Feature Selection for Trainable Multilingual Broadcast News Segmentation

David D. Palmer, Marc Reichman, Elyes Yaich
Virage Advanced Technology Group
300 Unicorn Park
Woburn, MA 01801
{dpalmer,mreichman,eyaich}@virage.com

Abstract

Indexing and retrieving broadcast news stories within a large collection requires automatic detection of story boundaries. This video news story segmentation can use a wide range of audio, language, video, and image features. In this paper, we investigate the correlation between automatically-derived multimodal features and story boundaries in seven different broadcast news sources in three languages. We identify several features that are important for all seven sources analyzed, and we discuss the contributions of other features that are important for a subset of the seven sources.

1 Introduction

Indexing and retrieving stories within a large collection of video requires automatic detection of story boundaries, and video story segmentation is an essential step toward providing the means for finding, linking, summarizing, and visualizing related parts of multimedia collections. In many cases, previous story segmentation research has focused on single stream analysis techniques, utilizing only one of the information sources present in news broadcasts: natural language, audio, image, and video (see, for example, (Furht *et al.*, 1995),(Fiscus and Doddington, 2002),(Greiff *et al.*, 2001), (O'Connor *et al.*, 2001)). Some segmentation research has included multimodal approaches that were capable of combining features from multiple information sources (Boykin and Merlino, 1999),(Hauptmann and Witbrock, 1998) . While this work was a significant improvement over single-stream approaches, they were rarely applied to non-English sources without closed captioning.

Previous work on story segmentation has identified many features useful for finding story boundaries, but feature selection is often model-dependent and does not account for the differences between broadcast sources. Specific features useful for video story segmentation vary widely from one source to the next, and the degree to which each feature is useful also varies across sources and even from one broadcast to the next within a single source. This variety suggests the need for trainable techniques in which salient source-specific features can be automatically learned from a set of training data. This data-driven approach is especially important in multilingual video processing, where native speakers may not be available to develop segmentation models for every language.

The goal of this paper is to provide a model-independent investigation of the correlation between a wide range of multimedia features and news story boundaries, in order to aid the development of improved segmentation algorithms. Our work seeks to complement recent work in model-dependent feature selection, such as (Hsu and Chang, 2003), without making assumptions about the dependencies between features.

The feature analysis we describe in this paper consisted of several steps. First, we created a data set for our experiments by capturing and digitally encoding a set of news broadcasts from seven video news sources in three languages. A native speaker manually labelled the story and commercial boundaries in each broadcast; we describe the data in Section 2. We ran several state-of-the-art audio and video analysis software packages on each recorded broadcasts to extract time-stamped multimedia metadata, and we defined a set of possible segmentation features based on the metadata values produced by the analysis software; we describe the software analysis and feature extraction in Section 3. Finally, we analyzed the patterns of occurrence of the features with respect to story and commercial boundaries in all the news broadcasts; the results of our analysis are described in Section 4.

2 Data

The data for our segmentation research consists of a set of news broadcasts recorded directly from a satellite dish between September 2002 and February 2003. The data set contains roughly equal amounts (8-12 hours) of news broadcasts from seven sources in three languages: Aljazeera (Arabic), BBC America (UK English), China Central TV (Mandarin Chinese), CNN Headline News (US English), CNN International (US/UK English), Fox News (US English), and Newsworld International (US/UK English).

Each broadcast was manually segmented with the labels "story" and "commercial" by one annotator and verified by a second, at least one of whom was a native speaker of the broadcast language. We found that a very good segmentation is possible by a non-native speaker based solely on video and acoustic cues, but a native speaker is required to verify story boundaries that require language knowledge, such as a single-shot video sequence of several stories read by a news anchor without pausing. The definition of "story" in our experiments corresponds with the Topic Detection and Tracking definition: a segment of a news broadcast with a coherent news focus, containing at least two independent, declarative clauses (LDC, 1999). The segments within broadcasts briefly summarizing several stories were not assigned a "story" label, nor were anchor introductions, signoffs, banter, and teasers for upcoming stories. Each individual story within blocks of contiguous stories was labeled "story." A sequence of contiguous commercials was annotated with a single "commercial" label with a single pair of boundaries for the entire block.

Table 1 shows the details of our experimental data set. The first two columns show the broadcast source and the language. The next two columns show the total number of hours and the number of hours labeled "story" for each source. It is interesting to note that the percentage of broadcast time devoted to news stories varies widely by source, from 62% for CNN Headline News to 90% for CNN International. Similarly, the average story length varies widely, as shown in the final column of Table 1, from 52 seconds per story for CNN Headline News to 171 seconds per story for Fox News. These large differences are extremely important when modeling the distributions of stories (and commercials) within news broadcasts from various sources.

