NEW APPROACHES TO AUDIO-VISUAL SEGMENTATION OF TV NEWS FOR AUTOMATIC TOPIC RETRIEVAL

U. Iurgel, R. Meermeier, S. Eickeler, G. Rigoll

Department of Computer Science — Faculty of Electrical Engineering
Gerhard-Mercator-University Duisburg
Duisburg, Germany
e-mail: {uri, ralfm, eickeler, rigoll}@fb9-ti.uni-duisburg.de

ABSTRACT

This paper presents two new real-time approaches to segmentation of TV news shows into topics. The goal of this research work is the high precision retrieval of topics from TV news. For that purpose, the detection of correct topic boundaries is of great importance. We introduce a stochastic and a rule-based topic model based on HMMs. The former combines features from the visual as well as from the audio channel of the news show, whereas the latter uses the video channel only. They are compared to the detection of topics using only the audio channel, which is common for many other approaches. The paper contains the following innovations: 1) The detected segment boundaries correspond directly to topics and not to video or audio cuts, as most other segmentation methods. 2) An advanced stochastic topic model is introduced that uses audio as well as video features. 3) The introduced HMM-based approaches both outperform the audio-based approach. One algorithm has a very good topic boundary detection rate, whereas the other minimizes the number of wrongly inserted boundaries without missing too many real boundaries.

1. INTRODUCTION

This work has been carried out in the context of a European project on automatic topic retrieval in multimedia documents (see [5]). If these documents are from radio news, only acoustic information is available and the corresponding audio stream is segmented using well-known audio segmentation techniques. After segmentation, the segments are transcribed with a large-vocabulary broadcast speech recognizer and text-based topic detection methods are applied to the transcription result. If the source of text-based documents is the Internet, only topic detection is applied to the available text. In case of TV news, the available information consists of the video and the audio track. In this case, it is reasonable to assume that the additional information of the video track will be helpful to segment the information into topics. In this context one has to consider that the well-known methods for audio segmentation are mainly capable of detecting boundaries that correspond to abrupt changes of the audio information in the stream. Many of these boundaries will not correspond to topic boundaries, but instead to events not related to the change of a topic, such as the beginning of a report or an interview. However, a correct segmentation of the information into topics is of course important for the success of the entire information retrieval system. Furthermore, the goal of our system is to present the user the detection of a desired topic together with a clip of the corresponding video segment, and thus the beginning and ending of this video clip should be as precise as possible.

One approach for topic segmentation is the detection of significant changes in audio characteristics. Thus, speaker turns can be extracted and considered as topic turns. However, most likely there will be an over-segmentation, because an interview will be separated from the surrounding report and identified as a topic of its own. A widely used method for audio segmentation is the Bayesian Information Criterion (BIC) [1], which has been extended in our work to deliver more accurate boundaries.

The use of the video information allows for a different segmentation approach. The scenes of a news show can be split up into several content classes such as Newscaster, Report, Interview, Weather Forecast and editing effects combining these classes, i.e. Cuts, Dissolves and Wipes. This scene segmentation may serve as a starting point to topic boundary detection in a subsequent step. Eickeler and Müller [2] presented a novel approach for scene classification based on Hidden Markov Models (HMMs) which we have extended in our work in two ways in order to extract topic boundaries.

The first extension consists of a rule-based approach that evaluates the above-described HMM-based segmentation result from the video track using additional knowledge. The second extension consists of an augmented HMM approach where the news is modeled as a stochastic sequence of topics using an HMM-based topic model that makes use of video as well as of audio features. In this paper, the extended HMM method is compared to the rule-based approach and to a modified BIC algorithm that attempts to detect topics on the basis of the audio information only.

2. AUDIO SEGMENTATION

Several methods for audio segmentation have been proposed, like Akaike's Information Criterion (AIC) the Bayesian Information Criterion (BIC) [1], the Consistent AIC (CAIC) and the Minimum Description Length (MDL). These and other methods have been compared in [3] and it has been shown that with optimal parameters, almost all algorithms perform comparably well. Among the best performing methods are the BIC, the CAIC and the MDL.

The audio segmentation algorithm deployed in our work uses the BIC. It follows the method of Tritschler and Gopinath [4] which will be described briefly. The algorithm takes a window of n audio features $x_1, \ldots, x_n$ and arbitrarily places a boundary at position i, resulting in two segments. It then decides whether it is more likely that one single model θ_1 has produced the output $x_1, \ldots, x_n$, or

that two different models θ_{21} and θ_{22} have generated the two segments' output $x_1 \dots x_i$ and $x_{i+1} \dots x_n$ respectively. The decision rule to check if there is a boundary at point i is

$$\Delta BIC_i \stackrel{!}{<} 0 \qquad \text{with} \tag{1}$$

$$\Delta BIC_i = -\frac{n}{2}\log|\Sigma_w| + \frac{i}{2}\log|\Sigma_f| + \frac{n-i}{2}\log|\Sigma_s| + \frac{1}{2}\lambda(d + \frac{d(d+1)}{2})\log n. \tag{2}$$

Σ_w denotes the covariance matrix of all window feature vectors $x_1, \dots, x_n$, Σ_f and Σ_s are the covariance matrices of the features of the first and second segment respectively. d is the feature vector dimension. According to theory, the penalty weight λ should equal 1, but practical applications show better results with $\lambda \neq 1$.

