

F. Vallet, S. Essid, J. Carrive and G. Richard, High-level TV talk show structuring centered on speakers' interventions, In book: TV Content Analysis: Techniques and Applications, Y. Kompatsiaris, B. Merialdo and S. Lian (Eds.), CRC Press, Taylor Francis LLC, 2012.

High-level TV talk show structuring centered on speakers' interventions

Félicien Vallet, Slim Essid, Jean Carrive and Gaël Richard

July 25, 2011

Chapter 1

High-level TV talk show structuring centered on speakers' interventions

Abstract

Archives professionals have high expectations for efficient indexing tools. In particular, the purpose of archiving TV broadcasts has created an expanding need for automatic content structuring methods. In this chapter, is addressed the task of structuring a particular type of TV content that has been scarcely studied in previous works, namely talk show programs. The object of this work is examined in the light of a number of sociological studies, with the aim to identify relevant prior knowledge on the basis of which the approach followed is motivated. Of particular interest is the fact that, while these broadcasts show a wide disparity of aspects, they also share common features that have been studied in semiological works. For instance, the presence of guests, the polymorphy of the program (a succession of interviews, performances, TV reports, etc.) or the role of the anchorman: using very specific and well-defined codes. These common features, but also the differences, are highlighted by a comparison between two French talk shows from two different periods: “*Le Grand Échiquier*” (1980s) and “*On n’a pas tout dit*” (2010s).

Building upon that effort and clarifying this rather vague notion of structuring, specific use cases are considered (e.g. the creation of a table of contents for efficient browsing). A taxonomy of the talk show components is then presented, including concepts such as the guests, the hosts, the performances, etc. that suggests a generic high-level structure retaining the main cross-program invariances. These considerations stress out the relevance of considering the speakers' interventions as elementary units in place of the video shots usually employed

in structuring studies.

Subsequently, is proposed a panorama of various segmentation and detection techniques aiming at isolating the talk show components stated above. The necessity of multimodal approaches, exploiting both audio and video streams, and the potential of discriminative kernel-based methods against more traditional generative approaches is highlighted.

Finally, in the last section, a specific structuring scheme for the talk show “*Le Grand Échiquier*” is presented, targeting a hierarchical structure at the basis of which speakers' interventions are placed. This brings up the difficult question of designing appropriate evaluation procedures. Hence, a metric is proposed for the evaluation of the quality of the structures obtained in the light of the task considered. It is also shown that a very valuable structure can be obtained without necessarily recurring to sophisticated inference or reasoning techniques once a number of critical talk show features have been detected with a realistic level of accuracy.

Introduction

In the televisual domain, talk show programs are defined as broadcasts where one or several persons discuss various topics put forth by a host. Despite its history and popularity, this genre, originally invented in the 1950s, has been scarcely investigated.

This study proposes an analysis of this particular type of TV content with a view to suggest a *generic* talk show structuring scheme. If it seems that there is no consensual definition for “*structuring*”, which is a dedicated term for this type of process, one may however agree that its goal is the organisation of content in sections conveying a proper type of information. In the field of audiovisual content analysis and indexing, several works were conducted on structuring. In [33] and [36], various approaches to extract the inner arrangement of TV streams were proposed. Also, in the domain of sports videos, [16] and [27] modeled structures for the automatic parsing of tennis broadcasts. Similar works may also be found for football [54], baseball [57] or basketball [58, 55] match videos. However, to our knowledge, no analogous studies exist for talk show programs.

A number of semiotic works have studied the mechanisms of the talk shows, hence they offer significant prior knowledge on the structure of such programs. They stress out invariant aspects that appear to be characteristic of this genre of TV content. Since using this type of insight, the approach followed here may be considered as semiotically motivated.

The outline of this chapter is then the following. After examining the talk show from a semiotic point of view, this task of talk show structuring is better specified by discussing specific use cases (e.g. the creation of a talk show table of contents for efficient browsing). These considerations stress out the relevance of considering the speakers' interventions as elementary structural units for the extraction of generic talk show components. Thus, a panorama of various state of the art segmentation and detection methods aiming at isolating the previously defined components is drawn, highlighting the importance of speaker oriented detectors. However, the use of automatic speech recognition approaches is discarded since it would lead to language dependent and thus non-generic systems. The accent is thus put on speaker diarisation methods. Finally, a specific structuring scheme is proposed. This scheme is assessed in light of a specific use case through a user-based evaluation and conclusions are suggested.

1.1 Prior knowledge on the talk show programs

Historically, after the Second World War, politicians and intellectuals defined the missions of television broadcasters as the trilogy: information, education and entertainment [8]. At this time, radio and television were mostly produced by public national agencies, such as ORTF (Office de Radiodiffusion Télévision Française) in France. With the multiplication of TV sets in households and the advent of private networks, the television world underwent a slow but constant transformation. Due to the logic of profit peculiar to any industry, entertainment took an increasingly important place to respond to the need of gathering together a broader public and winning the audience ratings battles [7]. Thus, game shows, talk shows, variety shows, and reality TV went to occupy an always larger chunk of the program grid.

1.1.1 A semiotic point of view

Talk shows appeared in the 1950s and became since greatly popular. Over the years, scholars got more and more interested in studying this type of broadcast as it became clearer that it was returning an interesting image of the society.

In regard to the taxonomies proposed in semiotic works [8, 14, 52], talk show programs are positioned at the crossroads of informative, cultural and entertaining TV programs. They generally feature celebrity guests talking about their professional and/or personal lives as well as their latest films, TV shows, music recordings or other projects they wish to promote to the public.

However, defining the talk show as a genre has been an important issue due to the difficulty of developing a coherent picture of forms that vary so much from one another. The difficulty lies in identifying generic components from one show to another due to the heterogeneity of existing shapes [8, 29]. Since its invention the talk show format has been adopted in about every country. Thus, some cultural differences occur in the form of the show. For instance, studio sets may vary quite largely from one show to another. In French shows, the public generally surrounds the center of attraction that the couple host/guest forms [28], while in the USA a stage set is usually placed behind it. Besides, the host may seat behind a desk as it usually happens with American talk shows or around a table, with the guests, for the French ones. Furthermore, the host may welcome these guests one at a time or all together. On American talk shows, the main presenter is often a humorist who opens the program with

a monologue while in France he/she usually gives an introduction of the guest(s) and a brief outline of the show. Finally, in the USA, talk shows fall in two categories: daytime talk shows dealing more with public affairs and interviewing experts, and the more traditional late night talk shows following the informal host/guest format and wrapped around with comedic or musical segments.



Figure 1.1: Various talk show studio sets — Creative Commons license   (see [1] for details).

However, these differences tend to vanish nowadays since a general globalisation process is observed, as it is in many other cultural fields. An important amount of semiotic studies, such as [28, 29, 31], tried to spell common features or components out of the disparity observed. The most striking of these generic components displayed by talk shows is the constant presence of host(s) and guest(s). The numbers may vary but the couple host/guest is always the center of attention in these programs. The polymorphic form of the show is another of the common features. Talk shows are usually built as a succession of interviews, musical passages, TV reports, film excerpts, jingles, etc. For each talk show, this succession reflects an inner

organisation of the content.

