Screenplay alignment for closed-system speaker identification and analysis of feature films

2 authors:

Robert Turetsky
Columbia University
5 PUBLICATIONS   **151** CITATIONS

SEE PROFILE

Nevenka Dimitrova
Philips Cambridge
252 PUBLICATIONS   **3,221** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project   Combinatorial Approach for Automated Platform Diversity Testing View project

Screenplay Alignment for Closed-System Speaker Identification and Analysis of Feature Films

Robert Turetsky[1,2], Nevenka Dimitrova[1]
rob@ee.columbia.edu, nevenka.dimitrova@philips.com
(1) Philips Research, 345 Scarborough Rd., Briarcliff Manor, NY
(2) Columbia University, 500 W. 120th St. Room 1312, New York, NY

Abstract

Existing methods for audiovisual and text analysis of videos perform "blind" recovery of metadata from the audiovisual signal. The film production process however is based on the original screenplay and its versions. Using this information is like using the recipe book for the movie. High-level semantic information that is otherwise very difficult to derive from the audiovisual content can be extracted automatically using both audiovisual signal processing as well as screenplay processing and analysis.

As a testbed of our approach we investigated the use of screenplay as a source of information for speaker/character identification. Our speaker identification method consists of screenplay parsing, extraction of time-stamped transcript, alignment of the screenplay with the time-stamped transcript, audio segmentation and audio speaker identification. As the screenplay alignment will not be able to identify all dialogue sections within any film, we use the segments found by alignment as labels to train a statistical model in order to identify unaligned pieces of dialogue. We find that the screenplay alignment was able to properly identify the speaker in 30% of lines of dialogue on average. However, with additional automatic statistical labeling for audio speaker ID on the soundtrack our recognition rate improves significantly.

1. Introduction

Almost all feature-length films are produced with the aid of a screenplay. The screenplay provides a unified vision of the story, setting, dialogue and action of a film – and gives the filmmakers, actors and crew a starting point for bringing their creative vision to life. For those involved in content-based analysis of movies, the screenplay is a currently untapped resource for obtaining a textual description of important semantic objects within a film coming straight from the filmmakers. Hundreds of copies of a screenplay are produced for any film production of scale. The screenplay can be reproduced for hobbyist or academic use, and thousands of screenplays are available online.

The difficulty in using the screenplay as a shortcut to content-based analysis is threefold: 1) The screenplay follows only a semi-regular formatting standard, and thus needs robust parsing to be a reliable source of data. 2) There is no inherent correlation between text in the screenplay and a time period in the film, and 3) Lines of dialogue or entire scenes in the movie can be added, deleted, modified or shuffled. We address these difficulties by parsing the screenplay and then aligning it with the time-stamped subtitles of the film. Statistical models can then be generated based on properly aligned segments in order to estimate segments that could not be aligned.

Our test-bed of this framework is in character/speaker identification. Unsupervised (audio) speaker identification on movie dialogue is a difficult problem, as speech characteristics are affected by changes in emotion of the speaker, different acoustic conditions, ambient noise and heavy activity in the background. Patel and Sethi [5] have experimented with speaker identification on film data for use in video indexing/classification, but require that training data be hand-labeled and that all dialogues be hand-segmented. Salway et al. describe the association of temporal information in a movie to collateral texts (audio scripts) [1]. Wachman et al have used script information from situation comedy for labeling and learning by example in interactive sessions [3].

The remainder of this paper will proceed as follows. First, we introduce the content and structure of the screenplay in Section 2. Subsequently, we detail extracting information from the screenplay and the alignment process in Section 3. We present a quantitative analysis of alignments in section 4, and preliminary results of audio speaker ID in section 5. We present concluding remarks in section 6.

2. Concerning screenplays

As mentioned above, the screenplay describes the story, characters, action, setting and dialogue of a film. Additionally, some camera directions and shot boundaries may be included but are generally ignored. The screenplay generally undergoes a number of revisions, with each rewrite looking potentially uncorrelated to prior drafts (see *Minority Report* [6]). After the principal shooting of a film is complete, the editors assemble the different shots together in a way that may or may not respect the screenplay.