If for a point i, $\Delta BIC_i < 0$, then also for some points j surrounding i there will be $\Delta BIC_j < 0$. The algorithm decides for the boundary to be at the point with the lowest ΔBIC value.

To detect all audio segments of a news show, the window is shifted over all feature vectors with varying lengths n and varying i. See [4] for details.

Implementing the above described algorithm, we have noticed that sometimes segment boundaries are set too early, roughly one or two syllables before the speaker finishes his or her utterance. Instead of considering the point i at which the minimum of ΔBIC occurs as a boundary, we choose the point k that lies in the middle of the adjacent two points at which the ΔBIC value crosses the 0 line:

$$k = \frac{l+m}{2} \quad \text{with} \quad \Delta BIC_l = 0, \Delta BIC_m = 0, l < k < m. \tag{3}$$

This modification improves the segmentation accuracy and reduces the number of boundaries appearing too early.

As feature vectors, we use 39-dimensional mel-cepstral feature vectors without mean subtraction. The penalty weight has been set to $\lambda = 3.0$.

3. AUDIO-VISUAL SEGMENTATION

The video segmentation approach presented in this paper is based on the work by Eickeler and Müller [2]. In their approach, videos are segmented into content classes that do not directly correspond to topics. We chose this approach because the overall correct classification rate is 97.3% for station dependent recognition and 93.7% for station independent recognition, thus providing a good basis for extension to a topic boundary detection system. The following six content classes are used:

Begin, End, Newscaster, Report, Interview (an interview of the newscaster and the interviewed person) and Weather Forecast.

Four classes are defined for the edit effects:

Cut (a hard cut), Dissolve, Wipe and Window Change (a change of the "topic window" next to the newscaster. This effect is used as separator between two news topics.)

Each of these classes is modeled by a Hidden Markov Model (HMM). A feature vector consisting of 12 video features represents each image. Among these features are the center, velocity and variance of motion, a modified intensity of motion, a modified difference histogram and a feature which improves the detection of dissolve edit effects. All these features are based on luminance only. Three more values are added to the vector, giving the average value of the luminance (Y) and the two chrominance (U,V) components.

The segmentation and classification of a show is the result of calculating the sequence of HMMs that most probably has generated the observed feature vector. This is supported by defining a video model which combines the scenes and the edit effects to a typical and flexible news show structure, using transition probabilities. A typical video model is depicted in Figure 1.

As mentioned previously, the described framework for news content classification does not allow for the detection of topic boundaries. Thus, extensions have to be made which we implemented in two different ways:

- After scene classification, some rules, which are based on the typical structure of a topic, are applied to extract the topic boundaries.
- Video features and audio segmentation are combined into the HMM structure. An adapted video model is used which represents topic structures.

3.1. Rule-based approach

After scene extraction by the HMM framework, which compared to [2] uses an updated video model, these rules are deployed:

- All edit effects except for the Window Change effect are ignored.
- The following transitions are considered a topic boundary:
  - Begin to Newscaster
  - Report to Newscaster
  - a Window Change which separates two Newscaster scenes
  - a Window Change with a preceding Interview and a following Newscaster
  - Report to Weather Forecast
- Everything that appears after the Weather Forecast is ignored, thus rejecting unimportant previews of other news shows.

Thus, topics that start with reports are not properly detected. But as this kind of topic rarely occurs in the most important news show of German TV, which serves as the test set for this paper, the resulting error can be accepted. A weather forecast preceded by a newscaster scene is not considered a new topic, because the newscaster announces the forecast.

3.2. Combination of audio and video into the HMM structure

To combine the audio and video information, we first segment the audio track using the modified BIC algorithm described in Section 2. The resulting boundary positions are rounded to the nearest video frame. An audio feature stream is extracted in such a way that the frame at which the audio boundary occurs gets a maximum predefined feature value (e.g. 1.0), whereas all other frames are initially assigned a value of 0. To a predefined number of frames surrounding each peak frame, values are assigned that decrease linearly and symmetrically to this frame. If two close audio boundaries cause an overlap of their feature values, the maximum value is taken.