Talk show programs share the particularity of being structured around natural conversation. As shown in [7, 20, 22, 46], no matter how spontaneous it seems to be, this talk is nevertheless highly planned and formatted. It is anchored around the host who is responsible for the tone, the direction and for setting limits on the talk. In [13, 34], it is shown that the language used during the talk show is very well defined. There are for instance very precise ways and codes to introduce a guest, launch a live performance or change the subject of a conversation. The aim is to set for the viewer all the aspects of a daily conversation between the host and his/her guest while rigorously following a script written in advance all along the program. Thus, while the course of the show may seem rather erratic, it is highly structured. Transitions are carried out by the host, but clear delimitations are also generally employed to switch from one part to the other. Jingles, cutaway shots, advertisements or even applause are used as demarcation lines. The talk itself is precisely monitored, the host evoking in general only one discussion subject at a time with the guest in a speech section. Therefore, the comprehension of the internal organisation of the speakers' interventions, and the talk in general, is the foundation for talk show automatic structuring.

1.1.2 Comparison between two talk shows

Having identified common characteristics of the talk show from the semiotic point of view, it is of interest to evaluate their validity by comparing different programs. For this purpose we consider two collections of French programs: “*Le Grand Échiquier*” that was broadcasted monthly between 1972 and 1986 and presented by Jacques Chancel and “*On n’a pas tout dit*” that went on air each week from monday to friday between 2007 and 2008 and was presented by Laurent Ruquier. Both shows were designed to appeal to a wide audience. They were recorded thirty years apart, which has to be taken into account since the televisual production approach went through notable changes in the meantime. These two programs are thus supposed to be sufficiently representative of the talk show category presented in 1.1.1. The statistics given on audiovisual events (Table 1.1) were obtained on manual annotations of 22 shows for “*Le Grand Échiquier*” and 5 for “*On n’a pas tout dit*”. The corpus “*Le Grand Échiquier*” has been the object of previous works (see for example [2, 10, 24, 49]) and is used again in this study while “*On n’a pas tout dit*” [6] is used to validate the genericity of the approach proposed and can thus be of smaller size. This explains the imbalance between the

number of annotated shows. Speech statistics in Tables 1.2 and 1.3 were computed using 6 speaker-annotated shows of “*Le Grand Échiquier*”.

The study of the main constitutive parts of the two talk shows highlights some major differences. Table 1.1 shows the importance of several audiovisual events. “*Musical performance*” stands for parts where live music is played, “*non-musical performance*” for live performances where no music is produced (e.g. poetry reading, circus act, etc.) and “*inserts*” for any off-sequence shown on screen such as duplexes, TV reports, film excerpts, etc. Some of these events are unspecified, either because they do not exist or because they cannot be computed, as the number of speech sections for instance. First, as expected, the formats of the programs are radically different. In the case of “*Le Grand Échiquier*”, the show lasts on average a bit less than 3 hours but with great variations while for “*On n’a pas tout dit*” it is very precisely set around 50 minutes. “*Applause*” and “*talk*” components also contribute to characterise the reproducible pattern of “*On n’a pas tout dit*”, the standard deviations being very low compared to those obtained for “*Le Grand Échiquier*”. No “*jingles*” are used in “*Le Grand Échiquier*” which could have been expected since it was not a common procedure in the TV world at that time.

event	avg. length		standard dev.		avg. number		standard dev.	
	GE	OAPTD	GE	OAPTD	GE	OAPTD	GE	OAPTD
corpus								
show	165'24" (100%)	50'41" (100%)	25'13" (15.2%)	4'03" (8.0%)	-	-	-	-
participants	-	-	-	-	16.2	11.6	5.7	3.5
music perf.	85'43" (51.8%)	36" (1.2%)	25'07" (15.2%)	0 (0%)	26.4	1	8.5	0
non-music perf.	9'04" (5.5%)	-	16'17" (9.8%)	-	2.4	-	3.9	-
inserts	16'55" (10.2%)	45" (1.5%)	10'47" (6.5%)	37" (1.2%)	7.8	1	5.8	0.7
speech	40'31" (24.5%)	39'32" (78.0%)	23'31" (14.2%)	4'37" (9.1%)	-	-	-	-
laughter	2'35" (1.6%)	3'59" (7.8%)	4'27" (2.7%)	2'35" (5.1%)	41.3	23.3	66.1	10.7
applause	10'34" (6.4%)	5'26" (10.7%)	4'26" (2.7%)	41" (1.3%)	38.2	33.0	18.6	9.1
jingles	-	24" (0.8%)	-	1" (0%)	-	1	-	0

Table 1.1: Audiovisual events statistics for 22 shows of “*Le Grand Échiquier*” (GE) and 5 shows of “*On n’a pas tout dit*” (OAPTD). Durations are given in minutes/seconds and percentages reflect the part of each event in the talk show.

Also, while a great diversity is observed in the form of the show itself for “*Le Grand Échiquier*” with numerous excerpts or musical passages and still a very important “*talk*” component, “*On n’a pas tout dit*” mostly features dialogues: 78% of the show duration against 25% or 39% (according to which of Table 1.1 or 1.2 is considered). This difference can be explained by the nature of the show, centered on the life and achievements of one or several guests in the first case, and on humoristic chronicles in the second one as assessed by the large amounts of applause, laughter and speech turns (see Table 1.2), where speech turns

account for the number of changes from one active speaker to another. The great amount of speech turns indicates very short exchanges between participants. These are typically the differences between the late-night and daytime talk shows presented earlier. It explains the disparity observed in the treatment of the talk itself. In “*Le Grand Échiquier*”, two persons share about 70% of the speech load - namely the host and the main guest - while in “*On n’a pas tout dit*”, the speech is more “democratically” distributed as can be seen in comparing the average speaker percentage (see Table 1.2). It is also notable that “*Le Grand Échiquier*” features on average four more participants than “*On n’a pas tout dit*” with 16 against 12.

event	average		standard dev.	
	GE	OAPTD	GE	OAPTD
corpus				
show	165'17" (100%)	50'41" (100%)	24'41" (14.9%)	4'03" (8.0%)
speech	64'22" (38.9%)	39'32" (78.0%)	21'15" (12.9%)	4'37" (9.1%)
speech overlap	5'10" (8.0%)	3'17" (8.3%)	2'31" (3.9%)	1'38" (4.1%)
speech turns	1264	1348	432	237
1st dominant speaker	26'08" (40.6%)	14'52" (37.6%)	4'03" (6.3%)	1'49" (4.6%)
2nd dominant speaker	17'31" (27.2%)	6'15" (15.8%)	6'57" (10.8%)	1'02" (2.6%)
average speaker	4'38" (7.2%)	3'43" (9.4%)	1'02" (1.6%)	1'11" (3.0%)
average speech segment	3.1"	2.2"	0.6"	0.5"

Table 1.2: Speech statistics for 6 shows of “*Le Grand Échiquier*” (GE) and 5 shows of “*On n’a pas tout dit*” (OAPTD). Durations are given in minutes/seconds and percentages reflect the part of each event. Except for the first two lines percentages are computed over the total speech duration.