Source	Lang.	Total Hours	Story Hours	Story Count	Ave. Length
ALJ	Ara	10:37	6:56	279	89 s
BBC	Eng	8:09	6:09	215	103 s
CCTV	Chi	9:05	7:14	235	111 s
CNNH	Eng	11:42	7:18	505	52 s
CNNI	Eng	10:14	9:13	299	111 s
Fox	Eng	13:13	9:14	194	171 s
NWI	Eng	8:33	6:12	198	113 s

Table 1: *Data sources (Broadcast source, language, total hours, hours of stories, number of stories, average story length).*

3 Feature extraction

In order to analyze audio and video events that are relevant to story segmentation, we encoded the news broadcasts described in Section 2 as MPEG files, then automatically processed the files using a range of media analysis software components. The software components represented state-of-the-art technology for a range of audio, language, image, and video processing applications.

The audio and video analysis produced time-stamped metadata such as "face Chuck Roberts detected at time=2:38" and "speaker Bill Clinton identified between start=12:56 and end=16:28." From the raw metadata we created a set of features that have previously been used in story segmentation work, as well as some novel features that have not been used in previous published work. The software components and resulting features are described in the following sections.

3.1 Audio and language processing

A great deal of the information in a news broadcast is contained in the raw acoustic portion of the news signal. Much of the information is contained in the spoken audio, both in the characteristics of the human speech signal and in the sequence of words spoken. This information can also take the form of non-spoken audio events, such as music, background noise, or even periods of silence. We ran the following audio and language processing components on each of the data sources described in Section 2.

Audio type classification segments and labels the audio signal based on a set of acoustic models: speech, music, breath, lip smack, and silence. **Speaker identification** models the speech-specific acoustic characteristics of the audio and seeks to identify speakers from a library of known speakers. **Automatic speech recognition** (ASR) provides an automatic transcript of the spoken words. **Topic classification** labels segments of the ASR output according to predefined categories. The audio processing software components listed above are described in detail in (Makhoul *et al.*, 2000). **Closed captioning** is a human-generated transcript of the spoken words that is often embedded in a broadcast video signal.

Story segmentation features automatically extracted from audio and language processing components were: speech segment, music segment, breath, lip smack, si-

lence segment, topic classification segment, closed captioning segment, speaker ID segment, and speaker ID change. In addition we analyzed the ASR word sequences in all broadcasts to automatically derive a set of source-dependent **cue phrase n-gram features**. To determine cue n-grams, we extracted all relatively frequent unigrams, bigrams, and trigrams from the training data and compared the likelihood of observing each n-gram near a story boundary vs. elsewhere in the data. Cue n-gram phrases were deemed to be those that were significantly more likely near the start of a story.

3.2 Video and image processing

The majority of the bandwidth in a video broadcast signal is devoted to video content, and this content is a rich source of information about news stories. The composition of individual frames of the video can be analyzed to determine whether specific persons or items are shown, and the sequence of video frames can be analyzed to determine a pattern of image movement. We ran the following image and video processing components on each of the data sources described in Section 2.

Face identification detects human faces in the image and compares the face to a library of known faces. **Color screen detection** analyzes the frame to determine if it is likely to be primarily a single shade, like black or blue. **Logo detection** searches the video frame for logos in a library of known logos. **Shot change classification** detects several categories of shot changes within a sequence of video frames.

Story segmentation features automatically extracted from image and video processing components were: anchor face ID, blue screen detection, black screen detection, logo detection, fast scene cut detection, slow scene transition detection, gradual scene transient detection, and scene fade-to-black detection.

3.3 Feature analysis methodology

Each feature in our experiments took the form of a binary response to a question that related the presence of raw time-stamped metadata within a window of time around each story and commercial boundary, e.g., "Did an anchor face detection occur within 5 seconds of a story boundary?" For processing components that produce metadata with an explicit duration (such as a speaker ID segment), we defined separate features for the start and end of the segment plus a feature for whether the metadata segment "persisted" throughout the time window around the boundary. For example, a speaker ID segment that begins at t=12 and ends at t=35 would result in a true value for the feature "Speaker ID segment persists," for a time window of 5 seconds around a story boundary at t=20.

For each binary feature, we calculated the maximum likelihood (ML) probability of observing the feature near a story boundary. For example, if there were 100 stories, and the anchor face detection feature was true for 50 of the stories, then $p(anchor|story) = 50/100 = 0.5$. We similarly calculated the ML probabilities of an anchor face detection near a commercial boundary, outside of both story and commercial, and inside a story but outide the window of time near the boundary.

Useful features for segmentation in general are those which occur primarily near only one type of boundary, which would result in a large relative magnitude difference between these four probabilities. Ideal features, f, for story segmentation would be those for which $p(f|story)$ is much larger than the other values. For our experiments we identified features for which there was at least an order of magnitude spread in the observation probabilities across categories.