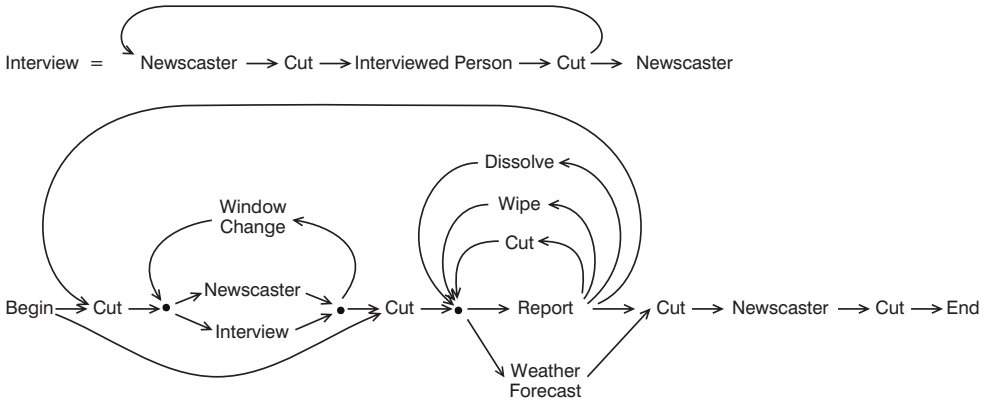

Fig. 1. Video model of a news show

This method outputs a 1-dimensional audio feature stream, which is added to the 12 video features described at the beginning of this Section. One new edit effect has been added, which we call AVcut (audio and video cut). It describes a hard video cut with an audio feature value greater than 0, i.e. a video cut with an audio cut nearby. This approach takes into account that in most cases, audio and video boundaries do not occur at the same frame but with a small displacement. Consequently, the Cut edit effect is defined as a hard video cut without any audio cut nearby.

Besides combining audio and video features, a novel news model has been introduced that reflects typical topic structures. Figure 2 depicts a slightly simplified version of this model with topic beginnings emphasized by gray circles. The model on the top represents a news show with embedded topic structures (N_topic and R_topic) and edit effects (ed_eff) that are defined below. The classes Newscaster and Report are abbreviated by N and R respectively. Square brackets denote optional elements. The interview subclass is the same as in Figure 1, except that Cuts have been replaced by AVcuts.

Transition probabilities between the model classes have not yet been incorporated into our algorithm.

Both video-based approaches not only output a topic segmentation, but they also inform about the video content classes, such as Report or Newscaster, which could be used as an additional information in a following topic identification step.

4. EXPERIMENTS AND RESULTS

As a test and training set we used recordings of the most important German TV news show. The video information was captured in a 192 x 144 pixel resolution at a frame rate of 12.5 fps. The audio track has been sampled with 16 kHz and 16 Bit resolution. All shows last 15 minutes. Nine shows, for a total time of 2:15 hours, have been used for training and testing each algorithm. For the experiments, we used the hold-out method, i.e. we tested each show with a system trained on the other eight.

For each of the three approaches, we have counted the numbers of boundaries in the test set that have no correspondence in the reference topic segmentation, giving the number of insertions. The number of insertions is divided by the total number of cuts in the test result, giving the relative number of test boundaries that do not correspond to real topic cuts (insertion rate). The number of boundaries in the reference that are not detected by the algorithm yields the number of deletions. The deletion rate is the relative number of real topic boundaries that remain undetected. A tolerance range of 2 s is applied when matching reference and test boundaries.

algorithm	audio	video	audio-visual
insertions	81.2 %	11.8 %	35.2 %
deletions	23.3 %	17.8 %	8.5 %
precision	18.8 %	88.2 %	64.8 %
recall	76.7 %	82.2 %	91.5 %

Table 1. Segmentation results for the three algorithms

All figures are shown in Table 1, together with the recall and the precision rates defined accordingly to [3]. The precision rate equals one minus insertion rate, the recall rate equals one minus deletion rate.

4.1. Audio topic detection using the BIC

The first test series checks the performance of our modified BIC audio segmentation which is described in Section 2. Each audio boundary is supposed to be a topic boundary, as no further information about the audio content is available. The results show the limited information that can be drawn from the audio track for topic segmentation. As expected, the number of insertions is very high. This is due to the fact that interviews within a report are considered topic cuts. Background noise or voices may result in additional audio cuts.

4.2. Video topic detection using HMMs and a rule-based approach

We then tested the performance of the rules applied to the HMM segmentation output (see Section 3.1). As they are quite simple and cannot detect topics beginning with reports, this algorithm tends to under-segment. Its insertion rate is significantly lower than the audio-visual approach, but more boundaries remain undetected.

The advantage of this algorithm is that it tends not to cut a topic into two or more, but it rather combines two topics to one.