However, while these two talk shows seem to display rather distinctive features, particularly in their inner construction, they still share a very similar distribution of the talk. In both programs, the spontaneity is put forward and the average speech intervention is about 2.5-second long. Besides, the importance of the overlap reinforces the feeling of natural conversation.

event	average		standard dev.	
	GE	OAPTD	GE	OAPTD
corpus				
speech	23'43" (36.8%)	12'38" (32.0%)	6'58" (10.8%)	1'08" (2.9%)
speech turns	498 (39.4%)	398 (29.5%)	151 (12.0%)	118 (8.8%)
average speech segment	2.9"	2.0"	0.5"	0.5"

Table 1.3: Speech statistics for the host for 6 shows of “*Le Grand Échiquier*” (GE) and 5 shows of “*On n’a pas tout dit*” (OAPTD). Percentages are computed over the total speech duration.

It has to be observed that the main host (Jacques Chancel or Laurent Ruquier) is always among two most prominent speakers, which enables him to orchestrate the talk with savvy transitions as stated in the previous section. It is also interesting to study his speech activity. Table 1.3 highlights the fact that he is the conductor of the show since his speech load is characterised by a great number of turns. 39% of the “*Le Grand Échiquier*” speech turns are from Jacques Chancel while 29% of those of “*On n’a pas tout dit*” are from Laurent

Ruquier. Besides, these speech turns are in general slightly shorter than the average over all the participants shown in Table 1.2 with 2.9 against 3.1 seconds for “*Le Grand Échiquier*” and 2.0 against 2.2 seconds for “*On n’a pas tout dit*”. This difference may look slim, it is however of importance since it accentuates the distinct roles of hosts and guests. The guest discusses his/her life and achievements while the host introduces and contextualises it, as detailed in [13, 34].

1.2 Specifying the task of talk show structuring

The previous section showed a number of features common to talk shows despite the great variability of this type of programs. These characteristics can be refined with a view of target applications. Therefore, it is crucial to assess specific examples of uses to circumscribe the properties of a good structuring process.

1.2.1 Use cases

In a context of preservation, conservation and dissemination of TV recordings, archiving industries have a strong need for indexing tools to be used in a variety of applications. The most prominent use cases include:

- The selection and retrieval of excerpts to be presented on a website (possibly for marketing or commercial purposes) which will be referred to as *UC1*: Use Case 1.
- The development of software to help archivists with the audiovisual indexing of programs (*UC2*: Use Case 2), which, to a large extent, is done manually.
- Complex cross-checking between various programs for intra-document or inter-document navigation (*UC3*: Use Case 3). This could, for instance, allow one to retrieve all the appearances of a given artist, the various versions of a particular song, etc.
- The timecoding of noticeable audiovisual events (*UC4*: Use Case 4).

The latter is quite popular. Many applications require the retrieval of audiovisual events given a corresponding description. Indeed, for archiving companies, most of the time a short outline or a selection of tags is the only available description of the programs stored. Thus, a very important chunk of the audiovisual archives lacks temporal information in the descriptions

provided. A tool performing a temporal alignment of events of interest to video content is therefore one that would be very valuable.

In all the previous use cases, the problem of structuring actually boils down to the extraction of temporal indexes for structural elements defined beforehand. The definition of these various use cases reveals what invariant aspects of the talk show have to be emphasised in order to proceed to the automatic structuring of this kind of programs.

1.2.2 Generic talk show components, the importance of the speakers

In light of the semiotic studies and the use cases mentioned earlier, we propose a set of *structural units* or *components* that characterise the vast majority of talk shows. These units can be considered as general “talk show structural invariances” that stand as particularly relevant for the implementation of use cases of interest. They are specified hereafter and their relative importance is discussed.

Content elements alternate over the duration of a show (between the opening and the final credits). They are organised into three generic entities (as suggested by Section 1.1) talk, performance and inserts.

- The *talk* component refers to every part where talk show participants (hosts and guests) are in an act of conversation. It is the skeleton of the talk show, linking together all constitutive elements.
- *Performance* refers to every live action that is not conversation, especially artistic actions. It includes musical performances, circus acts, but also comedy monologues or poetry recitation (that are not part of the talk component), etc.
- *Inserts* gather every sequence that is not shot inside the studio. They can be archives, reports, still images, etc.

It is worth noting that these three components can overlap or collide. Indeed, performances sometimes start before the host has finished announcing it; a guest may be asked to comment the insert being shown on the screen, etc.

Punctuation elements of diverse natures may also be defined. These markers are used to link together the various content units and make their succession smoother. They can take

a wide number of forms. Events such as *applause* and *laughter* are natural delimiters. From the point of view of the production, *jingles* and *cutaway shots* play the same role as they indicate a shift from one content part of the show to another. *Commercials* are also a clear punctuation element as they can similarly be seen as a delimiter or separator. They are generally introduced by the host beforehand. The host himself being the conductor of the show sets numerous punctuation elements as asserted in the previous section by the semiotic studies [13, 34] and the comparison of talk shows “*Le Grand Échiquier*” and “*On n’a pas tout dit*”. He announces performances or inserts, switches subjects in the conversation, introduces new guests, etc. The punctuation elements set by the host are the primary demarcation markers, since they delimit high-level “semantic” sections of the talk show. The rest of the structural elements belonging to the punctuation category generally only emphasise and put into perspective the information provided by the speakers and, most of the time, the host.

The last generic component of importance that can be defined here is the *location* of the action shown to the viewer which may happen either *outside* or *inside* the studio. Outside refers to the broadcast elements shot at the exterior of the TV studio such as excerpts, duplexes, reports, etc. while inside can take several values such as *set* or *stage*. There is generally a strong overlap between the elements *content* \rightarrow *insert* and *location* \rightarrow *outside*.

It is important to note, that while they explicitly appear only in the talk entity, speakers turn out to be crucial for the actual identification of most of the previous talk show components. This is particularly true for the host whose role as a conductor has, once more, to be emphasised. Practically, the speakers' interventions provide high-level indications on the occurrence of the various structural units. Performances, inserts and commercials are, as a matter of fact, unavoidably introduced and identified by the host, as prescribed by the basic editing TV rules. Thus, the viewer is given information about the type of content (song, circus act, film excerpt, etc.) and, if relevant to the situation, the performer, title of the piece, composer, musicians, author, director, etc.

1.2.3 Link with automatically extracted descriptors

Having defined a taxonomy for the generic structural units of the talk show programs, the focus is now put on how to retrieve them automatically. To this end, several detectors using audio and/or visual cues extracted from the video can be implemented. Table 1.4 gives a

hierarchical classification of the various structural units previously specified as well as known technologies that can be used to detect them.

	structural elements		detection method
<i>CONTENT</i>	<i>TALK</i>	speakers	speaker diarisation
		<i>MUSICAL</i>	music detector
	<i>PERFORMANCE</i>	<i>NON-MUSICAL</i>	speaker diarisation
			laughter detector
		applause detector	
<i>INSERTS</i>	photo	still image detector	
	film-report	environmental sound detector change in colour distribution black stripes detection	
<i>PUNCTUATION</i>	<i>APPLAUSE</i>		applause detector
	<i>LAUGHTER</i>		laughter detector
	<i>CUTAWAY SHOT</i>		high similarity shot detector
	<i>JINGLES</i>		music detector high similarity shot detector
	<i>COMMERCIALS</i>		color coherency detector monochrome frame detector shot boundary detector
<i>LOCATION</i>	<i>INSIDE</i>	<i>STAGE</i>	stage detector
		<i>SET</i>	set detector
	<i>OUTSIDE</i>		color coherency detector environmental sound detector

Table 1.4: Link between generic structural elements and their automatic detection.

