SoundRiver: Semantically-Rich Sound Illustration

H. Jänicke¹, R. Borgo¹, J. S. D. Mason², and M. Chen¹

¹ Computer Science Department, Swansea University, UK
² Electrical Engineering Department, Swansea University, UK

Abstract

Sound is an integral part of most movies and videos. In many situations, viewers of a video are unable to hear the sound track, for example, when watching it in a fast forward mode, viewing it by hearing-impaired viewers or when the plot is given as a storyboard. In this paper, we present an automated visualization solution to such problems. The system first detects the common components (such as music, speech, rain, explosions, and so on) from a sound track, then maps them to a collection of programmable visual metaphors, and generates a composite visualization. This form of sound visualization, which is referred to as SoundRiver, can be also used to augment various forms of video abstraction and annotated key frames and to enhance graphical user interfaces for video handling software. The SoundRiver conveys more semantic information to the viewer than traditional graphical representations of sound illustration, such as phonoautographs, spectrograms or artistic audiovisual animations.

Categories and Subject Descriptors (according to ACM CCS): I.3.4 [Graphics Utilities]: Picture description languages—I.3.4 [Graphics Utilities]: Paint systems—I.3.8 [Applications]: Sound illustration

1. Introduction

Imagine the difference if we could not hear the sound track when watching a movie. Imagine what sensuality hearing-impaired persons could gain if they can make some sense about the sound track accompanying a movie. While closed-captioning in cinemas or subtitles on DVDs have provided a great assistance in appreciating the speech and dialogs in a movie, there are many other sound features (e.g., mood of music, bird singing, traffic noise, etc.) that are difficult to convey using text-based audio descriptions. Sound is commonly depicted graphically. For example, the MS Windows Media Player transforms music to a colorful pattern animation synchronized with the beat of the music (Figure 1). There are graphical representations, such as the audio signal plot and sound spectrogram, commonly used in a technical context. However, none of these can provide semantically meaningful information to most, if not all, hearing-impaired persons to help them appreciate these rich sound features.

Figure 1: A graphical representation of a piece of music as provided by the MS Windows Media Player.

Figure 2: A video player with an automatically generated SoundRiver (“The Golden Compass”, ©New Line Cinema).
In this paper, we present an automated visualization solution to such problems. The system first extracts the common components (such as music, speech, rain, explosions) from a sound track, then maps them to a collection of programmable visual metaphors, and finally generates a compositional visualization. We refer to this form of sound illustration as SoundRiver, which was inspired by the theme river 

\[ \text{HWHN02}. \] A SoundRiver can be integrated into a movie player (Figure 2), and provides the viewer with a visualization of the sound track in normal play as well as in the fast forward or rewind mode. It can also be used to augment various forms of video abstraction and annotated key frames. Its application is thereby well beyond the scope of assisting hearing-impaired people. In this paper, we show the SoundRiver in conjunction with key frames of videos, for the convenience of static illustration.

Sound-related research has frequently been featured in computer graphics and visualization, including visualization of acoustic waves (e.g., [DBM"06]), automatic sound synthesis (e.g., [TH92]), sound-assisted interaction (e.g., [KWT99]). However, the visual display of the sound has been limited to geometric representations of various attributes of the sound signals, such as volumes, waves, and particles. There is little effort for depicting sound semantically. The contributions of this work thus include:

- A novel graphical solution for visualizing the contents of a sound track in a variety of situations when the viewers are unable to hear;
- The design of a collection of programmable visual metaphors and the method for audio-visual mapping;
- A system pipeline that transforms a sound signal to semantic sound components, to intuitive visual metaphors, to geometric primitives and finally to a semantically-rich SoundRiver illustration;
- An evaluation of the usefulness of the SoundRiver in video players for hearing-impaired viewers.

The remainder of the paper is organized as follows. In Section 2, we first provide a brief overview of the state of the art in sound processing, and sound-related research in computer graphics and visualization. The pipeline of the SoundRiver generation process is described in Section 3. In Section 4, we describe the techniques used in this work for analyzing the sound track of a video. In Section 5, we present the design of programmable visual metaphors, and describe our method for audio-visual mapping. In Section 6, we outline the system pipeline for generating SoundRivers, and discuss the results of semantically-rich sound illustration. We draw our concluding remarks in Section 6.