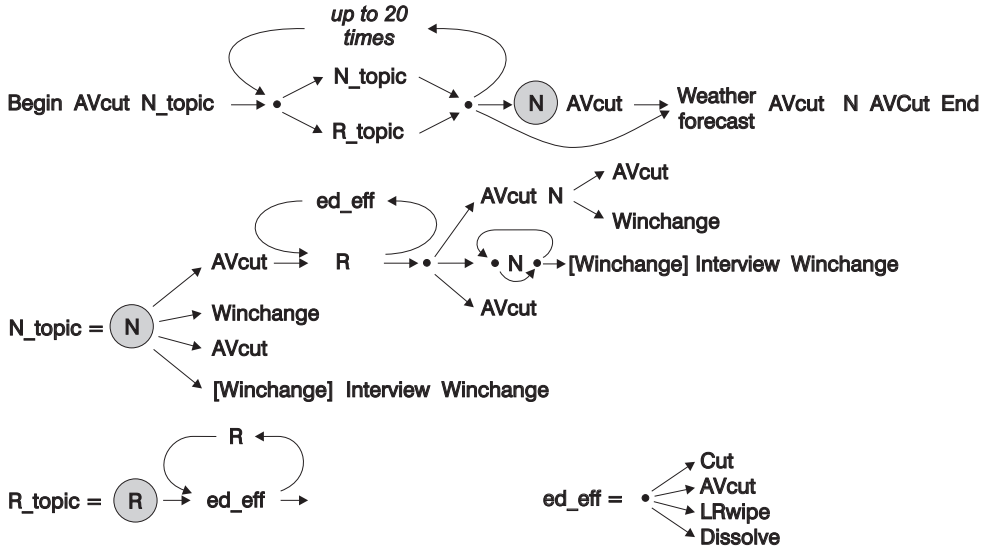

Fig. 2. Model of a news show with embedded topic structures

This is desirable for applications such as automatic media monitoring, for which it is better to deliver more than the wanted topic instead of delivering only a part of it. (A topic indexing system would have to consider oversized topics by assigning two or more topic keywords which describe *both* topics.)

4.3. Audio-visual topic detection using the BIC and HMMs

The results we achieved for the combined audio and video approach (see Section 3.2) show its great ability to detect almost all topic boundaries that are present in the news show. However, there are quite a high number of boundaries that are mistakenly inserted. This leads to over-segmentation, which could be compensated for by a subsequent topic identification step that clusters the segmented parts with the same topics. With its high recall rate, this algorithm is very well suited for cases that need to detect all boundaries and can cope with over-segmentation in the above described way.

5. CONCLUSIONS AND FUTURE WORK

We have presented two new approaches for audio-visual topic segmentation of TV news and compared them to an audio-based algorithm. Tests indicate their suitability for this task. The incorporation of visual information performs significantly better than the audio-only approach.

The results of the audio only segmentation approach could be improved using more sophisticated algorithms. For example, two audio cuts near each other (within a certain range) could be considered as one. Speaker classification and clustering could reduce the bad effect of interviews or background noise. However, we believe that using the audio track only poses limitations as to the ability of detecting topic boundaries.

The results of the video, HMM- and rule-based algorithm could be improved by a better consideration of the different topic structures. More editing effects, not only the Window Change, should be considered, although they do not imply a topic turn in the same way. Also, the duration of a topic or the length of preceding scenes could be taken into account. As for the audio-visual HMM approach, we expect better results if we modified the news model, for example by changing its structure or incorporating transition probabilities. As one important reason for over-segmentation is that the algorithm loops too many times through the R_topic-subclass (see Figure 2), we are confident that we can significantly reduce the insertion rate.

It should be emphasized here again that our new algorithms are capable of detecting real *topic* boundaries instead of boundaries consisting mainly of audio or video edit effects, as most other algorithms do. The promising results of our approaches convince us that they make a good basis for future improvements in the field of topic segmentation.

6. ACKNOWLEDGMENTS

The work presented in this paper has been carried out within the European Union project ALERT (IST-1999-10354) [5].

7. REFERENCES

[1] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[2] Stefan Eickeler and Stefan Müller, "Content-based video indexing of tv broadcast news using hidden markov models," in *Proc. IEEE ICASSP*, 1999, pp. 2997–3000.

[3] Mauro Cettolo and Marcello Federico, "Model selection criteria for acoustic segmentation," in *Proc. of the ISCA ITRW ASR2000 Automatic Speech Recognition*, Paris, France, 2000, pp. 221–227.

[4] Alain Tritschler and Ramesh Gopinath, "Improved speaker segmentation and segments clustering using the bayesian information criterion," in *Proc. EUROSPEECH*, 1999, vol. 2, pp. 679–682.

[5] "Alert homepage," http://alert.uni-duisburg.de/start.html.