News Talk-Show Chaptering with Journalistic Genres

Delphine Charlet, Géraldine Damnati, Jérémy Trione
Orange Labs, Lannion, France
{firstname.lastname}@orange.com

Abstract
Modern TV or radio news talk-shows include a variety of sequences which comply with specific journalistic patterns, including debates, interviews, reports. The paper deals with automatic chapter generation for TV news talk-shows, according to these different journalistic genres. It is shown that linguistic and speaker-distribution based features can lead to an efficient characterization of these genres when the boundaries of the chapters are known, and that a speaker-distribution based segmentation is suitable for segmenting contents into these different genres. Evaluations on a collection of 42 episodes of a news talk-show provided by the French evaluation campaign REPERE show promising performance.

Index Terms: Chapter generation, Structural segmentation

TV-Content Segmentation

1. Introduction
Segmenting audio or audiovisual contents into homogeneous parts is a key step for indexing these data. This segmentation can be directly used to smartly browse the content, and is also a pre-processing step for further classification or meta-data extraction.

For video contents, many types of segmentation are possible, according to the description unit considered. Speaker Diarization is one of the possible segmentations, that describes the content in terms of “who speaks when”.

Many research works are also conducted on higher level segmentation, more prone to semantic interpretation. For instance, story segmentation has received much attention in the last decade (e.g. [1][2][3] on TV Broadcast News) and more recently, video scenes segmentation has also been studied (e.g. [4] on documentary films and [5] on Ally McBeal TV series). Story segmentation usually relies on an initial fine-grain segmentation in breath groups while video scene segmentation relies on a segmentation in video shots.

Once the segmentation is provided, identifying the topics of these homogeneous chapters is the next step for a better description of the content, where the topic identification is seen as a mono or multi-label classification task (e.g. [6]). Identifying the genre of the video has also been explored, for homogeneous (or already segmented) contents (e.g. [7] on very distinct genres such as sport, musics, commercials, cartoons and news).

In this paper, we are interested in structural segmentation and characterization of News contents. By “structural”, we mean that a description of the content is extracted, independently of the thematic content. Indeed, modern TV and radio news talk-shows are long contents (often 1 hour), involving different journalistic sequences, which obey specific patterns, for instance debates, interviews, reports, etc. It is thus of interest to automatically extract the structure of these contents, according to a set of journalistic genres, to enable a better browsing in the content, or to apply further specific processing on each journalistic sequence.

In this work, we propose a specification of journalistic genres for news talk-show. We investigate how these genres can be automatically segmented and characterized.

Section 2 explains the problem, on a concrete case. In Section 3 to 5, methods for segmenting and classifying these chapters are investigated. Experiments and discussion are reported in Section 6.

2. Journalistic genres for news talk-shows
Modern TV or radio news talk-shows are divided chapters, corresponding to different journalistic genres. We adopt the following typology of genres:

- Interview: a rather long interaction between one guest and at least one journalist. This discussion can include short reports and excerpts of other people to make the guest react.
- Debate: a rather long interaction between at least two guests and at least one journalist. This discussion can also include short reports and excerpts of other people to make the guests react.
- News: all the informational content which is not Interview nor Debate; news can contain reports, anchor speech, short interaction between anchor and guest or reporter, journalists speech, ...
- Advertising: Commercials, Jingle, Announcement of next shows...

For instance, in the corpus provided by the French evaluation campaign REPERE [8], a collection of 42 episodes of the program BFM-Story is available. BFM-Story is a daily one-hour-long news program that implies sequences of Interview, Debate, News and Advertising according to the definitions given above.

Figure 1 represents 30 episodes of the corpus, in terms of genre segmentation. Each line represents the timeline of a show with the corresponding genre with a one-minute resolution.

It can be seen that, although there is some similarity between different episodes of the show, the number of boundaries, as well as their time location differ from one episode of the show to another. The aim of our work is to automatically retrieve these chapters and we avoid using a priori structure knowledge in order to be robust to editorial changes, and to be more easily transposed to other shows.

The problem of segmenting and characterizing Advertising is discarded, as highly efficient methods based on audio and
video fingerprinting exist [9] to identify jingles that are broadcasted at the beginning and at the end of an Advertising segment. Hence, in this paper, the proposed task is to automatically find the genre structure of portions of show in between 2 Advertising segments.