2. Related Work

Audio signal processing became an essential technical component for radio broadcasting a century ago. Over the last four decades, the technology of audio signal processing and analysis has advanced rapidly, resulting in many technical products, ranging from echo cancellation in theaters to speech recognition on PDAs. The literature in this field is vast, and for a comprehensive coverage, readers may consult textbooks in the field (e.g., [RS78, Poh95]).

Several aspects of the research in audio signal processing and analysis have provided this work with necessary underpinning techniques, including in particular, content-based search and retrieval of audio data [WBKW96] and audio content classification and detection [MAHT98]. From the literature, we can observe the following technical advances, which provide the basis for this project.

- Several methods have been played central roles in audio signal processing and analysis, including the Hidden Markov Model (HMM) and the Gaussian Mixture model (GMM). They have reached a matured status, with many enhanced variations, and their usability has been confirmed by many applications. The availability of software tools, such as Mistral [d’A09], enables us to consider graphical applications based on the results of audio signal processing and analysis.
- The reported detection accuracies are promising, typically ranging from 60% to 100% (e.g., [MAHT98] for music and speech and [CNK09] for environmental sound). This suggests that it is technically feasible to extract audio features automatically from a sound track, and then visualize such features.
- The detection of semantic audio information is in general more reliable than that of semantic visual information. Thus, it is common to improve video handling using audio signal processing and analysis [WBKW96]. This suggests that it is technically advantageous to visualize semantic audio information in applications such as video segmentation, analysis and abstraction.

Visual representations of sound fall into three major categories mostly according to their \textit{modus operandi}. The first category includes various forms of signal representations, among which signal plots are the most common form (Figure 3 (top)). \textit{Spectrograms}, which are used commonly in sound and speech analysis, depict a sound signal as an image showing short-term spectral densities as a function of time. Figure
3 (bottom) shows a audio spectrogram where each pixel’s color represents the amplitude of a particular frequency (y-axis) at a particular time (x-axis). Chromagrams, which became a popular form of visualization recently in sound analysis, display a grid of colored blocks depicting the strength of that pitch chroma in the signal, assisting in the analysis of the harmonic structure of audio samples [CIDM09].

The second category includes visual representations that depict semantic information of a sound signal using textual, iconographical or metaphoric illustrations. Textual illustrations are commonly seen in comics [Bru84], and geometric primitives have been used in music-related abstract paintings [Kan26]. Ferguson et al. used metaphoric visualization as a feedback in musical training [FMC05].

The third category focuses on the use of dynamic imagery patterns to create an artistic impression of the audio signal. Techniques in this category typically map the attributes of an input signal, such as frequency and loudness, to abstract geometries and color patterns. Such visual representations are commonly used in commercial media players, such as MS Windows Media Player (Figure 1) [Mic09]. When synchronized with the music, the animated visual representations can provide listeners with an entertaining visual experience.

In recent years, video handling has become a topic in computer graphics and visualization. For example, Goldman et al. proposed to create a storyboard from a video using video abstraction [GCS06]. Bartoli et al. developed a method to align video frames to create a motion panorama [BDH04]. However, the resulting visual representations do not feature any auditory component. Ho-Ching et al. considered using two visual designs, namely a spectrogram and a 2D map showing sound ripples, to provide information about environmental sound to the hearing-impaired [HCMLL03]. The spectrogram is found non-intuitive, while the 2D map requires the specification of a 2D floor plan. Azar et al. proposed a multi-view visualization design for aiding the hearing impaired in daily life [ASAA07]. The design is a mixture of signal plots, 2D sound source maps, spectrograms, icons for sound classes, text, and animation of sign languages for recognized speech. They conducted an evaluation of the design in a school for the hearing impaired, and confirmed the potential usability of sound visualization in this particular context. However, their visual design exhibits several drawbacks namely (a) it takes the whole screen, (b) it contains non-intuitive visualization such as spectrogram, (c) it does not support temporal recalls, demanding full attention from the viewer. It is thus unsuitable for annotating video sound track.

3. SoundRiver Pipeline

Figure 4 shows the pipeline for the computation of the SoundRiver. It consists of four steps: sound processing, audio-visual mapping, graphical composition and optional combination of the different components in a video player.
In an optional final step, the resulting SoundRiver can be integrated into a video player (Figure 2), a keyframe storyboard (see Section 6 for details), or other forms video abstractions and video handling software.

