Chapter 1

SCENE DETERMINATION USING AUDITIVE SEGMENTATION MODELS OF EDITED VIDEO

Silvia Pfeiffer
CSIRO Mathematical and Information Sciences
Locked Bag 17, North Ryde NSW 1670, Australia
Silvia.Pfeiffer@cmis.csiro.au

Uma Srinivasan
CSIRO Mathematical and Information Sciences
Locked Bag 17, North Ryde NSW 1670, Australia
Uma. Srinivasan@cmis.csiro.au

Abstract This chapter describes different approaches that use audio features for determination of scenes in edited video. It focuses on analysing the sound track of videos for extraction of higher-level video structure. We define a scene in a video as a temporal interval which is semantically coherent. The semantic coherence of a scene is often constructed during cinematic editing of a video. An example is the use of music for concatenation of several shots into a scene which describes a lengthy passage of time such as the journey of a character. Some semantic coherence is also inherent to the unedited video material such as the sound ambience at a specific setting, or the change pattern of speakers in a dialogue. Another kind of semantic coherence is constructed from the textual content of the sound track revealing for example the different stories contained in a news broadcast or documentary. This chapter explains the types of scenes that can be constructed via audio cues from a film art perspective. It continues on a discussion of the feasibility of automatic extraction of these scene types and finally presents existing approaches.

Keywords: scene determination, audio content analysis, sound classes, shot clustering, scene types

Introduction

Video structure extraction is essential for effective search, retrieval and browsing of video. The segmentation of a video into its shots and scenes provides a better semantic access to video structure than the pure frame-level access.

Example applications that may make use of the higher-level segmentation of a film into scenes are:

- navigation and browsing applications: shots and scenes may be used to create a table of contents for a film and for direct access.

- search and retrieval applications: annotations such as keywords, full transcripts or meta-data like MPEG-7 may be used to perform a content-based search on films; in this context, the retrieved entity may be a scene giving a semantically richer access to the film than a shot.

- summarisation applications: scenes may be used as semantically richer basic entities in the creation of a film abstract or summary.

In this chapter we explore only auditive techniques for unification of several shots into a coherent scene. The first section describes our meta-model framework that represents video semantics at various levels of abstraction taking automatic analysis results from low level features to high level semantics. In the second section, we adapt the model to the subject of the chapter. We analyse current audio editing practices and cinematic techniques which create perceivable scene structures and thus explain from a production viewpoint how scenes are constructed using audio effects. In the third section, we discuss the feasibility of automatic extraction for the previously identified scene types. This creates the link between the high-level cinematic scene structure and signal analysis. The fourth section presents existing approaches toward automatic scene determination using audio features. We conclude the chapter with a summary.

1. The Meta-model Framework

Most research to date on video content analysis is based on retrieving low level perceptual features in both the audio and visual domain. These features can be useful only if they are represented in the context of higher level semantics that are meaningful to viewers. In order to understand and represent the sophisticated semantics of films, we need a model that represents audio-visual (av) objects, meanings associated with these av

objects and other associated objects, actions, and events depicted within these objects.

In this section, we first describe a meta-model that helps to model video semantics at different levels of abstraction. The meta-model presented here draws from semiotics and film theory [1], and allows users to develop and specify their own semantics while simultaneously exploiting the results of video analysis techniques. The term "film semiotics" in this context describes the analysis of film on a level of audio-visual signs that communicate meaning to the viewers.

The video meta-model represents the spatio-temporal dimensions of a video on the horizontal axis and the semantics on the vertical axis (see Figure 1.1). The third dimension shows the multiple levels of cinematic codification that cover audio-visual features, objects, actions, and events depicted in the images together with semiotic aspects of the meaning of images. We believe that modeling video semantics requires describing a video object at any semantic level at any of these levels of interpretation.

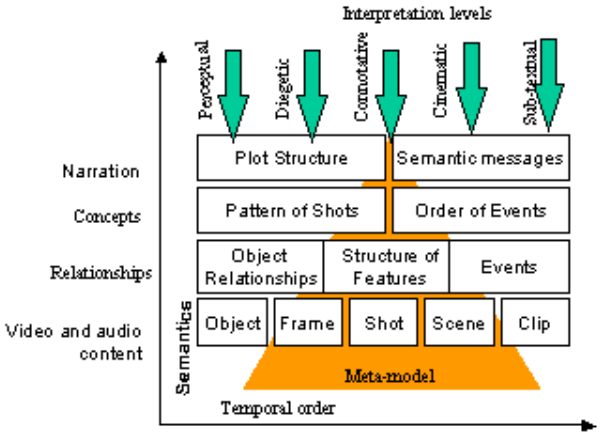

Figure 1.1. Video meta-model.

Using the meta-model framework, video semantics can be incrementally developed by establishing relationships across av objects in any or all of the layers shown in Figure 1.1. The bottom layer shows the standard way in which video content is organised into temporal levels of granularity. The terms used in the diagram are defined as follows:

- **Object:** Represents a single av object within a frame. The object here is not necessarily region-based. It could be feature-based, for example colour, loudness, etc.
- **Frame:** A single image of a video sequence.
- **Shot:** "A series of frames produced by the camera in an uninterrupted operation" ([2], p. 10).
- **Scene:** A scene is a series of consecutive shots constituting a unit from the narrative point of view [3]. This happens when they are shot in the same location or they share some thematic content.
- **Clip (video):** An arbitrary excerpt of a video used for a specific purpose.

The next layer in the semantic dimension helps to model semantics based on different types of relationships in the spatio-temporal domain.

- **Object relationships:** Relationship among objects represented in a frame, shot, or scene. The relationship could be spatial, temporal, visual, aural, or semantic in nature.
- **Structure of features:** Pattern of features that represents a semantic construct. For example, a group of regions connected through some spatial relationships may lead to the shape of an object. (Basically it represents knowledge about the features).
- **Events:** Specific events that occur over a temporal interval. For example, camera pan is a cinematic event; a sudden burst of sound is a perceptual auditory event.