![Fig.1 TV show structuration](image)

3. **Journalistic genre classification for news talk-shows**

The very first question raised is that if these above-defined journalistic genres can be characterized automatically. In a preliminary approach, it is assumed that the content has already been segmented, and that the task is to classify a homogeneous segment into News, Interview and Debate.

In this work, we restrict our investigations to audio features, for the sake of generalization to radio contents. It is assumed that for a given content, the automatic speech transcription is available, as well as the speaker diarization (speaker segmentation and clustering). Two families of potential features are investigated to characterize News, Interview and Debate, derived from speech transcription and speaker diarization: linguistic features and speaker-based features. Such families of features have been investigated for instance in the task of video genre classification and have shown interesting performance [10]. Then, these features will feed a classification process based on icsiboost [11] (a large margin classifier based on a boosting method of weak classifiers).

3.1. **Linguistic features**

Debates and Interviews are interactions, whereas News are more narrative. It is thus suspected that some linguistic features derived from part of speech (POS) tagging could help to characterize genres. To be independent of the size of the segment to classify, ratio of use of specific POS are computed. We focus on the use of personal pronouns. As a matter of fact, in interactions, first-person and second-person pronouns are used, whereas they are hardly used in narrative news. We also investigate on the use of verbs, distinguishing the infinitive and the conjugated verbs.

3.2. **Speaker-based features**

Speaker-turns distribution is different for each journalistic genre. In Interviews, speaker-turns are mainly exchanges between one guest and at least one journalist, where the guest speaks longer than the journalist(s). In a Debate, there are several guests, and the speech time should be equally balanced between the guests. In a News part, there may have an anchor speaker which appears regularly throughout the chapter, and many speakers which appear punctually on a given subject. Moreover, overlapped speech could be more important in debates than in the other genres. These remarks lead to investigate the following potential features:

- coverage rate of the N main speakers (the ratio between the speech duration of the N speakers who speak most and the total speech duration)
- dual interaction rate, computed as follows:
  - for each pair of adjacent speakers (independently of their order), compute the number of speaker changes covered by this pair, select the pair with the maximum number.
  - compute the ratio between the number of speaker changes due to this maximal pair and the total number of speaker changes
- number of speaker-turns per second
- ratio between overlapped speech duration and total speech duration

4. **Segmenting into homogeneous journalistic genre segment**

In [12], we have proposed to extend a topic segmentation algorithm based on lexical features to a topic segmentation based on speaker distribution features. In contents where speaker structure was closely related to topic structure, this approach yielded very good performance. In the same vein, we argue that the speaker structure is closely related to the journalistic genre structure, and that finding significant changes in the speaker distribution could help to find the change of journalistic genre.

The proposed segmentation algorithm follows a classical sliding window-based segmentation scheme. It is composed of 3 steps:

- **Speaker cohesion curve**: to each speaker turn is associated a score of potential segmentation boundary. This score is based on a dissimilarity measure between a fixed-length window ending at the speaker turn, and a fixed-length window beginning at the speaker turn. The length of the window can be defined in number of speaker turns or in seconds.
- **Local maxima detection**: a recursive process detects local maxima on the curve, within a given segment, and, if the local maximum is above a given threshold, the process splits this segment into 2 segments. The process is initialized with one segment containing the whole curve and is recursively applied on each new segment, while the segments are long enough (minimal segment length constraint) and the local maxima are above the threshold.
- **Validation step**: to confirm or reject potential boundaries detected from the previous local maxima detection step. For each potential break \( k \), a dissimilarity measure is computed, between the previous segment \([k-1, k] \) and the next segment \([k, k+1] \). Potential breaks are sorted according to their dissimilarity measure. The minimal dissimilarity break is deleted from the list of breaks, if its value is under a given threshold. The algorithm is repeated iteratively until all the potential breaks have a score above the threshold. The interest of the validation step is to compute dissimilarity measure on windows...
In section 3, the features used for classification are computed on the set of speaker turns of a given segment. Here, the same type of features are used, but, as they have to be associated to each speaker turn, they are actually computed on a window of speaker turns, centered on the current speaker turn. Experiments are conducted with the toolkit CRF++ (crfpp.googlecode.com), which requires symbolic features. Thus, a discretization step is necessary to use our classification features in CRF++. This step is performed thanks to [13], which implements [14].

6. Experiments and results

6.1. Corpus

The corpus is made of TV shows provided by the French ANR REPERE challenge [8], and focused on BFMStory shows. BFMStory is a news talk-show which combines several genres in the same show, as shown in section 2.