4 Results

The overarching goal of our analysis was to identify multimedia events for each source that could be used to distinguish stories from commercials and other non-story segments in the broadcast. The results of our feature selection experiments revealed several features that were important for all seven sources we analyzed, as well as other features that were important for certain sources but not others. In this section we discuss our results.

Table 2 shows the selected features for each of the broadcast sources. The first two columns show the name and type of each feature, as defined in Section 3, with start, end, and persist for durational metadata features, where relevant. The cells in the remaining columns show a "+" if the feature was automatically selected for the corresponding source; the cells are empty if the feature was not selected. The cell contains "n/a" if the feature was not available for the source; this was the case for the English-language TDT topic classification

Of the hundreds of features we analyzed, only 14 were selected for at least one of the broadcast sources. The selected features varied greatly by source, with some features being used by only one or two of the seven sources. There are only three features that were selected for each of the seven sources: *music segment persist*, *video fade start*, and *cue n-gram detected*. Two other features, *broadcaster logo detected* and *blue screen detected*, were selected for all but one of the sources. One interesting result is that these features selected for all or most sources come from all four information sources: audio, language, image, and video.

The significance of the selected features also varied by sources. For example, the *blue screen detected* feature was selected for all but one source; this feature thus has a much higher probability of occurring at certain points

Feature	Type	ALJ	BBC	CCTV	CNNH	CNNI	FOX	NWI
Cue n-gram detected	language	+	+	+	+	+	+	+
Closed captioning start	language				+		+	
TDT topic start	language	n/a		n/a			+	
Music segment start	audio	+			+			
Music segment persist	audio	+	+	+	+	+	+	+
Breath	audio	+				+		
Speaker change	audio	+		+				
Anchor face detected	image	+			+			
Blue screen detected	image	+	+	+	+		+	+
Black screen detected	image	+		+	+		+	
Broadcaster logo detected	image		+		+	+	+	+
Video transition start	video			+	+	+		
Video transient start	video		+	+				
Video fade start	video	+	+	+	+	+	+	+

Table 2: *Features automatically selected for each of the seven sources.*

than others. For two sources (ALJ and CNN), the presence of a blue screen is much more likely to occur during a commercial. For NWI it is most likely to occur at the start of a story, and for BBC it is most likely to occur outside stories and commercials. For CCTV it is equally likely in commercials and at the start of stories. For none of the sources is the blue screen likely to occur within a story.

One of the most important features for all seven sources is the *cue n-gram detected* feature derived from the automatic speech recognition output. Interestingly, the n-grams that indicate story boundaries were extremely source-dependent, with almost no overlap in the lists of words derived across sources. Table 3 shows some examples of the highest-ranked n-grams from each of the sources (Arabic and Mandarin n-grams are shown manually translated into English).

Source	Top n-gram feature
Aljazeera	here is a report
BBC America	hello and welcome
CCTV	Chinese news broadcast
CNN Headline News	stories we're following
CNN International	world news I'm
Fox News	exclusive
Newsworld International	hello everybody I'm

Table 3: *Top n-gram feature derived for each source.*

References

S. Boykin and A. Merlino, "Improving Broadcast News Segmentation Processing," *Proceedings of IEEE Multi-media Systems*, Florence, Italy, June 7-11, 1999.

J. G. Fiscus and G. R. Doddington, "Topic detection and tracking evaluation overview," In J. Allan, editor, *Topic Detection and Tracking: Event-based Information Organization*, pages 17-31. Kluwer Academic Publishers, Boston, 2002.

B. Furht, S. Smoliar, and H. Zhang, *Video and Image Processing in Multimedia Systems*, Kluwer Academic Publishers, 1995.

W. Greiff, A. Morgan, R. Fish, M. Richards, A. Kundu, "Fine-Grained Hidden Markov Modeling for Broadcast-News Story Segmentation," *Proceeding of First International Conference on Human Language Technology Research* (HLT 2001).

A. Hauptmann and M. Witbrock, "Story Segmentation and Detection of Commercials in Broadcast News Video," *Advances in Digital Libraries Conference* (ADL'98), Santa Barbara, CA, April 22 - 24, 1998.

W. Hsu, S. Chang, "A Statistical Framework for Fusing Mid-level Perceptual Features in News Story Segmentation," *IEEE International Conference on Multimedia and Expo* (ICME) 2003.

Linguistic Data Consortium, *TDT2 Segmentation Annotation Guide*, 1999. http://www.ldc.upenn.edu/Projects/TDT2/

J. Makhoul, F. Kubala, T. Leek, D. Liu, L. Nguyen,R. Schwartz, and A. Srivastava, "Speech and Language Technologies for Audio Indexing and Retrieval," in *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1338-1353, 2000.

N. O'Connor; C. Czirjek; S. Deasy; S. Marlow; N. Murphy; A. Smeaton, "News Story Segmentation in the Fischlar Video Indexing System," *Proceedings of IEEE International Conference on Image Processing* (ICIP-2001), Thessaloniki Greece.