The actual content of the screenplay generally follows a (semi) regular format. Figure 1 shows a snippet of a

screenplay from the film *Contact*. The first line of any scene or shooting location is called a *slug line*. The slug line indicates whether a scene is to take place inside or outside (`INT` or `EXT`), the name of the location (e.g. 'TRANSPORT PLANE'), and can potentially specify the time of day (e.g. `DAY` or `NIGHT`). Following the slug line is a description of the location. Additionally, the description will introduce any new characters that appear and any action that takes place without dialogue. Important people or objects are made easier to spot within a page by capitalizing their names.

The bulk of the screenplay is the dialogue description. Dialogue is indented in the page for ease of reading and to give actors and filmmakers a place for notes. Dialogues begin with a capitalized character name and optionally a (V.O.) or (O.S.) following the name to indicate that the speaker should be off-screen (V.O. stands for "Voice-over"). Finally, the actual text of the dialogue is full-justified to a narrow band in the center of the page.

The continuity script, a shot-by-shot breakdown of a film, is sometimes written after all work on a film is completed. A method for alignment of the continuity script with closed captions was introduced by Ronfard and Thuong [1]. Although continuity scripts from certain films are published and sold, they are generally not available to the public online. This motivates analysis on the screenplay, despite its imperfections.

```
INT.  TRANSPORT PLANE

The Major Domo leads Ellie down a corridor of the plane's
custom interior.  Through one door we can see into a room
where several very beautiful, very young women sit
watching TV with vacant eyes.  Ellie only catches a
glimpse before the Major Domo nods for her to enter the
main room and takes his post outside.

Ellie cautiously enters the interior of what appears to be
a flying Dascha; dark, heavy on the chintz.  Bookshelves,
an exercise machine... and a wall of monitors, filled top
to bottom with scrolling hieroglyphics.  Ellie reaches out
to them --

                    HADDEN (O.S.)
              Dr. Arroway, I presume.

Ellie turns to see S.R. HADDEN.  Thick glasses, wearing a
cardigan, he stands by a silver samovar.

                    ELLIE
              S.R. Hadden...
                    (beat)
              You compromised our security codes.

                    HADDEN
              Once upon a time I was a hell of an
              engineer.  Please, sit, Doctor.  I
              have guests so rarely, it's
              important to me they feel welcome in
              my home.
                    (turns to the
```

Figure 1: Example segment from a screenplay

3. System overview

One reason why the screenplay has not been used more extensively in content-based analysis is because there is no explicit connection between dialogues, actions and scene descriptions present in a screenplay, and the time in the video signal. This hampers our effectiveness in assigning a particular segment of the film to a piece of text. Another source of film transcription, the closed captions, has the text of the dialogue spoken in the film, but it does not contain the identity of characters speaking each line, nor do closed captions possess the scene descriptions which are so difficult to extract from a video signal. We get the best of both worlds by aligning the dialogues of screenplay with the text of the film's time stamped closed captions

Second, lines and scenes are often incomplete, cut or shuffled. In order to be robust in the face of scene re-ordering we align the screenplay to the closed captions one scene at a time and filter out potential false positives through median filtering.

Finally, we are unable to find correlates in the screenplay for every piece of dialogue. Thus, it becomes imperative to take information extracted from the time-stamped screenplay, combined with multimodal segments of the film (audio/video stream, closed captions, information from external websites such as imdb.com, other films), to create statistical models of events not captured by the screenplay alignment.

A system overview of our test-bench application, which includes pre-processing, alignment and speaker identification throughout a single film, is shown in Figure 2. First we parse the text of a film's screenplay, so that scene and dialogue boundaries and metadata are entered into a uniform data structure. Next, the closed caption and audio streams are extracted from the film's DVD. In the crucial stage, the screenplay and closed caption texts are aligned. The aligned dialogues are now time-stamped and associated with a particular character. These dialogues are used as labeled training examples for generic machine learning methods (in our case we've tested neural networks and GMM's) which can identify the speaker of dialogues which were not labeled by the alignment process.
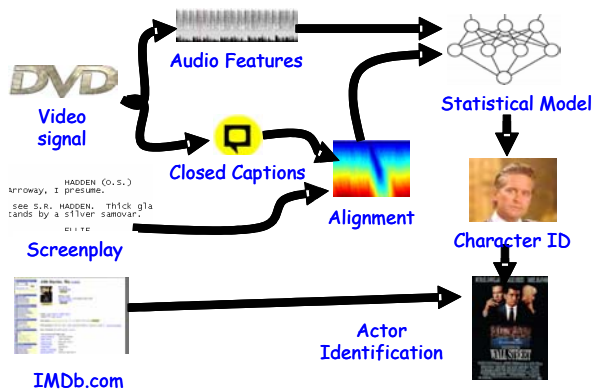
Figure 2. System overview