The next level of conceptualization occurs when certain patterns can be perceived over the spatio-temporal space, which is most often the case in narrative forms in films. Bordwell and Thompson [2] define narrative to be a chain of events in cause-effect relationship occurring in time and space. Film theorists realised that many of these cause-effect relationships can be achieved through cinematic effects introduced through powerful film editing techniques. Viewers' experiences are shaped using editing techniques that control shot lengths and relationships between shots.

- **Pattern of shots:** This represents different types of relationships between shots planned during the process of continuity editing. An editor uses four basic ways of arranging patterns of shots to produce a desired effect on the viewer [2]. These arrangements

are based on graphic relationships, rhythmic relationships, spatial relationships and temporal relationships. For example, Adams et al. [4] measure tempo of a movie using temporal relationships on shot lengths and motion characteristics.

- **Order of events:** This represents the order of cinematic and perceptual events arranged in a certain predetermined sequence. For example, a close-up followed by a loud sound could be used to produce a dramatic effect. In the case of narrative structure, the order of events occurring in succession is seen as a kind of surface structure that conceals deeper logic of the narrative story.
- **Plot structure:** This describes a systematic ordering of plot events for narrative progress and development.
- **Semantic messages:** The message refers to the meaningful sequences generated by the process of communicative utterances. Semantic messages deals with the relation of signs and messages produced by the narrative to the larger cultural system, which gives it meaning.

The third dimension in the meta-model shows the different semiotic levels at which a video is interpreted. Based on film semiotics pioneered by the film theorist Christian Metz [1], we identify five levels of cinematic codification that must be represented in the meta-model [5]. These are:

1 **Perceptual level:** This is the level at which visual phenomena become perceptually meaningful, the level at which distinctions are perceived by the viewer. This is the level that is concerned with features such as colour, loudness and texture.

2 **Cinematic level:** This level is concerned with formal film and video editing techniques that are incorporated to produce expressive artifacts. For example, arranging a certain rhythmic pattern of shots to produce a climax, or introducing voice-over to shift the gaze.

3 **Diegetic level:** This refers to the four-dimensional spatio-temporal world posited by a video image or a sequence of video images, including spatio-temporal descriptions of objects, actions, and events that occur within that world.

4 **Connotative level:** This level of video semantics is the level of metaphorical, analogical and associative meanings that the objects and events in a video may have. An example of connotative significance is the use of facial expression to denote some emotion.

5 **Sub-textual level:** This is the level of more specialized, hidden and suppressed meanings of symbols and signifiers that are special to cultural and social groups.

Increasingly the generation of cinematic and perceptual level descriptions is being automated in current video analysis techniques. Subtextual and connotative descriptions must still be created manually. The diegetic level represents an interface between what may be detected automatically and what must be defined manually [6].

The main idea of this meta-model framework is to allow users to develop their own application models, based on their semantic notion, by specifying objects and relationships of interest at any level of granularity. A given application may use any subset of the model. In [7], we have shown how this meta-model is used to develop an application for the sports domain.

Our focus in this chapter is to show how sound contributes to video semantics represented in the meta-model. It is well understood by film makers that sounds complete the film and video experience because they bring reality to the illusion of image. Sounds in films induce the audience to respond at any or all the interpretation levels shown in the meta-model.

The tripartite division of sound track into speech (or dialogue), music and noise drawn from the vocabulary of film making practice is based only on the perceptual level and is hardly adequate to the analysis of the audio-visual logic of the represented world of the film [8].

At the diegetic level, we need to hear the sounds that match the images on the screen: For example an actor knocking on the door with no sound is not really a knock at all.

At the connotative level, a sound can subtly affect how we respond to a scene emotionally. A night scene of a couple in the woods can be entirely different depending on the sound we hear. The sound of a howling wolf as against the sound of gentle breeze gives a completely different sense of drama to the same visual scene.

As Bordwell and Thompson [2] point out, we must learn to 'listen' to films, as sound can achieve very strong visual effects and yet remain quite unnoticeable. Sound can actively shape how we perceive and interpret the image. For example, music in films contributes to the interpretation at all the levels shown in the meta-model.

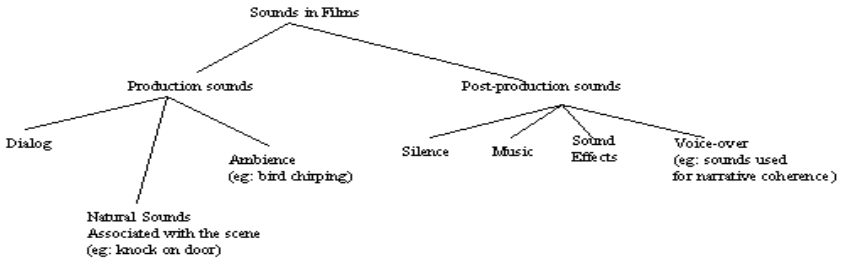

Figure 1.2. Sound Categories.

2. Audio Editing Practices for Scenes

We classify sounds in films into two broad categories: production sounds and post-production sounds (see Figure 1.2). The sound recorded on the set is called production sound and should be recorded and cut with the same care as the visuals. Sound created later and used to fill in gaps or add to existing production sound is called post-production sound. Both these categories contribute to the development of coherent scenes in films.