Some of the structural elements have been subdivided. For instance, performance leads to two categories: *musical*, where live music is produced, and *non-musical* in the case of monologues, theatre or circus acts, etc. This specific arrangement has been proposed to put in relation generic components and corresponding automatic descriptors. As a matter of fact structural elements defined in this section often correspond directly to the decisions output by specific detectors that can be automatically run on the audio and/or the video signals. Dedicated classifiers can be exploited to detect events such as “applause”, “laughter” [25, 51] or “music” [38, 40] using only the audio modality.

In other cases, such as “non-musical performances” or “commercials”, the structural units can be obtained using reasoning or inference techniques based on the outputs of the previously defined detectors. These outputs then serve as “mid-level descriptors”, on the basis of which the occurrence of the higher-level structural elements is inferred. For instance, the work presented in [4] proposes a method for detecting the presence of commercials through dynamic Bayesian networks [32]. Similarly, generic inference methods are shown in [50] to detect “semantic” events in sport videos. A similar approach can be adopted for the detection of non-musical performances such as theatre acts, poetry reading, etc. where speech occurs. In a preliminary study, an inference system was built upon the combined outputs obtained from

the applause and laughter detectors and speaker diarisation and led to very convincing results for their detection. Thus, while there is only one talk related structural element, one should keep in mind that the speakers' interventions convey a large proportion of the subjective information.

These descriptors turn out to be particularly efficient when dealing with talk show videos. For instance, in [40], a F-measure of 96.5% is obtained for speech/music discrimination on radiophonic streams. Obviously, while a radiophonic stream is not exactly the same content as a TV talk show, common characteristics are shared that ensure a pretty good reproducibility of the results. Commercials can also be extremely well detected, as shown in [4] with a F-measure of 91% for 10 hours of video from French TV channels. In [25], the authors propose new features for the classification of impulse responses and obtain great results at detecting applause. In this case, sports videos are used, however, similar results would be expected with talk show videos. Finally, great performances can also be reached for speaker diarisation methods on talk show content as detailed in 1.3.3.

1.3 Review of key relevant technologies

Any effort in segmenting, indexing or classifying an audiovisual document can be seen as a step in the process of automatic structuring as emphasised in state of the art reviews [12] and [17]. Therefore, a huge variety of potentially useful technologies exist, with varying levels of “semantic” value. Some problems, such as shot boundary detection or music/speech segmentation, have become cornerstones of audiovisual analysis. Besides, methods can follow various learning schemes, ranging from supervised to unsupervised learning.

1.3.1 The use of multimodal approaches

In [45], the accent is put on the importance of the use of multimodal approaches. The joint exploitation of audio, video or even text modalities has shown great results in the resolution of automatic structuring problems. Since a gap is observed between the low level of interpretation provided by basic detectors and the human-level of interpretation of the content being analysed (usually referred to as the “semantic” gap), a combination of different low-level detectors is expected to lead to the resolution of more complex problems.

The TRECVID¹ benchmarking activity has become since 2001 a reference in the field of video information retrieval by providing large test collections and uniform scoring procedures for organisations interested in comparing their results. In this program, the detection of key audiovisual events is performed through the joint analysis of audio and video cues. At the same time, a lot of multimodal approaches for automatic structuring have been proposed as for instance [16, 27] for tennis broadcasts or [35] for Formula 1 races.

1.3.2 Segmentation methods

The canonical segmentation problem in the field of audiovisual content analysis is the shot boundary detection problem [44]. Indeed, the shot is very often considered as the fundamental structural unit for audiovisual documents. In [23, 56], the authors put forward the limitations of the use of the shot as a structuring element, owing to the lack of “semantic” value, and suggest to consider the “audiovisual scene” or “story unit” as an alternative. Story units are nevertheless gathered through the clustering of video shots, hence requiring a good shot boundary detection.

In the audio community, *audio diarisation* is a popular segmentation problem [39]. Different tasks are related to this general problem that are addressed either in a supervised or non-supervised fashion. For instance, speech/non-speech discrimination (where non-speech generally refers to music and/or environmental sounds) is achieved using supervised approaches, often based on Gaussian Mixture Models (GMM) [41] or Support Vector Machines (SVM) [40], while tasks such as speaker turn segmentation are tackled in a non-supervised fashion (see for example [15, 21, 24]). The previous are often used as pre-processing blocks for the task of speaker diarisation that will be addressed specifically in the following section.

1.3.3 Speaker diarisation and identification

Having highlighted the crucial role of the speakers for structuring talk show programs, a deeper look at the task of speaker information retrieval is proposed. Indeed, they are the fundamental description unit for building a coherent structure.

Speaker diarisation is the process of partitioning an input audio stream into homogeneous segments according to the speaker identity. Speaker diarisation systems are widely based on a GMM-Hidden Markov Model (GMM-HMM) architecture where two alternative approaches

¹TREC Video Retrieval Evaluation - <http://trecvid.nist.gov/>

have been proposed: the bottom-up [53] and the top-down [9]. This task is usually exclusively audio-oriented as stressed out in [47]. The descriptors that are traditionally used in these systems are Mel Frequency Cepstrum Coefficients (MFCC) [37] along with their first and second derivatives as they represent the short-term power spectrum of a sound in a homomorphic mapping. However, authors have recently proposed to include additional visual cues such as colour histograms [10], grey-scale difference images [48] or visual activity features [19]. Indeed, the visual information is known to have a much steadier behaviour than the audio and thus to allow better speaker model generation.

An alternative approach was proposed in [49], introducing a novel way of identifying speakers in a talk show program. For this task, the joint use of audio-visual features, namely MFCC, dominant colors [30] and motion features [30] was investigated. A semi-automatic method was chosen in order to lead to the actual identification of the speakers. Indeed, while speaker diarisation methods lead to good results, the problem of assigning a speaker to each cluster still needs to be addressed with the inconvenience that some speakers may not be found among the clusters due to the unsupervised nature of these approaches. Therefore, the scheme proposed is comparable to methods known as relevance feedback ones.

The originality lies in the use of a kernel-based method for a task generally assigned to generative ones. Once a segment of 5 to 15 seconds is given by the user (an archivist in the scenario proposed) for each speaker, a SVM classifier is run with several sets of audio and visual features. On top of classical MFCC features, visual speaker oriented features were developed. Based on the assumption that most of the time the person shown on screen is the one speaking, several video features were introduced, namely the average dominant color of the clothing or motion features for points of interest localised in the region below the face (see Figure 1.2).

Indeed, during a talk show program participants are not expected to change outfit. Besides, some of them show a very distinctive body language while talking, which is what the motion features are supposed to account for. The bounding boxes are obtained using a face detection algorithm [11] and a set of heuristic rules for the location of the costume [26]. The experiment was run on one talk show of the corpus "*Le Grand Échiquier*" and a 100-fold cross-validation was performed on the various candidate speech segments to assess the validity of the approach. Results show that the use of multimodal features leads to an encouraging 8% increase of the performance compared to an audio only system using MFCC features. Indeed,



Figure 1.2: Face and costume bounding boxes — Creative Commons license   (see [1] for details).

the error rate shifts from 35.4% with MFCC only to 27.4% for the combination of MFCC, average dominant color and motion features.