In our experiments, we were working towards very high speaker identification accuracy, despite the difficult noise conditions. It is important to note that we were able to perform this identification using supervised learning methods, but the ground truth was generated automatically so there is no need for human intervention in the classification process.

3.1 Screenplay parsing

From the screenplay, we extract the location and time and description of a scene, the individual lines dialogue and their speaker, and the parenthetical and action direction for

the actors, and any transition suggestion (cut, fade, wipe, dissolve, etc) between scenes. We used the following grammar for parsing most screenplays:

SCENE_START:	.* | SCENE_START | DIAL_START | SLUG | TRANSITION
DIAL_START:	\t+ <CHAR NAME> (V.O.|O.S.)? \n \t+ DIALOGUE | PAREN
DIALOGUE:	\t+ .*? \n\n
PAREN:	\t+ (.*?)
TRANSITION:	\t+ <TRANS NAME> :
SLUG:	<SCENE #>?<INT/EXT><ERNAL|.> ? - <LOC> <- TIME>

A similar grammar was generated for two other screenplay formats which are popular on-line.

3.2 Time-stamped subtitles and metadata

For the alignment and speaker identification tasks, we require the audio and time stamped closed caption stream from the DVD of a film. In our case, four films, *Being John Malkovich, Magnolia, L.A. Confidential* and *Wall Street*, were chosen from a corpus of DVDs. When available, subtitles were extracted from the User Data Field of the DVD. Otherwise, OCR (Optical Character Recognition) was performed on the subtitle stream of the disc.

Finally, character names from the screenplay are converted to actor names based on fields extracted from imdb.com. If no match from a charcater's name can be found in IMDB's reporting of the film's credit sequence (e.g. 'Hank' vs. 'Henry'), we match the screenplay name with the credit sequence name by matching quotes from the "memorable quotes" section with dialogues from the screenplay.

3.3 Screenplay Alignment

The screenplay dialogues and closed caption text are aligned by using dynamic programming to find the "best path" across a similarity matrix. Alignments that properly correspond to scenes are extracted by applying a median filter across the best path. Dialogue segments of reasonable accuracy are broken down into closed caption line sized chunks, which means that we can directly translate dialogue chunks into time-stamped segments. Below, each component is discussed.

The similarity matrix is a way of comparing two different versions of similar media [4], [9]. In our similarity matrix, every word i of a scene in the screenplay is compared to every word j in the closed captions of the entire movie. In other words, we populate a matrix:

$$SM(i, j) \leftarrow screenplay(scene\_num, i) == subtitle(j)$$

$SM(i,j)=1$ if word i of the scene is the same as word j of the closed captions, and $SM(i,j)=0$ if they are different. Screen time progresses linearly along the diagonal i=j, so when lines of dialogue from the screenplay line up with lines of text from the closed captions, we expect to see a solid

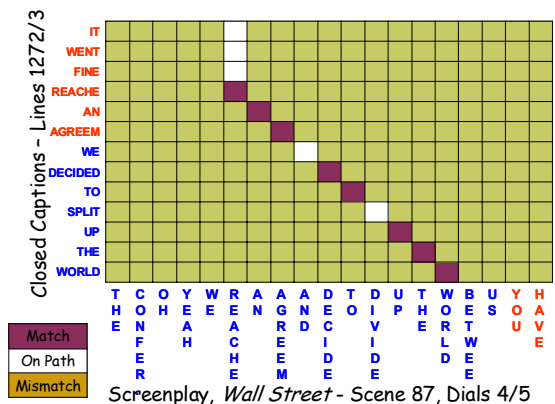diagonal line of 1's. Figure 33 shows an example similarity matrix segment.