The corpus contains 42 episodes of BFMStory show, of about 55 min each, recorded between May 2011 and November 2012. We consider only the portions in between Advertising. Thus, the set of 42 episodes translates into a set of 122 sub-episodes without Advertising. This corpus contains a total set of 295 boundaries between News, Interview or Debate. Table 1 describes the corpus in terms of its composition in journalistic genres. Here, a segment is a journalistic genre homogeneous chapter. It is composed of multiple speech turns.

<table>
<thead>
<tr>
<th>genre</th>
<th>number of segments</th>
<th>average segment duration [min-max]</th>
<th>number of speech turns</th>
</tr>
</thead>
<tbody>
<tr>
<td>News</td>
<td>126</td>
<td>381s [28-931]</td>
<td>2335</td>
</tr>
<tr>
<td>Interview</td>
<td>150</td>
<td>405s [129-1146]</td>
<td>2887</td>
</tr>
<tr>
<td>Debate</td>
<td>36</td>
<td>569s [115-944]</td>
<td>1294</td>
</tr>
</tbody>
</table>

Table 1. BFMStory segment distribution per genre

For the classification algorithm with icsiboost, and for the joint segmentation/classification algorithm with CRFs, a training phase is required. For the segmentation algorithm, there is no training phase, but a set of free parameters (sliding window size, minimal segment length and thresholds for boundary detection) has to be optimized on a given corpus, which can be seen as a training corpus for these parameters. As the TV corpus is too short to be divided in train/test, the whole corpus of 42 shows is divided in 3 folds of 14 shows. Each fold is considered as a test set, and processed with the set of parameters trained on the 2 other folds considered as train set. The results presented in the tables are obtained on the whole corpus of 42 shows, when cumulating the results obtained separately on each fold.

6.2. Corpus pre-processing

For each show, automatic speech transcription as well as speaker diarization are provided. Speech transcription is performed with the Vocapia speech to text system[15], which gives for this corpus a WER of 19%. Part of Speech tagging is performed with lia_tagg software[16]. Speaker diarization is a classical BIC/CLR two-step process, with special attention paid for overlapped speech [17]. The Diarization error rate, for this set of TV shows, is 10.3%.

5. Segmentation and classification

For a News talk-show composed of several segments from different journalistic genres, the task consists in both segmenting and classifying into journalistic genres. This can be done sequentially, by applying first the segmentation step described in section 4 and then the classification step of section 3 on automatic segments. A process that jointly performs segmentation and classification is also possible. Conditional Random Fields (CRF) are widely used for the task of joint segmentation and labelling. Thus, they are used here as a baseline process, for sake of comparison with the sequential approach.

CRFs are a powerful tool to train and predict sequence labelling. Here, the observation unit is the speaker turn, and we want to predict the labels (News, Debate or Interview) for sequences of speaker turns.

In this paradigm, each speaker turn is associated to a set of descriptors, and for the training phase, to a label of journalistic genre. In the test phase, the CRFs are applied on the unlabeled speaker turns, to find the sequence of labels.
6.3. Evaluation Measure

Segmentation task:
The evaluated task is a boundary detection task, with 2 types of errors: miss detection and false alarm, which leads to 2 evaluation measures (precision and recall), along with their harmonic mean, the F-measure. As usual in this kind of evaluation, a tolerance margin around the reference is considered. Here, as the potential unit for segmentation is the speaker turn, the tolerance margin is equal to one speaker turn, or 10s (if the speaker turns adjacent to the boundary are shorter than 10s).

Classification task:
For reference segmentation, genre labelling evaluation is a closed-set evaluation task, which is evaluated through a correct classification rate. It is performed at the segment level, and can be translated into an evaluation at the speech turn level. For each genre, a detailed evaluation of its detection is also performed with Precision (Prec), Recall (Rec) and F-measure.

For automatic segmentation, the evaluation of genre labelling is restricted to an evaluation at the speech turn level, as it is always possible to give a reference genre label at the speech turn level, which is not the case for the automatic segment level (which may merge several different genre reference segments).

6.4. Results and discussion

Table 2 presents the classification results, at the segment level, obtained for reference segment labeling, when considering different subsets of features and using icsiboost classifier. Results between brackets are the corresponding results, when evaluated at the speech turn level. The results show the potential of each set of features, linguistic or speaker-based, and their good complementarity. The Debate genre is the most difficult to detect, but it can also be due to the fact that it is the less numerous class, which may be poorly modeled in the training phase.