1.3.4 Concept detection systems

Audiovisual structuring can gain from using semantically richer descriptors. Some automatic systems, either supervised or unsupervised, are built to detect audiovisual *concepts*. For instance, a number of works prospected the field of key event recognition in sports videos. For instance, studies on football match videos addressed this type of tasks using Hidden Markov Model (HMM) structures, exploiting either the video only [3], the audio only [5] or both modalities [54]. The audiovisual key events to be retrieved are here typically the ones that would be included in a summary of the match: goals, penalties, faults, etc. In TRECVID, the task described as “high-level feature extraction” is also an instance of concept detection. In this case, participants are asked to automatically retrieve scenes such as *Sports*, *Desert*, *US flag*, *Car*, etc. from a database of videos. A key issue with this task is the variability in concept interpretation making the development of efficient detectors very challenging. Another popular concept detection task is person detection. In [42], the authors retrieve the

appearances of each person present in a TV sitcom using weak supervision from automatically aligned subtitles and script text with very encouraging results. Action recognition [18] is also a very active topic. It consists in temporally annotating a program with human actions in an automatic fashion. Action classes detected are in general of the following type: *opening door*, *drinking*, *sitting down*, etc.

Finally, audio concepts can also be detected in a similar manner. For instance, in [38], the authors propose discriminative methods based on SVM for the detection of vocal parts in music signals.

Thus, it is clear that there is substantial background that can be exploited for the implementation of the detectors to be used in extracting the talk show components described earlier.

1.4 Structuring method based on speakers' interventions

After defining generic structural elements and pointing to methods to retrieve them automatically, the focus is now put on their combination for the resolution of a specific use case.

1.4.1 Creation of the structure

It has been shown that generic talk show components related to noticeable audiovisual events can be retrieved automatically using appropriate concept detectors. It is worth noting that these detectors become very reliable when their local outputs are integrated over longer temporal windows for decision making. The choice of the lengths of these temporal decision windows typically depends on the type of event to detect. For example, while 1 to 2-second length windows should be used for taking decision about the occurrence of "laughter", up to 1-minute length windows could be used for musical or non-musical performances.

The critical role of the speakers has also been largely commented since the structural units proposed are mostly based on speaker related descriptors. It has been demonstrated that the talk component convey the majority of the "semantic" knowledge. Though automatic speech recognition systems would deliver precious information on the structure, their use here was discarded. Indeed, these systems would lead to language dependent and therefore non generic approaches. The focus is thus put on the study of the physical properties of the speech that may allow high level reasoning such as the speech repartition between speakers and the

prosody. The structure proposed here results from the combination of the various generic components defined in Table 1.4 (talk, musical performance, applause, etc.). This list can be refined if some structural units are obviously useless for specific talk-show instances, as could have been jingles for a talk show like “*Le Grand Échiquier*” for which this procedure was not common at the time the program was released. The set of generic components defined here is thus expected to solve a wide variety of use cases on top of those presented earlier in 1.2.1. To actually evaluate our structuring scheme to correctly capture the organisation of a talk show document, we chose to apply it to the specific task of timecoding the noticeable audiovisual events selected by the archivists. This refers to the use case *UC4* presented in 1.2.1 where archivists' notes have to be temporally aligned with the corresponding video. The structural elements defined are expected to facilitate the navigation through the various events occurring during the course of the talk show and therefore simplify the task.

1.4.2 Evaluating the structure

We propose an evaluation of our structuring scheme by applying it to the use case *UC4* (described in Section 1.2). To this end, we considered the corpus “*Le Grand Échiquier*” composed of talk show videos and corresponding audiovisual notes. Indeed, this corpus displays a great variety of audiovisual events of interest and has thus been the object of several studies (see for example [2, 10, 24, 49]). Sentences referring to audiovisual events were selected in the archivist's notes.

The task consisted in asking users to retrieve as fast as possible the audiovisual events corresponding to the chosen sentences. Each subject had to retrieve one half of these events using structural elements provided beforehand and the other half without. Reference sentences were manually selected to reflect a large range of aspects of the audiovisual programs. Some were for instance strongly linked to visual cues while some others were more complex in the sense of their high-level human interpretation. Such a choice was made to efficiently test the relevance of the approach proposed. A manual annotation locating the matching events was performed and is hereafter referred to as groundtruth.

Protocol

Users had to retrieve 16 audiovisual events from each of 4 shows (4 excerpts per show) within a time limit of 8 minutes for each event, thus, making it 2 shows with structural elements and

2 shows without. 20 subjects took the test and the group was divided in two parts: A and B. Users belonging to the group A had to retrieve events with the help of structural elements for the shows 1 and 3 and without them for the shows 2 and 4. It was the opposite for the subjects from the group B.

In order to put the focus on the evaluation of the relevance of the chosen structure (rather than on the performance of low-level detectors) manually annotated structure components were actually used in this test. Indeed, using automatically obtained results could have caused interpretation issues since two major effects would then have been observed: the correctness of the detected elements and their actual usefulness for the retrieval task. Table 1.5 presents the sentences selected for the task.

show / guest	<i>Orchestra of the Opéra national de Paris (1982)</i>	<i>Jean-Pierre Rampal (1985)</i>	<i>Michel Sardou (1982)</i>	<i>Michel Berger (1985)</i>
length	2 hours 58 min 37 sec	2 hours 40 min 3 sec	2 hours 51 min 0 sec	2 hours 17 min 10 sec
excerpt 1	Ella Fitzgerald sings "smoke gets in your eyes" (insert)	Jean-Pierre Rampal plays the flute solo in Gluck's "orpee"	seal taming act by Robby Gasser	Patrick Vigier presents a guitar with microprocessor
excerpt 2	the orchestra directed by Alain Lombard plays "alleluia"	excerpts of an American program showing Rampal with the Muppets	Mireille Darc reads the poem "Colloque sentimental" by Verlaine	Michel Berger sings "Y'a pas de honte"
excerpt 3	Jacques Chancel talks with the union representative of the orchestra	the famous flute player presents his family: Françoise his wife, his daughter and his son	Jean-Jacques Debout criticises Jack Lang	24 hours of violence: broadcast news from A2
excerpt 4	mechanic dismantling the piano in the studio	Alexandre Lagoya talks about Django Reinhardt	Michel Audiard post-syncing a scene of "Quest for Fire" by Jean-Jacques Annaud	Daniel Balavoine talks about his 1 st album "mur de berlin"

Table 1.5: Excerpts to retrieve in four shows of "*Le Grand Échiquier*".

Ideally, a dedicated navigation tool would have been designed for the task. However, since the evaluation was focused on measuring the contribution of various structural elements (regardless of presentation considerations), an existing software was chosen instead. Thus, the users were asked to retrieve the excerpts using the ELAN² program, a professional tool for the creation and the visualisation of audiovisual annotations (see [43]). This has actually raised ergonomic issues that were ignored since the default layout was considered as satisfactory. Whether they were using the structural components or not, the subjects were always allowed to use the basic controls for video navigation (slider, play/pause, shift of one second, etc.). To get accustomed with the tool, the users were given a 15-minute training with and without the structure elements on two other shows before proceeding to the actual test.

²EUUDICO Linguistic Annotator - <http://www.lat-mpi.eu/tools/elan/>

Once the training phase had finished, the users were asked to proceed to the actual testing. They had to retrieve in a row all four excerpts of each program with pauses allowed only when switching shows. The objective was to limit the “learning” factor. Indeed, users would otherwise increase their score over time since they would jointly learn how to use the ELAN software and the organisation of the show itself.

content			punctuation				location			
performance		insert	talk	applause	laughter	shot 1	shot 2	studio		exterior
musical	non-musical		speakers					stage	set	

Table 1.6: Structural elements available to the users during the test.