How does a scene get recorded? Bordwell and Thompson [2] describe the production of a scene (the recording itself) in three steps:

1 The director films a master shot, which records the entire action and dialogue of a scene.

2 Portions of the scene are restaged and shot in closer views or from different angles.

3 The script supervisor ensures continuity details on the image track.

This results in several takes of the scene out of which the editor composes the image track. It also creates the production sounds. Production sounds may be dialogues, natural sounds associated with the scene - for example the sound of a door bang when a person shuts the door -, and ambient sound - such as the sound of a crowd during a fair. From an aesthetic point of view it is important to capture as much of the ambience and dialogues on the location. However, the recording levels of sounds captured on location may be uneven, there may be some 'dead' spots, sound effects may have to be replaced or enhanced, ambient sound may not be continuous, etc. [9]

In post-production, clean dialogues, sound effects and music will be created on several sound tracks for the scene. The aim of post-production sounds is to lend coherence, complete the picture and enhance the story. An example of good coherence between the image and the sound track is the fading out of music in parallel to a fade out of a shot. Post-production sounds are used in a variety of ways during editing to support continuity in narrative films. As shown in Figure 1.2 there are different categories of post production sounds that are used in films. Each of the post-production sounds can be used individually or collectively, for a variety of purposes, such as creating drama in a scene, stringing together multiple shots into a coherent a scene, building a sense of anticipation for a scene that is yet to appear, set up the emotional state of the audience for coming events and so on.

The sound recordist needs to find a way to achieve the best sound possible in the context of camera movements, lighting setup and other visual constraints. As part of production planning, the sound preparation stage involves planning how much dialogue to use, how many characters, nature of locations, extra sounds that must be recorded at the location, other ambient sounds and live effects such as rustling of clothes, noise of steps, etc. An important goal of recording dialogue is that it should be consistent with the point of view of the camera and from the perspective of the lens used for the shot. In order to make the sounds consistent with the camera, dialogue and sound effects may be on different tracks giving control over loudness, and used appropriately to match the visual perspective. For example when closing up on a talking couple within a crowd, the crowd sound may be dominating the conversation at first. As soon as the camera focuses on the couple, their conversation becomes the foreground sound and the crowd sound recedes and becomes the background sound.

Eventually, the sound editor composes the production and post-production sound tracks together to create the final mix. The main audio editing approach that he uses to construct scenes is called **sound overlap** and describes a certain sound (be that speech, music, sound effect, silence, or ambience) continuing over a shot boundary. This indicates to a viewer that the two shots are connected into a scene. It is most prominently used in shot/reverse shot dialogue scenes where the dialogue and the ambient sound continues over the shot boundaries. Another example, not drawn from feature films but rather from documentaries or news storeis, is the continuation of a narrator's voice over several shots.

Current editing practice also uses so-called **sound bridges**. A sound bridge is a sound that either belongs to a previous scene but is kept longer, or begins at the end of the previous scene but belongs to the

next. This effect is used to bridge the viewer into the next scene either by letting him ponder longer about the consequences of the previous scene, or by anticipating the next action.

The use of the mentioned sound classes silence, music, sound effects, voiceover, and speech for the construction of scenes is now illustrated on an example: the movie 'Titanic'. Through discussion of this example, we can also explain in more detail the general use of the different sound classes for creation of narrative coherence.

Silence. In 'Titanic', silence is not used to link shots together into a scene. It is rather used to create a sense of tension in the viewer before an important decision or discovery, or to make a transistion between scenes of the past and the present. An example is the transition between the scene that shows the main actor (Jack) kissing the main actress (Rose) which is followed by a period of total silence. The next scene shows Rose as an old woman reminiscing about the episode. And later, there is again silence when there is a flash back and the scene goes back to the ship. Here silence is used as a punctuation to create a temporal ellipsis.

Music. Music is a dominant sound that is used throughout the 'Titanic' movie to connect multiple shots. Music is often used to bind the picture together, particularly over cuts and transitions. It connects shots that may not have apparent connections. An example are the film titles. They consist of shots from the start of the journey of the Titanic in 1912 and the present in which divers are discovering the sunk ship. The music thus transports the audience to another place and time.

Another prominent example is the scene in which Jack draws a sketch of Rose. This scene consists of a number of distinct shots taken from different angles as he sketches her. What binds these shots into a coherent scene is the continuity of the music throughout the drawing period.

When music is used to connect shots, it may also drive the shot sequence, instilling it with energy. When Rose and Jack flee from Rose's room after the drawing scene, the music changes to an Irish tune, which becomes the dominant music during the following chase. The subsequent shots show the workers in the engine room, where they arrive while trying to get away from their pursuers. The piece of Irish music acts as a sound bridge that transports the viewers from the deck to the engine room.

It is mostly post-production music which is used to connect shots together. However, production music can serve the same purpose. On

the 'Titanic', for example, the shots of the Sunday sermon on the ship are unified by the choral singing of the attendants.

Sound effects. Sound effects are often used to direct, guide and shift the attention of the viewer both in the temporal and spatial dimension, thus creating a sense of drama in a scene. For example, during the conversation at the dinner scene, there is a sharp sound of cutlery, and the viewer's focus is immediately turned on Jack and his speech about how he won a free ticket to get on board the ship.

The ways in which sound effects can be combined to create a continuous stream of information is illustrated during the scenes that show the evacuation of passengers from the ship to the lifeboats. Here, a variety of sound effects - such as sounds of breaking glass, water waves at various loudness levels, and screaming sounds of people - are introduced to highlight the commotion and confusion of the passengers. This recurring mix of sound effects unifies the sequence into a coherent set.

In the 'Titanic' sound effects are however rarely used to connect different shots into coherent scenes. That also stems from the commonly short-timed nature of sound effect. For example, the toot of the ship's siren is used several times to introduce a sharp break from an indoor scene to an outdoor scene.

Voiceover. An special category of post-production sound is the voiceover. A voiceover is a separate voice that is not in sync with the picture. It may represent the main character commenting or narrating the story. For example, the voice of Rose as the old woman serves as a voice over to move between two periods of time and lends poignancy to the romance on the Titanic.