Performances at speech turn level are slightly better. This can express the fact that the longest segments are easier to label, and thus, they have more correctly labelled speech turns.

<table>
<thead>
<tr>
<th></th>
<th>Global classif. rate</th>
<th>News F-measure (Prec/Rec)</th>
<th>Interview F-measure (Prec/Rec)</th>
<th>Debate F-measure (Prec/Rec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>linguistic features</td>
<td>75.8</td>
<td>81.6 (83.9/79.3)</td>
<td>76.7 (73.7/80.0)</td>
<td>50.0 (54.5/46.2)</td>
</tr>
<tr>
<td>speaker-based feat.</td>
<td>85.5</td>
<td>92.6 (89.8/95.6)</td>
<td>86.8 (89.0/84.8)</td>
<td>62.8 (64.0/61.5)</td>
</tr>
<tr>
<td>linguistic+ speaker-based feat.</td>
<td>92.4</td>
<td>94.0 (93.5/94.6)</td>
<td>92.5 (91.8/93.3)</td>
<td>85.7 (91.3/80.8)</td>
</tr>
</tbody>
</table>

Table 2: Genre labelling performance, for reference segmentation

Table 3 shows the performance of automatic genre segmentation, with the speaker-based algorithm presented in section 4, and with the CRF presented in section 5. CRFs perform very poorly to detect genre change, whereas the speaker-based segmentation gives interesting results. For the speaker-based segmentation, precision errors are mostly false alarms during news containing long reports. Indeed, when the reports are long enough, they lead to a specific speaker distribution in these reports, which is captured by the algorithm, and leads to genre change decision. When it comes to miss detection, they are mainly due to debate/interview changes.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRF</td>
<td>34.6</td>
<td>19.1</td>
<td>24.6</td>
</tr>
<tr>
<td>Speaker-based</td>
<td>71.4</td>
<td>64.7</td>
<td>67.9</td>
</tr>
</tbody>
</table>

Table 3: Genre segmentation performance (boundary detection)

When including Ads boundaries in the evaluation set, the F-measure performance of the speaker-based segmentation algorithm increases from 67.9% to 76.1%, which means that boundaries of News, Interview or Debates with Ads are particularly well detected based on speaker distribution.

Table 4 shows the result of genre labeling on automatic segments, with an evaluation at the speech turn level. The sequential approach refers to the approach where the classification is performed with icsiboost classifier, on the segments resulting from the speaker-based segmentation algorithm. Although performing very poorly for segmentation, CRFs are not so bad when restricting the evaluation to the classification, which means that CRFs have captured some of the characteristics of the journalistic genres. Indeed, their very bad performances in segmentation are due to the miss detection of short segments (of few speech turns), which have a strong impact on the segmentation evaluation, and a low impact on the classification evaluation. Finally, the sequential method performs best, and the decrease of performance due to automatic segmentation is moderate (an improvement on Debate is even observed, where automatic segments have focused on parts of Debates which have been better classified than the whole reference segment).

<table>
<thead>
<tr>
<th></th>
<th>Global classif. rate</th>
<th>News F-measure (Prec/Rec)</th>
<th>Interview F-measure (Prec/Rec)</th>
<th>Debate F-measure (Prec/Rec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRFs</td>
<td>76.3</td>
<td>81.4 (84.4/78.6)</td>
<td>77.6 (74.1/81.3)</td>
<td>63.7 (67.0/60.7)</td>
</tr>
<tr>
<td>sequential approach</td>
<td>90.0</td>
<td>90.4 (87.7/93.2)</td>
<td>90.1 (91.1/89.2)</td>
<td>89.1 (91.9/86.3)</td>
</tr>
</tbody>
</table>

Table 4: Genre labelling performance, for automatic segmentation

7. Conclusion

We have proposed to describe news talk-show in terms of sequence of journalistic genres, such as News, Interview, Debate. Automatic characterization of such genres, based on linguistic and speaker-based features is investigated. An algorithm of genre segmentation, directly inspired from a topic segmentation algorithm translated to speaker-space is proposed. Evaluations on a collection of 42 episodes of a French TV talk-shows show that linguistic and speaker-based features enable a good characterization of such genres when the segmentation is provided, suggesting that it could be successfully applied to mono-genre content characterization. Furthermore, the proposed genre segmentation algorithm on the basis of speaker distribution provides good performances and enables promising results on fully automatic journalistic genre segmentation and classification.
8. References