When an excerpt was identified, the users had to provide a timecode for its beginning and another for its end before being shown the next one. To be considered as correctly attributed, a video sequence had to overlap with at least 70% of the groundtruth. By this mean, we ensured the correct identification of the events without putting the focus on their slicing. In case of misdetection, the user was given by default the maximum retrieval time allowed for a single excerpt i.e. 8 minutes.

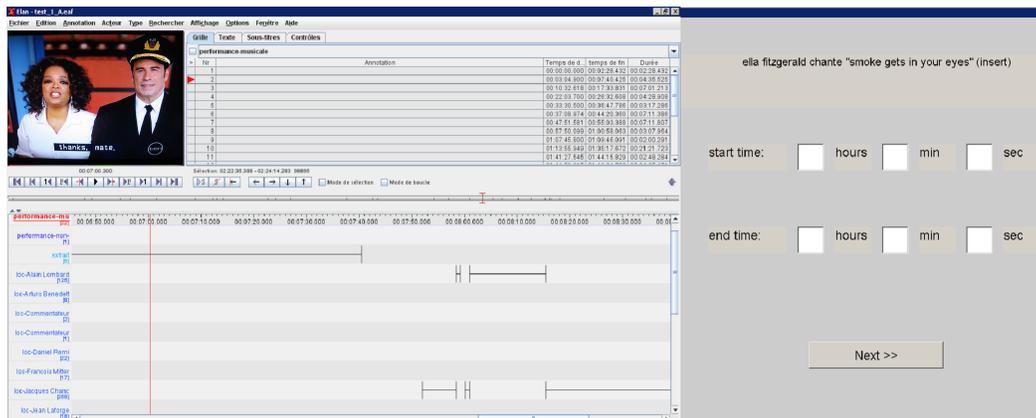


Figure 1.3: Illustrations of the ELAN software and of the interface collecting the timecodes — Creative Commons license (see [1] for details).

Results and discussion

The maximum time allowed for retrieving a given excerpt being 8 minutes, the time limit for the evaluation was theoretically of 2 hours and 8 minutes (without counting the training phase and the pauses between the shows). However, the average time spent was slightly more than

one hour (1 hour 1 minute 38 seconds with a standard deviation of 6 minutes 52 seconds). Group A performed on average in 1 hour 4 minutes 7 seconds with a standard deviation of 6 minutes 4 seconds and group B in 59 minutes 9 seconds with a standard deviation of 7 minutes. The results in Table 1.7 highlight the fact that on average it takes less time to find a given excerpt with structuring elements than without: 3 minutes 31 seconds against 4 minutes 11 seconds.

excerpt	with structuring elements				without structuring elements			
	average retrieval time (sec)	standard deviation (sec)	% wrong	% over time	average retrieval time (sec)	standard deviation (sec)	% wrong	% over time
1	170.7	62.2	0	0	229.1	122.8	0	10
2	438.8	73.7	0	60	442.4	102.1	10	70
3	380.6	131.8	20	30	357.2	108.9	0	30
4	209.1	115.0	0	10	130.3	41.3	0	0
5	392.7	100.9	0	40	387.9	76.4	0	30
6	107.2	39.9	0	0	90.1	34.5	0	0
7	300.2	112.7	0	20	326.5	119.5	10	10
8	115.3	33.3	0	0	327.6	139.5	10	30
9	107.8	87.9	0	0	85.2	22.9	0	0
10	93.5	38.2	0	0	197.7	102.9	0	0
11	164.9	39.6	0	0	241.9	84.1	0	0
12	193.2	44.5	0	0	249.4	105.9	0	0
13	102.3	35.0	0	0	201.9	57.8	0	0
14	319.3	128.0	10	20	236.5	113.6	10	0
15	82.7	36.3	0	0	123.5	63.5	0	0
16	195.8	119.3	20	0	395.4	129.0	20	40
average	210.9	118.4	2.5	11.3	251.4	112.3	3.8	14.8

Table 1.7: Structural elements available to the users during the test. **% wrong** stands for the percentage of users who entered an incorrect audiovisual segment while **% over time** indicates the percentage of users who did not locate the event within the allowed time.

Surprisingly, for several excerpts, such as 3,4,6,9 and 14, it actually took more time in average to retrieve the excerpts with the structural elements than without. However, the presence of a generally large standard deviation has to be noted. For some excerpts, such as 4 (“mechanic dismantling the piano in the studio”), structural elements did not really help users since it was not clearly linked to any of the proposed generic structural components. Subjects tended to spend some time to figure out which components could be useful (they wondered if it was an insert, a performance, if it happened on stage, etc.) instead of directly using the usual video navigation tools. Also, for excerpts easily identifiable visually, such as 6 (“the Muppets”) or 9 (“seal taming act”), it actually took more time to think about what field the event actually belonged to than to just scroll along the show with the slider. This is another evidence that users had sometimes the tendency to forget to use basic navigation tools when provided with the structural components. This inclination seems however rectifiable.

For some other excerpts such as excerpt 3 for instance, many users relying on the structural

elements tended to proceed hastily, got misled, and assigned the role of “union representative” to the wrong person. This is highlighted by the particularly high number of wrong detections for this excerpt. Thus, users with structural elements showed bad answers at a higher rate than users without. Since errors were penalised with the maximum allowed time 8 minutes, greater average retrieval times were observed. It is a clear sign that users leaned towards trusting exceedingly the structural elements. They seemed to take a less active part in the search of the events. Feeling that they had the information at hand, they were indeed more eager to answer the question within the time limit.

The 16 excerpts can be classified in the four following categories that were defined as generic components earlier: talk, inserts, musical performance and non-musical performance. Due to their nature, some excerpts can belong to more than one single category. For instance, “Michel Audiard post-syncing a scene of “Quest for Fire” by Jean-Jacques Annaud” belongs to non-musical performance as much as insert (since the film is shown on screen). Thus, it is possible to compute the average retrieval time according to the type of the excerpts (Table 1.8).

excerpt	with structuring elements				without structuring elements			
	average retrieval time (sec)	standard deviation (sec)	% wrong	% over time	average retrieval time (sec)	standard deviation (sec)	% wrong	% over time
talk	193.2	101.2	3.8	6.3	287.2	74.2	5	13.8
insert	152.6	55.0	0	2	164.5	70.3	0	2
musical perf.	285.7	142.4	2	24	277.2	140.1	4	22
non-musical perf.	165.2	62.7	0	3.3	192.5	59.7	0	0

Table 1.8: Retrieval time with and without structuring elements according to the type of excerpt.

Except for the musical performances where users without the structural elements at hand performed slightly better than the others, the results confirm the usefulness of the structuring for this application. Users with structure information clearly outperformed the others in the case of talk excerpts. This is furthermore asserted by the low percentages for answers exceeding the time limit, indicating that the information is, in this case, much easier to retrieve.

Following the experiment, users were asked to provide a feedback on the way they handled it. They were in particular asked what structural elements they used. The results of this questionnaire are summarised in Table 1.9.