Speech. Speech overlaps occur frequently in the 'Titanic'. As an example take the scene at the ship's stern where Jack stops Rose from jumping off by talking to her. Their dialog covers several shots. Most other dialogues in the movie function similarly. However, not that music is often used to fill in gaps in speech breaks or as background sound to a dialogue.

As the sounds that accompany the moving image have become increasingly sophisticated, the final sound track - which is a judicial and artistic combination of all categories of sounds shown in Figure 1.2 - has a profound impact on the audience's response to the world inhabited by the characters.

3. Automatic extraction

After identifying the kind of scenes that are built for narrative coherence using production sounds and post-production sound editing, we now proceed to a discussion of the feasibility of automatic determination of these scene types. The aim here is to illustrate how to link high-level cinematic scene structures and low-level signal analysis.

We first enumerate the assumptions on which our discussion is based:

1 Material: only the final edited and mixed film is available - there is no access to the different video and sound tracks out of which the film was composed.

2 Format: digitised sound and video tracks are available, possibly in a compressed format such as MPEG-2.

3 Shot segmentation: a (highly reliable) set of shot boundaries is available, which might have been determined via video analysis or manual extraction.

Based on these preconditions, we will now examine how we can automatically determine scenes by analysing the sound track.

3.1 Scenes created by narration

Wang et al state that "the clustering of 'shots' into 'scenes' depends on subjective judgement of semantic correlation." ([10], p.20). This semantic correlation is built by the narration. For example, in a documentary a topic may correlate shots semantically, while in a news broadcast this is done by the news stories. In feature films it may be a certain action or event that produces the semantic correlation.

On the sound track, the narration of a video is presented mainly via the spoken words. Therefore, the textual transcription of a video may be used for an analysis of the semantic relationship between shots and a grouping of shots into scenes. In this case, automatic speech recognition (ASR) and a linguistic analysis of the resulting transcription are required to link shots through the context of the narration. ASR may be used successfully on clean studio speech recordings as are common for documentaries or studio news broadcasts. Unfortunately, current ASR results are not very reliable on general film sound tracks because of the large amount of other sounds present at the same time. Therefore this approach is bound to be not very successful on most types of film material nowadays.

3.2 Scenes created by editing

Instead of performing a linguistic analysis, Wang et al [10] propose another approach: "(..) sometimes it is possible to recognize shots that are related in locations or events, without actually invoking high-level analysis of semantic meanings." ([10], p. 20). Our analyses of film production practices confirm this statement: sound editing is often used to convey the narrative structure to the viewer. It is therefore possible to attempt identification of scenes which are created through sound overlaps.

There are two fundamentally different approaches to the automatic determination of such scenes: the first is a top-down approach which starts from the art of film making, and uses signal analysis on the sound track for identification of sound overlaps in a way similar to a human analyser. The second approach is a data-driven bottom-up approach and starts from the kinds of audio features that are available. It investigates change patterns of the features that help in determining relations between consecutive shots and clustering them into scenes.

3.2.1 Top-down approach. We describe five basic sound classes which are distinguished during film production:

- speech,
- music,
- sound effects,
- sound ambience, and
- silence.

A human analyser who tries to identify sound overlaps will start by identifying these sound classes in the film and determining the temporal segments during which they occur. Some algorithms have been developed to perform that task automatically. Implementations of sound segmentation approaches usually extract features on short time frames (10-50 ms) and classify them into one of their considered sound classes. This often results in a highly segmented sound track; so some publications propose that a more accurate segmentation can be derived by integrating sequential frames into longer segments according to some heuristics (such as an n-gram approach where rows of n segments of the same class are detected). This approach determines non-overlapping segments with one specified sound class only. So, at intervals during which some of these classes occur simultaneously, it can only determine the dominating class.

Unfortunately, no current publication on sound segmentation and classification distinguishes between the same sound classes as the ones listed above. Zhang and Kuo [11] get the closest by distinguishing between silence, speech, music, environmental sound with special features, and other environmental sound. Most regard music and speech only, see for example [12, 13, 14, 15, 16, 17, 18, 19]. Some also include silence and other sounds, as in [20, 21, 22, 23, 24, 25].

Let us presume that automatic identification of the above listed sound classes is feasible. The most straight-forward approach toward identification of sound overlaps is then based on the segmentation of the sound track into intervals classified as one of these sound classes by identification of the dominant sound. Integration of sequential shots into scenes is performed where a determined segment of one class overlaps a shot boundary. Figure 1.3 illustrates this approach.

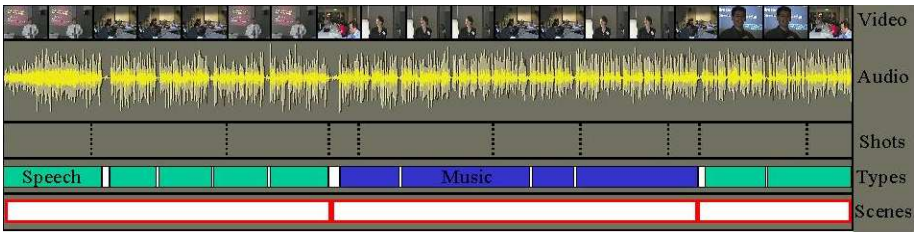

Figure 1.3. Simple top-down sound overlap identification.

There are a number of drawbacks with this approach:

1 it cannot handle changes of sound within one sound class,

2 it cannot handle sound overlaps which are interrupted at the shot boundary by another dominating sound class, and

3 it cannot handle sound bridges.

Let us look at each one of the drawbacks.

A few examples demonstrate the first one: assume that one shot ends with music and the next one starts with music, though a completely different type. The simple approach would merge them into one scene although they really belong to different ones. Another example is a dialogue between two people that ends in one shot followed by a dialogue between two different people that starts in the next one. Grouping shots connected by the same sound class will often associate shots that should be kept separate. Therefore, there is a requirement to not only segment into the given sound classes but to subsegment within a sound class.