Figure 3. Screenplay vs. closed captions similarity

In order to find the diagonal line which captures the most likely alignment between the closed captions and the screenplay scene, we perform dynamic programming. Figure 44 visualizes the successful alignment of scene 30 of *Magnolia*. The three diagonal regions indicate that much of this scene is present in the closed captions, but there are a few dialogues which written into the screenplay but were not present in the finished film and vice versa.

We are able to remove these missing dialogues by searching for diagonal lines across DP's optimal path, as in [2]. We use an m-point median filter on the slope of the optimal path, so that in addition to lining up with the rest of the scene, the dialogue must consistently match at least $(m+1)/2$ (e.g. 3) out of m (e.g. 5) words to be considered part of a scene. In Figure 44, proper dialogues are demonstrated in the shaded regions. The bounds of when the alignment is found to be proper are used to segment aligned scenes from omitted ones.

We can then warp the words of the screenplay to match the timing of the closed captions. An example warping, as taken from the film *Wall Street*, is shown in **Figure 55**.
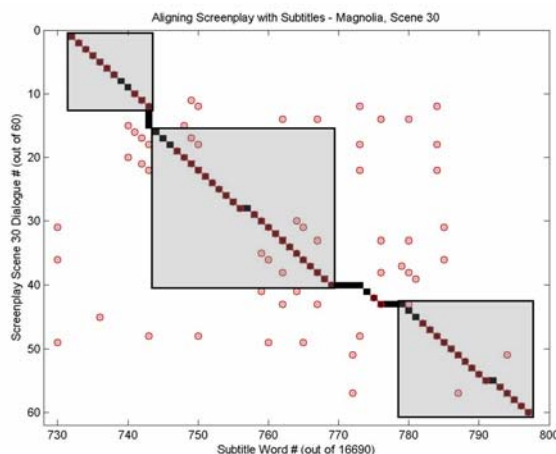

Figure 4. Alignment visualization

```
1:3:50.233  1:3:54.829  DARIEN  IF I COUID HAVE ANYTHING THIS WOUID ALMOST DO
1:3:56.239  1:3:58.707  BUD     ALMOST
1:3:58.775  1:4:1.676   DARIEN  SO HOWD YOUR CONFERENCE GO WITH GORDON
1:4:1.744   1:4:4.178   BUD     IT WENT FINE REACHED AN AGREEMENT
1:4:4.247   1:4:7.705   BUD     WE DECIDED TO SPLIT UP THE WORLD BETWEEN US
1:4:7.784   1:4:9.684   DARIEN  YOU HAVE MODEST WANTS
1:4:9.752   1:4:12.186  DARIEN  I LIKE THAT IN A MAN
1:4:12.255  1:4:14.189  BUD     WHAT DO YOU WANT
1:4:14.257  1:4:20.218  DARIEN  IETS SEE A TURNER A PERFECT CANARY diamond
1:4:20.296  1:4:23.697  DARIEN  WORID PEACE THE BEST OF EVERYTHING
1:4:23.766  1:4:25.734  BUD     OH WHY STOP AT THAT
1:4:25.802  1:4:27.360  DARIEN  I DONT
1:4:46.289  1:4:48.655  BUD     ILL PARK MONEY IN YOUR ACCOUNT
```

Figure 5. Timestamped dialogues as generated by the alignment process

4. Quantitative analysis of alignments

Below we evaluate the performance of the alignment process. Here we define the *coverage* of the alignment to be the percentage of lines of dialogue in the film in which the alignment was able to identify the speaker. The *accuracy* of the alignment is the percentage of speaker IDs generated by the alignment which actually correspond to dialogue spoken by the tagged speaker. Accuracy is a measure of the purity of the training data and coverage is a measure of how much data will need to be generated by classification.

	CRA	LES	LOT	MAL	MAX	?
CRA	82	0	1	1	0	11
LES	0	41	0	0	0	0
LOT	0	0	40	0	0	2
MAL	0	0	0	25	0	2
MAX	0	0	1	0	71	4

Table 1: Confusion Matrix for segment speaker label accuracy, *Being John Malkovich*.

Table 1 presents lines of dialogue that were identified as belonging to a main character in *Being John Malkovich*. We were able to achieve a high degree of accuracy in labeling the speaker for each segment.