As expected, content elements proved to be very useful to retrieve audiovisual events. This is due to the nature of these events. Indeed, since extracted from the archivists' notes, they pointed to an action of actual interest. There were for instance no events such as “X laughs at Y's joke” or “W applauds Z's performance”, etc. The decrease in usefulness for non-musical performance can be explained by the lack of non-musical performances for users belonging to the group B (since they did not have the structural elements for the show 3 where two non-musical performances happened). It is interesting to observe that while it took slightly more time to find musical performances with the structural elements than without, it was still, on average, considered as very useful by the subjects.

element	musical perf.	non-musical perf.	insert	speakers	applause	laughter	cutaway shot 1	cutaway shot 2	stage	set	exterior
usefulness	3.9	2.9	3.6	3.9	1.5	1.4	1.4	1.4	1.9	1.9	1.8

Table 1.9: Usefulness of the various structural components according to the users (4: very useful, 3: quite useful, 2: not very useful, 1: useless).

It has to be noted also that the wide standard deviation on all results indicates that the ergonomics of the ELAN program plays a great role as well as the capacity of the users to browse at best the annotations available at hand. It is worth noting that it is difficult to single out the only merit of the structure proposed. Several aspects are evaluated at the same time: the ability of the subject to use the ELAN software efficiently, his/her retrieval strategy as well as the “chance” factor. All the users were rather familiar with computers. However large differences occurred in the way they retrieved excerpts. The fastest users tended to scroll through the whole show independantly of whether they had the structure available or not to get a rough idea of the show. It turned out that the time spent in doing so allowed them to achieve a much faster retrieval since they had already figured out the form of the show. Indeed, the “learning” aspect is very important and hard to isolate. While looking for a given excerpt, users accumulate information that can allow them sometimes to retrieve future events (that are unknown to them at this point) much faster. This is clearly related to the user's capacity of memorising and identifying what he/she sees on screen. It is also related to the “chance” factor that is as well impossible to eliminate. Besides, despite the training phase, users also learned how to handle the software the most efficiently as possible along the experiment.

One has to keep in mind that this evaluation targetted specifically the use case *UC4*. The

appeal of the generic structuring scheme proposed lies in the possibility to use it for other tasks, such as the retrieval of excerpts to be presented on a website (*UC1*) or the cross-checking of various shows, for instance to retrieve the appearances of a given artist that is known to be present in the considered programs (*UC3*). In these circumstances as well the structure proposed would be expected to simplify and speed up the execution of tasks that are usually fully manual. This evaluation is also a good starting point for the reflection on the construction of an efficient navigation tool as proposed in the use case *UC2*.

The results of this user-oriented experiment therefore prove the usefulness of the generic structural elements proposed in Section 1.2. Users retrieved audiovisual events more easily with the structural elements than without. It was particularly obvious for talk-oriented excerpts. The difference was not as marked for other types of events such as musical performances. However, the browsing through the video shows with the ELAN software seems to be improvable, indicating that ergonomics problems have to be carefully addressed.

Conclusion

In this chapter a TV talk show structuring scheme has been proposed that does not depend on automatic speech recognition technologies. Inspired by semiotic works, a number of talk show common characteristics have been drawn that has led to the identification of key structural components. Use cases of interest have been proposed, on the basis of which the most relevant of those components have been cast out to form a generic structuring scheme. The link between the components defined and corresponding automatic detectors has been established that emphasised the importance of speakers' related processing modules (owing to the fact that the inference of most structural components depend on their outputs). Finally, a user evaluation has been proposed on a specific use case (of particular interest to archivists' activity) that confirmed the usefulness of the structuring scheme. This evaluation has pointed out the importance of presentation aspects for a proper exploitation of the structure. Efforts should be dedicated in this direction towards the design of efficient interfaces as part of the effort of implementation of particular use cases.

References

- [1] flickr images - <http://www.flickr.com/photos/65482984@N04/galleries/72157627130487057/>.
- [2] José Anibal Arias, Julien Piquier, and Régine André-Obrecht. Evaluation of classification techniques for audio indexing. In *European Signal Processing Conference*, 2005.
- [3] Jürgen Assfalg, Alberto del Bimbo, Walter Nunziati, and Pietro Pala. Soccer highlights detection and recognition using hmms. In *International Conference on Multimedia and Expo*, 2002.
- [4] Siwar Baghdadi, Guillaume Gravier, Claire-Hélène Demarty, and Patrick Gros. Structure learning in bayesian network based video indexing. In *International Conference on Multimedia and Expo*, 2008.
- [5] Mark Baillie and Joemon M. Jose. An audio-based sports video segmentation and event detection algorithm. In *Conference on Computer Vision and Pattern Recognition Workshop*, 2004.
- [6] Meriem Bendris, Delphine Charlet, and Gérard Chollet. Talking faces indexing in tv-content. In *Content-Based Multimedia Indexing*, 2010.
- [7] Pierre Bourdieu. *Sur la télévision*. Raisons d'agir, 1996.
- [8] Jérôme Bourdon. Propositions pour une semiologie des genres audiovisuels. *Quaderni*, 4:19–36, 1988.
- [9] Simon Bozonnet, Nicholas Evans, and Corinne Fredouille. The lia-eurecom rt'09 speaker diarization system : enhancements in speaker modelling and cluster purification. In *International Conference on Acoustics, Speech, and Signal Processing*, 2010.
- [10] Simon Bozonnet, Félicien Vallet, Nick Evans, Slim Essid, Gael Richard, and Jean Carriev. A multimodal approach to initialisation for top-down speaker diarization of television shows. In *European Signal Processing Conference*, 2010.
- [11] Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, 2008.
- [12] Roberto Brunelli, Ornella Mich, and Carla-Maria Modena. A survey on the automatic indexing of video data. *Journal of Visual Communication and Image Representation*, 10:78 – 112, 1999.
- [13] Sabine Chalvon-Demersay and Dominique Pasquier. Le langage des variétés. *Terrain*, 15:29–40, 1990.
- [14] Patrick Charaudeau. Les conditions d'une typologie des genres télévisuels d'information. *Réseaux*, 15:79–101, 1997.
- [15] Scott S. Chen and P.S. Gopalakrishnan. Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *DARPA Broadcast News Transcription and Understanding Workshop*, 1998.
- [16] Manolis Delakis, Guillaume Gravier, and Patrick Gros. Audiovisual integration with segment models for tennis video parsing. *Computer Vision and Image Understanding*, 111:142 – 154, 2008.
- [17] Nevenka Dimitrova, Hong-Jiang Zhang, Behzad Shahraray, Ibrahim Sezan, Thomas Huang, and Avidesh Zakhor. Applications of video-content analysis and retrieval. *Multimedia, IEEE*, 9:42–55, 2002.
- [18] Olivier Duchenne, Ivan Laptev, Josef Sivic, Francis Bach, and Jean Ponce. Automatic annotation of human actions in video. In *International Conference on Computer Vision*, 2010.
- [19] Gerald Friedland, Hayley Hung, and Chuohao Yeo. Multi-modal speaker diarization of real-world meetings using compressed-domain video features. In *International Conference on Acoustics, Speech and Signal Processing*, 2009.
- [20] Rodolphe Ghiglione and Patrick Charaudeau. *La Parole confisquée. Un genre télévisuel: le talk show*. Dunod, 1997.
- [21] Herbert Gish, Man-Hung Siu, and Robin Rohlicek. Segregation of speakers for speech recognition and speaker identification. In *International Conference on Acoustics, Speech,*