Music, for example, needs to be segmented in case of strong changes, speech in case of speaker changes (such as in [26]), and ambience as soon as the background sound environment changes substantially. Figure 1.4 illustrates this approach.

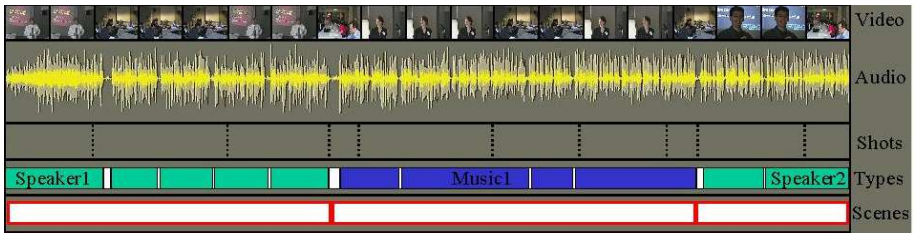

Figure 1.4. More detailed top-down sound overlap identification.

The second problem often occurs with dialogues. Pauses in dailogs are very common. Such pauses contain the sound ambience, but in films are often filled by music. As a result, dialogues usually come in bursts with either silence, ambience, or music segments in between. While dialogue overlaps are common in continuity editing, their automatic identification therefore may not be simple. A similar problem occurs where the sound ambience might be dominated by a short sound effect at the shot boundary. One approach to overcome some of these problems is to look at each sound class separately and identify its occurrance. Instead of creating only a single segmentation of a sound track, this will create five segmentations which overlap at times. Figure 1.5 illustrates this approach. In this way, some of the original sound tracks' composition can be restored and exploited more indepth.

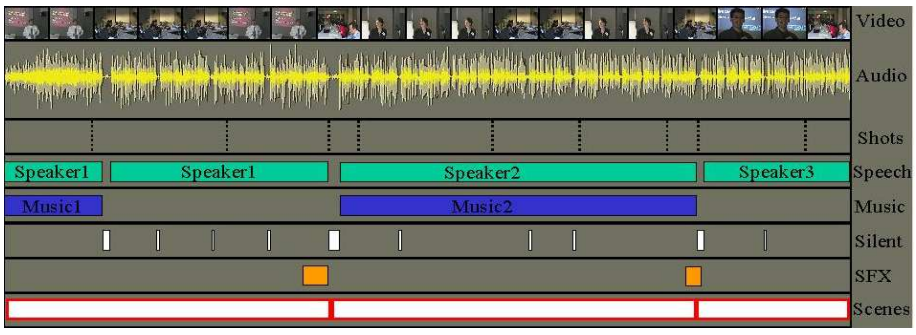

Figure 1.5. Multitrack top-down sound overlap identification.

Sound bridges of course are a counterexample to the use of sound for continuity editing. They create a problem as they connect shots which really belong to different scenes. This can often be irritating to human viewers, who overcome the situation through analysis of the narration. Our heuristic approach, however, will be quite helpless in this situation. One may be tempted to think that sound bridges start (or end) closer to a scene boundary than sound overlaps. Consequently, one would only need to define a temporal interval around a shot boundary during which overlapping sound is taken as a sound bridge and ignored. There is however no general rule to which sound editors adhere for timing differences between sound overlaps and sound bridges. Therefore, the best way to distinguish between a sound bridge and a sound overlap is probably during a post-processing step where identified scenes are re-examined for their sound ambience consistency.

3.2.2 Bottom-up approach. Even without determination of the five basic sound classes, patterns in features extracted from the sound data may indicate the shots that belong together into a scene.

One approach is to generically segment the audio stream into intervals containing consistent feature patterns. It is expected that a sound overlap is covered by an audio segment such that the shots of a scene get fused by audio segments overlapping the shot boundaries. A scene is then characterised by both a video and an audio segment boundary.

In analogy to the video segmentation approach, research has approached the problem of generic audio segmentation as a problem of finding significant changes in feature vectors. There is a large set of possible audio features that may be used in a feature vector:

- **transform-based features** such as spectral [27], cepstral [28, 29, 30, 31], linear predictive coefficients [32], or linear spectral frequencies [33],

- **physical features** such as energy statistics, zero crossing statistics, spectral centroid, spcetral bandwidth, or spectral peak, (see [34, 35, 36, 37]),

- and **perceptual features** such as pitch, tonality, harmonicity, pulse metric, or silence density (see [11, 28, 32, 38, 13]).

This is an open list as new audio features emerge every day.

As in the top-down approach, features are calculated on short analysis windows (10-100ms). They are integrated into a compact feature vector on larger temporal windows (1-3s) using statistical methods. Distances

between these feature vectors are calculated and significant changes are characterised by large distances, which result in segment boundaries.

This approach works amazingly well for music, silence, or ambience overlaps. The reason is that the spectral composition of these sound classes is relatively stable. In contrast, human speech is composed of speech bursts and pauses. Therefore, speech overlaps are more difficult to grasp with this generic approach.

One way to overcome this problem is by calculating the distance between all audio segments of two neighbouring shots and only cluster them into one scene in case of a very strong similarity of at least one general audio segment in each shot. This approach also provides a handle to sound bridges because the fact that a sound overlaps a shot boundary is less important than the fact that the neighbouring shots contain highly similar sound.

A completely different approach is to use a sequence of feature vectors to calculate a model for specific scene tyes. A model (as in a hidden Markov model, HMM) consists of a probabilistically trained sequence of feature vectors which represent a typical pattern for a specific scene type. Thus, heuristics on patterns or training of feature vector sequences on scene types may be used to determine which shots to group together.