While the alignment process affords a high level of confidence in terms of the accuracy (approximately 90%) of segment speaker label generation, the liquid nature of the screenplay means we are unable to label most of the film. Table 2 presents a measure of how much dialogue the alignment is able to cover in each of our four films. This motivates creating a speaker-identification system based on statistical models generated from the segment labeling as found by alignment.

Movie	#CC's	Coverage	Accuracy
Malkovich	1436	334 (23%)	311 (93%)
LA Conf	1666	548 (33%)	522 (95%)
Wall St.	2342	954 (41%)	850 (89%)
Magnolia	2672	843(32%)	747 (89%)

Table 2: Accuracy and coverage of alignments

5. Statistical models for speaker ID

Our speaker identification system examines the behavior of audio features over time. We performed extensive testing with various combinations of audio features reported to have high discriminability (MFCC, LSP, RASTA-PLP), incorporating mean subtraction and deltas. In our case we have a good deal of training data so we can use simple classifiers. The goal of our classifier is to allow for different clusters in feature-space that correspond to the voice characteristics of an actor under different emotional and acoustic conditions.

While our method is under large scale benchmarking, initial results are promising. Table 3 presents frame accuracy on unlabeled data. Here we are using the first 13 MFCC components at 12.5msec intervals stacked across a .5sec time window. Principal Component Analysis [8] was used to reduce the dimensionality of feature-space, and classification was performed using an 8-component Gaussian Mixture Model. Note that the table demonstrates that identification accuracy is highly speaker dependant.

	CRA	LES	LOT	MAL	MAX
CRA	57 %	17 %	21 %	27 %	28 %
LES	5 %	77 %	4 %	15 %	1 %
LOT	8 %	0 %	49 %	10 %	13 %
MAL	7 %	2 %	8 %	31 %	4 %
MAX	23 %	4 %	19 %	17 %	55 %

Table 3: Confusion Matrix for percentage of frames labeled by automatic speaker identification

6. Conclusions and Future Work

We have described a method for using screenplay information to extract high level semantic information about a film, and to create models to describe segments where other description is not available or unreliable. This is advantageous because the screenplay contains data about the film, that is not extractable at all by audiovisual analysis or if it can be extracted then the reliability is very low. These high level concepts are closer to the human understanding of the film and to the potential methods of searching of audiovisual content. We used screenplay information for speaker ID, which has limited coverage of about 30%. Then we used the same framework for generating labels for a statistical approach to audio speaker ID.

There are a limitless number of potential applications for this alignment. Aside from speaker ID, it is possible to get a textual description of scenes. Statistical analysis of the texts of thousands of screenplays can be mined to associate emotions with words and words with timestamps. It will then be possible to derive statistical models of different affects as present across many films.

7. References

[1] Salway, A., Tomadaki, E.,"Temporal information in collateral texts for indexing moving images," in *Proc. of LREC 2002 Workshop on Annotation Standards for Temporal Information in Natural Language*, 2n002.

[2] Ronfard R., Thuong, T. T., "A Framework for Aligning and Indexing Movies with their Script," Proc. of ICME 2003, Baltimore, MD, July 6-9, 2003.

[3] Wachman J., and Picard, P.W., "Tools for browsing a TV situation comedy based on content specific attributes," *Multimedia Tools and Applications*, vol. 13, no. 3, pp. 255–284, 2001.

[4] R. Turetsky and D. P. W. Ellis. *Ground-Truth Transcriptions of Real Music from Force-Aligned MIDI Syntheses*. ISMIR 2003.

[5] N. Patel and I. Sethi. *Video Classification Using Speaker Identification*. IS&E SPIE Proc.: Storage and Retrieval for Image and Video Databases V, Jan 1997, San Jose, California

[6] S. Frank, *Minority Report*. Early and revised Drafts, available from Drew's Script-o-rama, http://www.script-o-rama.com

[7] Continuity script for closed captions: http://www.cfv.org/caai/nadh33.pdf

[8] Jolliffe, I. T., *Principal Component Analysis*, SpringerVerlag, New York, 1986.

[9] J. Foote. *Methods for the Automatic Analysis of Music and Audio*, Technical Report FXPAL-TR-99-038. 1999