- and *Signal Processing*, 1991.
- [22] Erving Goffman. *Forms of Talk*. University of Pennsylvania Press, 1981.
 - [23] Alan Hanjalic, Reginald Lagendijk, and Jan Biemond. Automated high-level movie segmentation for advanced video-retrieval systems. *IEEE Transactions on Circuits and Systems for Video Technology*, 9:580–588, 1999.
 - [24] Zaïd Harchaoui, Félicien Vallet, Alexandre Lung-Yut-Fong, and Olivier Cappé. A regularized kernel-based approach to unsupervised audio segmentation. In *International Conference on Acoustics, Speech and Signal Processing*, 2009.
 - [25] Cyril Hory and William J. Christmas. Cepstral features for classification of an impulse response with varying sample size dataset. In *European Signal Processing Conference*, 2007.
 - [26] Gaël Jaffré and Philippe Joly. Costume: A new feature for automatic video content indexing. In *International Conference on Adaptivity, Personalization and Fusion of Heterogeneous Information*, 2004.
 - [27] Ewa Kijak, Guillaume Gravier, Lionel Oisel, and Patrick Gros. Audiovisual integration for tennis broadcast structuring. *Multimedia Tools and Applications*, 30:289 – 311, 2006.
 - [28] Marie Lhéroult and Erik Neveu. Quelques dispositifs de talk-shows français (1998-2003). *Réseaux*, 118:201–207, 2003.
 - [29] Guy Lochard. Débats, talk-shows : de la radio filmée? *Communication et langages*, 86:92–100, 1990.
 - [30] B.S. Manjunath, Philippe Salembier, and Thomas Sikora, editors. *Introduction to MPEG-7 - Multimedia Content Description Interface*. Wiley, 2002.
 - [31] Wayne Munson. *All Talk: The Talk Show in Media Culture*. Temple University Press, 1993.
 - [32] Kevin Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD thesis, UC Berkeley, Computer Science Division, 2002.
 - [33] Xavier Naturel and Patrick Gros. Detecting repeats for video structuring. *Multimedia Tools and Applications*, 38:233 – 252, 2008.
 - [34] Hermine Penz. *Language and Control in American TV Talk Shows: An Analysis of Linguistic Strategies*. Gunter Narr, 1996.
 - [35] Milan Petkovic, Vojkan Mihajlovic, Willem Jonker, and S. Djordjevic-Kajan. Multi-modal extraction of highlights from tv formula 1 programs. In *International Conference on Multimedia and Expo*, 2002.
 - [36] Jean-Philippe Poli. An automatic television stream structuring system for television archives holder. *Multimedia systems*, 14:255–275, 2008.
 - [37] Lawrence Rabiner and Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice Hall PTR, 1993.
 - [38] Mathieu Ramona, Gael Richard, and Bertrand David. Vocal detection in music with support vector machines. In *International Conference on Acoustics, Speech and Signal Processing*, 2008.
 - [39] Douglas A. Reynolds and Pedro A. Torres-Carrasquillo. Approaches and applications of audio diarization. In *International Conference on Acoustics, Speech, and Signal Processing*, 2005.
 - [40] Gaël Richard, Mathieu Ramona, and Slim Essid. Combined supervised and unsupervised approaches for automatic segmentation of radiophonic audio streams. In *International Conference on Acoustics, Speech and Signal Processing*, 2007.
 - [41] Eric Scheirer and Malcolm Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *International Conference on Acoustics, Speech, and Signal Processing*, 1997.
 - [42] Josef Sivic, Mark Everingham, and Andrew Zisserman. Who are you?: Learning person specific classifiers from video. In *International Conference on Computer Vision and Pattern Recognition*, 2009.
 - [43] Han Sloetjes and Peter Wittenburg. Annotation by category - elan and iso dcr. In

- International Conference on Language Resources and Evaluation*, 2008.
- [44] Alan Smeaton, Paul Over, and Aiden Doherty. Video shot boundary detection: Seven years of trecvid activity. *Computer Vision and Image Understanding*, pages 1–25, 2009.
 - [45] Cees G.M. Snoek and Marcel Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25, 2005.
 - [46] Bernard Timberg. *Television Talk: A History of TV Talk Show*. University of Texas Press, 2002.
 - [47] Sue E. Tranter and Douglas A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14 (5):1557–1565, 2006.
 - [48] Himanshu Vajaria, Tanmoy Islam, Sudeep Sarkar, Ravi Sankar, and Ranga Kasturi. Audio segmentation and speaker localization in meeting videos. In *International Conference on Pattern Recognition*, 2006.
 - [49] Félicien Vallet, Slim Essid, Jean Carrive, and Gael Richard. Robust visual features for the multimodal identification of unregistered speakers. In *International Conference on Image Processing*, 2010.
 - [50] Fei Wang, Yu-Fei Ma, Hong-Jiang Zhang, and Jin-Tao Li. A generic framework for semantic sports video analysis using dynamic bayesian networks. In *Multimedia Modelling Conference*, 2005.
 - [51] Peter Wilkins, Tomasz Adamek, Daragh Byrne, Gareth J. F. Jones, Hyowon Lee, Gordon Keenan, Kevin McGuinness, Noel E. O'Connor, Alan F. Smeaton, Alia Amin, Zeljko Obrenovic, Rachid Benmokhtar, Eric Galmar, Benoît Huet, Slim Essid, Rémi Landais, Félicien Vallet, Georgios Th. Papadopoulos, Stefanos Vrochidis, Vasileios Mezaris, Ioannis Kompatsiaris, Evaggelos Spyrou, Yannis Avrithis, Roland Morzinger, Peter Schallauer, Werner Bailer, Tomas Piatrik, Krishna Chandramouli, Ebroul Izquierdo, Martin Haller, Lutz Goldmann, Amjad Samour, Andreas Cobet, Thomas Sikora, and Pavel Praks. K-space at trecvid 2007. In *TRECVID 2007 - Text Retrieval Conference TRECVID Workshop*, 2007.
 - [52] Raymond Williams. *Television: Technology and Cultural form*. Fontana, 1974.
 - [53] Chuck Wooters and Marijn Huijbregts. The icsi rt07s speaker diarization system. *Multimodal Technologies for Perception of Humans*, 4625:509 – 519, 2008.
 - [54] Ziyou Xiong, Regunathan Radhakrishnan, and Ajay Divakaran. Generation of sports highlights using motion activities in combination with a common audio feature extraction framework. In *International Conference on Image Processing*, 2003.
 - [55] Min Xu, Ling-Yu Duan, Changsheng Xu, Mohan Kankanhalli, and Qi Tian. Event detection in basketball video using multiple modalities. In *Pacific-Rim Conference on Multimedia*, 2003.
 - [56] Minerva Yeung and Boon-Lock Yeo. Time-constrained clustering for segmentation of video into story units. In *International Conference on Pattern Recognition*, 1996.
 - [57] Dongqing Zhang and Shih-Fu Chang. Event detection in baseball video using superimposed caption recognition. In *ACM Conference on Multimedia*, 2002.
 - [58] Wensheng Zhou, Asha Vellaikal, and C.-C. J. Kuo. Rule based video classification system for basketball video indexing. In *ACM workshops on Multimedia*, 2000.