4. Implemented approaches

This section gives an overview of publications on existing research approaches using audio analysis for scene detection. It follows the three subsections established in the previous section. We start by examining systems that implement scene determination by linguistic analysis, thus detecting narration breaks.

4.1 Scenes determined by linguistic analysis

The Informedia Digital video library project makes use of externally available transcripts or ASR transcribed soundtracks for scene segmentation. Hauptmann et al. [39] first identify shots (which they call scenes) via video cut detection. In a second step they identify scene boundaries (which they call video paragraphs) via natural language processing and silence analysis. With ASR, they have to cope with a word error rate in the range from 20% to 70% depending on the quality of the speech recording. Results seem quite promising.

Because of the large word error rates associated with current ASR systems, closed captions have been used as another source for getting high-accuracy textual transcripts. One such system has been presented by Huang et al. [40]. They automatically segment TV news into news

stories, story introduction, augmented news stories, and news summary using audio analysis only. General audio features such as the non-silence ratio and the standard deviation of the zero crossing rate (ZCR) are first used to separate out commercial breaks from the recorded newscast. Then, they identify anchor person shots using text-independent, closed-set speaker identification with a trained Gaussian mixture model. Finally, the news stories are extracted via a discourse-based segmentation on the closed captions. The accuracy that they achieve is high: on a 2 hour test database, they determine all scene boundaries correctly while finding no incorrect boundaries.

Another system that uses only audio information for topic segmentation of news broadcasts has been presented by the SRI MAESTRO team [41]. To find the topic boundaries, they extract prosodic information from the speech waveform (pause and pitch patterns) and calculate word usage statistics from the ASR transcript. Both sources of information are combined in a HMM to calculate the topic segmentation. In a query they would thus prompt the user for keywords and return associated scenes.

Most systems performing scene segmentation by linguistic analysis have worked on news broadcasts. The last such system that we would like to mention here is called Rough'n'Ready [42]. It segments radio news by speaker gender and speaker, creates a transcript with ASR, augments this transcript with punctuation and capitalisation, and segments it into paragraphs and stories via language analysis. The extracted information is used as an index into a collection of news broadcasts, and for search and retrieval on this collection. Extracted story boundaries seem to be highly reliable.

Before continuing on from the use of linguistic analysis to sound classification for scene segmentation, we would like to mention that many publications perform speaker segmentation on TV or radio broadcasts also with the aim of accessing news topics more easily. However, as they do not explicitly detect scene or story breaks but rather imply such a break at a speaker change, they are not regarded here.

4.2 Scenes determined by sound classification

Segmentation of TV news broadcasts has also been a main target of scene segmentation publications using sound classification. All use shot boundaries calculated from video track analysis.

Jiang et al. [24] first classify sound segments on a 1 s resolution into speech, music, environmental sounds, and silence. Speech segments are further distinguished into different speakers. This results in a set of audio

breaks determined as sound segment boundaries. Shot boundaries that coincide with an audio break are taken as scene change candidates. They are confirmed if they coincide with color correlation breaks calculated on the video track. With this approach they reach a recall rate of 91.9% and a precision of 86.8% on a set of 800 shots and 100 scenes.

Nam et al. [43] also work on TV news broadcasts. They determine silence and speaker segments and accept such shot boundaries as scene breaks where a speaker change coincides with a shot boundary. Results are promising.

Other publications that use sound classification for scene segmentation focus on analysing more general film material such as movies or documentaries. All of these also use shot boundaries computed on the video track.

Saraceno and Leonardi [44] distinguish between dialogues, stories, action scenes, and generic scenes. They regard the following sound classes: silence, speech, music, and noise. Video shots are grouped such that audio and visual characteristics follow predefined visual patterns for the four scene types:

- dialogue scenes are determined where the audio contains mostly speech and the shot pattern is alternating ABABABA...
- story scenes contain mostly speech and have a shot pattern that repeats some content ABCADEFGAH...
- action scenes contain mostly non-speech and a non-repeating shot pattern ABCDEF...
- generic scenes are all the rest.

Shot patterns are identified via a similarity measure between shots which are represented as a vector quantisation (VQ) codebook with distortion.

Automatic segmentation of dialogue scenes is at the core of the publication of Alatan et al. [21]. They detect dialogue scenes in movies using a multi-modal HMM-based approach. To that end, sounds are classified into speech, music, and silence. Face detection is performed on the video track. Shots at the same location are clustered based on colour similarity. Then they set up a HMM containing different stages of a dialogue scene (establishing scene, dialogue scene, transitional scene) and train it with feature vectors containing tokens from the audio classification (music / silence / speech), face detection (face / no-face), and location change (changed / unchanged). They achieve an accuracy of about 95%.

4.3 Scenes determined by feature patterns

Accuracy in automatic sound classification is still very low. Some publications have therefore relied less on semantic analysis. The structure of low-level feature sequences was used for determination of scenes instead.

Pfeiffer, Lienhart, and Effelsberg [45] compute audio clips by determining significant changes in the sound track. To that end, background segments of 0.5 s minimum duration are calculated first. Audio cuts are then identified as significant changes. Transitions between fore- and background and audio cuts determine the boundaries for audio clips. Each such clip is represented via an audio feature vector calculated on 100 ms windows. Differences of feature vectors between all the audio clips of subsequent shots are calculated. Shots are clustered together if differences between them are small. The hit rate on two full feature films was around 63%.

Nam and Tewfik [46] propose audio segmentation by detection of sharp changes between spectra of 80 ms windows. If a segment boundary coincides with a shot break, a scene boundary is established. They only examine this approach for the detection of TV commercials and are successful in detecting three commercial boundaries.

Huang et al. [47] detect audio breaks by computing significant changes in sound using a dissimilarity index with 12 features (such as non-silence ratio, volume dynamic range) on 1 s clips. Frames with both, shot breaks and audio breaks, are declared scene boundaries. They examine news, commercials, and sports film material and achieve a 100% hit rate with 30% false detections.

