows). We would expect to have to edit the rules of the narrative theme analyzer for other sitcoms and other genres. Additional user studies and comparison with human annotation would also provide further insight into the usability and generalizability of the methods presented.

[3] http://nist.gov/speech/tests/rt/rt2004/fall

References

[1] ACM Multimedia Grand Challenge. `http://www.acmmm09.org/MMGC.aspx`.

[2] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo. Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system. In *Proceeding of the NIST MLMI Meeting Recognition Workshop, Edinburgh*, 2005.

[3] S. Berrani, G. Manson, and P. Lechat. A non-supervised approach for repeated sequence detection in TV broadcast streams. *Signal Processing: Image Communication*, 23(7):525–537, 2008.

[4] G. Friedland, L. Gottlieb, and A. Janin. Joke-o-mat: Browsing sticoms punchline-by-punchline. In *Proceedings of ACM Multimedia*, page to appear. ACM, October 2009.

[5] G. Friedland and O. Vinyals. Live speaker identification in conversations. In *Proceedings of ACM Multimedia*, pages 1017–1018. ACM, October 2008.

[6] G. Friedland, C. Yeo, and H. Hung. Visual speaker localization aided by acoustic models. In *Proceedings of ACM Multimedia*, page to appear. ACM, October 2009.

[7] NIST Rich Transcription evaluation. `http://www.itl.nist.gov/iad/mig//tests/rt`.

[8] NIST TRECVid evaluation. http://www-nlpir.nist.gov/projects/trecvid/.

[9] F. Niu, N. Goela, A. Divakaran, and M. Abdel-Mottaleb. Audio scene segmentation for video with generic content. In *Proceedings of SPIE*, volume 6820, page 68200S, 2008.

[10] D. A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 1-2:91–108, 1995.

[11] D. A. Reynolds and P. Torres-Carrasquillo. Approaches and applications of audio diarization. In *Proceedings of the IEEE ICASSP*, 2005.

[12] S. Satoh, Y. Nakamura, and T. Kanade. Name-it: Naming and detecting faces in news videos. *IEEE Multimedia*, 6(1):22–35, 1999.

[13] O. Vinyals and G. Friedland. Towards semantic analysis of conversations: A system for the live identification of speakers in meetings. In *Proceedings of IEEE International Conference on Semantic Computing*, pages 456–459, August 2008.

[14] H. Wactlar, T. Kanade, M. Smith, and S. Stevens. Intelligent access to digital video: Informedia project. *Computer*, 29(5):46–52, 1996.

[15] C. Wooters and M. Huijbregts. The ICSI RT07s speaker diarization system. In *Proceedings of the RT07 Meeting Recognition Evaluation Workshop*, 2007.