Finally, we examine two feature vector based scene segmentation approaches that only use audio information for segmentation. They both achieve promising results, but a combination with shot boundaries from video analysis should bring much higher accuracy.

Liu et al. [48] use feature vectors on 1 s clips. Distances between the current clip and several previous and following clips are calculated. A clip is declared a scene change point if it is similar to the six following clips and different from the six preceding ones. The accuracy of their approach is high (most scene boundaries are found), but with about 100% false alarms.

Kemp et al. [49] examine three different types of generic audio segmentation: model-based, metric-based, and hybrid segmentation for story segmentation in TV news broadcasts. The model-based segmentation creates a set of models for different acoustic classes, trains them,

and classifies the audio stream using the model (HMM or GMM) nominating class boundaries as story boundaries. The metric-based segmentation uses maxima in distances between feature vectors of small frames as segment boundaries. Their best result is achieved with a hybrid segmentation, performing metric-based clustering on larger frames (chunks of size 1 s) and use of the clusters to train models and perform segmentation. This integration increases the F-measure from about 60% to 78%.

5. Conclusions

This chapter described the process of constructing scenes using auditive effects introduced during film production. It continued into a discussion on approaches for automatic determination of identified scene types. Finally, an overview of existing implementations was presented. Our survey shows that none of the existing implementations are capable of identifying all scene types discussed, but that some very fundamental techniques have been developed to support that task.

References

[1] C. Metz, *Film Language: A Semiotics of the Cinema*, The University of Chicago Press, 1974, trans. by M. Taylor.

[2] D. Bordwell and K. Thompson, *Film Art: An Introduction*, McGraw-Hill, New York, 5th edition, 1997.

[3] P. Aigrain, H. Zhang, and D. Petkovic, "Content-based representation and retrieval of visual media," *Multimedia Tools and Applications*, vol. 3, pp. 179–202, 1996.

[4] B. Adams, C. Dorai, and S. Venkatesh, "Study of shot length and motion as contributing factors to movie tempo," in *Proc. ACM Multimedia 2000*, Los Angeles, CA, USA, November 2000, pp. 353–355.

[5] C. Lindley and U. Srinivasan, "Query semantics for content-based retrieval of video data: An empirical investigation," in *Storage and Retrieval Issues in Image- and Multimedia Databases, in conjunction with 9th International Conference DEXA98*, Vienna, Austria, Aug 1998.

[6] U. Srinivasan, C. Lindley, and B. Simpson-Young, *Database Semantics- Semantic Issues in Multimedia Systems*, chapter A Multi-Model Framework for Video Information Systems, pp. 85–108, Kluwer Academic Publishers, Jan 1999.

[7] U. Srinivasan, S. Nepal, and G. Reynolds, "Modelling high level semantics for video data management," in *Proceedings of ISIMP 2001*, Hong Kong, May 2001, pp. 291–295.

[8] R. Stam, R. Burgoyne, and S. Flitterman, *New Vocabularies in Film Semiotics: Structuralism, Post-Structuralism, and beyond*, Routeledge, 1996.

[9] Rea and Irving, *Producing and Directing the Short Film and Video*, Focul Press, 1995.

[10] Y. Wang, Z. Liu, and J.-C. Huang, "Multimedia content analysis," *IEEE Signal Processing Magazine*, vol. 17, no. 6, pp. 12–36, Nov 2000.

[11] T. Zhang and J. C.-C. Kuo, "Heuristic approach for generic audio data segmentation and annotation," in *Proc. ACM Multimedia*, Orlando, 1999, pp. 67–76.

[12] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Atlanta, Georgia, USA, 1996, pp. 993–996.

[13] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, München, April 1997.

[14] K. Minami, A. Akutsu, H. Hamada, and Y. Tonomura, "Video handling with music and speech detection," *IEEE Multimedia*, vol. 5, no. 3, pp. 17–25, July-September 1998.

[15] David Gerhard, "Ph.D. depth paper: Audio signal classification," Tech. Rep., School of Computing Science, Simon Fraser University, Burnaby, Canada, February 2000.

[16] Arnaud Philibert, "Speech/music discriminator," Tech. Rep., Tampere University of Technology, Department of Information Technology, 1999.

[17] G. Williams and D.P.W. Ellis, "Speech/music discrimination based on posterior probability features," in *Proc. EuroSpeech*, Budapest, Hungary, September 1999, pp. 687–690.

[18] G. Tzanetakis and P. Cook, "Sound analysis using MPEG compressed audio," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing ICASSP 2000*, Istanbul, Turkey, 2000, IEEE, vol. 2, pp. 761–764.

[19] G. Lu and T. Hankinson, "An investigation of automatic audio classification and segmentation," in *Proc. 5th Intl. Conf. on Signal Processing WCCC-ICSP 2000*. IEEE, 2000, vol. 2, pp. 776–781.

[20] N.V. Patel and I.K. Sethi, "Audio characterization for video indexing," in *Proc. SPIE, Storage and Retrieval for Still Image and Video Databases IV*, San José, CA, USA, February 1996, vol. 2670, pp. 373–384.

[21] A.A. Alatan, A.N. Akansu, and W. Wolf, "Comparative analysis of hidden markov models for multi-modal dialogue scene indexing," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing ICASSP 2000*, Istanbul, Turkey, 2000, IEEE, vol. 4, pp. 2401–2404.

[22] H.S.M. Beigi and S.H. Maes, "Speaker, channel and environment change detection," in *Proceedings of the World Congress on Automation, 1998*, Anchorage, Alaska, May 1998, pp. 18–22.

[23] K. El-Maleh, A. Samouelian, and P. Kabal, "Frame-level noise classification in mobile environments," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Phoenix, AZ, USA, Mar 1999, pp. 237–240.

[24] H. Jiang, T. Lin, and H. Zhang, "Video segmentation with the assistance of audio content analysis," in *Proc. IEEE Intl Conf. on Multimedia and Expo, ICME 2000*. IEEE, 2000, vol. 3, pp. 1507–1510.

[25] Y. Nakajima, Y. Lu, M. Sugano, A. Yoneyama, H. Yanagihara, and A. Kurematsu, "A fast audio classification from MPEG," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Phoenix, Arizona, USA, May 1999, vol. IV, pp. 3005–3008.

[26] S. Tsekeridou and I. Pitas, "Audio-visual content analysis for content-based video indexing," in *Proc. IEEE Intl. Conf. on Multimedia Computing and Systems (ICMCS)*, 1999, vol. 1, pp. 667–672.

[27] T. Zhang and J. C.-C. Kuo, "Hierarchical classification of audio data for archiving and retrieving," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 1999, vol. IV, pp. 3001–3004.

[28] M.J. Carey, E.S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," in *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 1999, vol. 1, pp. 149–152.

[29] D. Kimber and L. Wilcox, "Acoustic segmentation for audio browsers," in *Proc. Interface Conference*, Fairfax, 1996, pp. 295–304.

[30] M.S. Spina and V.W. Zue, "Automatic transcription of general audio data: Preliminary analyses," in *Proc. Intl. Conf. on Spoken Language Processing, ICSLP 96*, Philadelphia, PA, Oct 1996, vol. 2, pp. 594–597.

[31] J. Foote, "Content-based retrieval of music and audio," in *Proc. SPIE, Multimedia Storage and Archiving Systems II*, C.-C.J. Kuo and others, Eds., San José, CA, USA, 1997, vol. 3229, pp. 138–147.

[32] G. Tzanetakis and F. Cook, "A framework for audio analysis based on classification and temporal segmentation," in *Proc. 25th EUROMICRO Conference, 1999*. IEEE, 1999, vol. 2, pp. 61–67.

[33] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal, "Speech/music discrimination for multimedia applications," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing ICASSP 2000*, Istanbul, Turkey, 2000, IEEE, vol. 4, pp. 2445–2449.

[34] R.M. Aarts and R.T. Dekkers, "A real-time speech-music discriminator," *J. Audio Eng. Soc.*, vol. 47, no. 9, pp. 720–725, 1999.

[35] L. Wyse and S. Smoliar, "Toward content-based audio indexing and retrieval and a new speaker discrimination technique," in *Proc. International Joint Conference on Artificial Intelligence IJCAI*, Montreal, Aug 1995, pp. 149–152.

[36] A. Samouelian, J. Robert-Ribes, and M. Plumpe, "Speech, silence, music and noise classification of TV broadcast material," in *Proc. Intl. Conf. on Spoken Language Processing*, Sydney, 1998, pp. 1099–1102.

[37] C. Saraceno and R. Leonardi, "Audio as a support to scene change detection and characterization of video sequences," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, München, Mai 1997, pp. 2597–2600.

[38] S. Venugopal, K.R. Ramakrishnan, S.H. Srinivas, and N. Balakrishnan, "Audio scene analysis and scene change detection in the MPEG compressed domain," in *IEEE Third Workshop on Multimedia Signal Processing, MMSP 19999*. IEEE, 1999, pp. 191–196.

[39] A.G. Hauptmann and M.A. Smith, "Text, speech, and vision for video segmentation: The Informedia project," in *AAAI-95 Fall Symposium on Computational Models for Integrating Language and Vision*, November 1995, pp. 90–95.

[40] Q. Huang, Z. Liu, A. Rosenberg, D. Gibbon, and B. Shahraray, "Automated generation of news content hierarchy by integrating audio, video, and text information," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Phoenix, AZ, USA, Mar 1999, pp. 3025–3028.

[41] SRI MAESTRO Team, "Maestro: Conductor of multimedia analysis technologies," *Communications of the ACM*, vol. 43, no. 2, pp. 57–63, Feb 2000.

[42] F. Kubala, S. Colbath, D. Liu, A. Srivastava, and J. Makhoul, "Integrated technologies for indexing spoken language," *Communications of the ACM*, vol. 43, no. 2, pp. 48–56, Feb 2000.

[43] J. Nam, A. Cetin, and A. Tewfik, "Speaker identification and video analysis for hierarchical video shot classification," in *Proc. IEEE Intl. Conf. on Image Processing (ICIP)*, Santa Barbara, CA, USA, Oct 1997, vol. 2, pp. 550–555.

[44] C. Saraceno and R. Leonardi, "Identification of story units in AV sequences by joint audio and video processing," in *Proc. Intl. Conf. Image Processing (ICIP-98)*, Chicago, IL, Oct 1998, vol. 1, pp. 363–367.

[45] S. Pfeiffer, R. Lienhart, and W. Effelsberg, "Scene determination based on video and audio features," *Multimedia Tools and Applications*, vol. 15, pp. 363–384, 2001.

[46] J. Nam and A. H. Tewfik, "Combined audio and visual streams analysis for video sequence segmentation," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Munich, Germany, Apr 1997, vol. IV, pp. 2665–2668.

[47] J. Huang, Z. Liu, and Y. Wang, "Integration of audio and visual information for content-based video segmentation," in *Proc. IEEE Intl. Conf. Image Processing (ICIP-98)*, Chicago, IL, Oct 1998, vol. 3, pp. 526–530.

[48] Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," *Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, vol. 20, no. 1/2, Oct 1998.

[49] T. Kemp, M. Schmidt, M. Westphal, and A. Waibel, "Strategies for automatic segmentation of audio data," in *Proc. IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing ICASSP 2000*. IEEE, 2000, vol. 3, pp. 1423–1426.