4. Sound Feature Extraction

In this section, we describe the first processing step of the pipeline for generating the SoundRiver. The primary function of this step is to augment various sound components in a sound track with semantically meaningful tags. There is a huge collection of previous work on techniques for sound classification. The focus of this work has invariably been placed on the graphical illustration of sound, and the use of sound analysis techniques provides a means to deliver semantically-meaningful information to the illustration process. For self-containment, we briefly describe the techniques used in this processing module.

4.1. Sound Analysis

At a very abstract level, we can distinguish three major components in a typical movie sound track: speech, music and sound effects [MAHT98]. Their combination forms an auditory scene, which is an integral part of the story line of a movie and has direct influence upon the viewers’ impression of the corresponding imagery scene.

Methods for sound classifications fall into two main categories, low-level signal descriptors and high-level sound models. The former usually involves the decomposition of a sound signal into a collection of descriptors, for instance, decomposing a spectrogram into a vector representation, representing both the basis function (e.g., eigenvectors) and the projected features (e.g., eigen coefficients). The latter usually organizes different types of sound using a hierarchical classification scheme, defining the relationship between different categories and sub-categories. Figure 5(a) shows a small section of such a hierarchy.

A binary classification scheme is commonly used to recognize relatively unadulterated sound segments. For mixed sound, the detection rate declines rapidly in relation to the level of variety in the composite sound [MAHT98]. In this work, we assume that a sound segment is a mixed auditory scene, and different categories of sounds are not necessarily mutually exclusive. Conceptually, the organization of sound categories and sub-categories occurring in a sound track are better represented by a Venn or Euler diagram, as illustrated in Figure 5(b). We thus employ two approaches for sound feature extraction. For the classification of main categories, that is, music, speech and sound effects, we use a spectrum-based analysis, which can handle the contamination by other sound categories better. For individual sub-categories, we use a number of sound classification models, each of which is trained to detect a specific class of sound, such as bird singing, plane taking-off, etc.

4.2. Sound Spectrum

Both classification schemes used in this work are built upon the spectrogram representation. Figure 3 (bottom) shows a spectrogram of a one minute sequence of a movie. Each line in the image represents a certain frequency. The x-axis encodes time and the color tells how strongly a certain frequency contributes to the signal at a certain instance in time. Going from left to right, we can observe how the sound changes over time being sometimes heavily influenced by a few frequencies and sometimes being a combination of a large variety.

In this work, each column in the spectrogram is computed from a short sample (100ms) of the original audio signal sampled at 8kHz. Each sample is Fourier transformed using the well-known DFT equation:

\[
y_k = \sum_{n=0}^{N-1} X_n e^{-\frac{2\pi}{N} i n k}, \quad k = 0, \ldots, N-1,
\]

where \(X\) is the complex input signal consisting of \(N\) samples which is transformed into \(N\) frequency components \(Y_k\) and \(i\) is the imaginary unit. Thus, we obtain information about frequencies within the range 10 to 4000 Hz. To control the level of adjacent spectral artifacts, the original data is convolved with a Hamming window in the Fourier domain. The temporal distance between columns in the spectrogram is 250 samples which is approximately 31ms.

In the SoundRivers, we wish to illustrate the sound on a scale reflecting human hearing, i.e., as perceived by the human ear. Therefore, the derived spectrum is adjusted to reflect the perceived loudness for the volume computation of music and speech. Psychoacoustics is the study of subjective human perception of sounds [Pla05]. Research in this field revealed that the human hearing is logarithmic. The human ear is most sensitive at frequencies between 250 and 5000 Hz, which cover the range of speech. To account for these findings, the color scale is logarithmic using unit dBFS (decibels relative to full scale):

\[
a2 = 20 \times \log_{10}(a1/\text{max}),
\]
Figure 6: Spectral patterns of different sounds, from left to right, music, speech and an explosion.

where \( a_1 \) is the original value, \( a_2 \) the one measured in dBFS and \( max \) is the maximal entry in the spectrogram. Moreover, the values are modified according to the A-weighting defined in the international standard IEC 61672:2003.

Applying this procedure to the sound of a movie, we receive different visual patterns for different auditory events as depicted in Figure 6, which shows three example spectrograms representing music, speech and an explosion respectively. Music and related sounds such as chimes are characterized by horizontal lines in the spectrogram. Human speech uses a large range of individual frequencies that oscillate conjointly. Hence, with some training it is quite easy for a human to detect certain audio events. As most sounds such as music and speech feature a large variance, a model-based system will most likely be under-trained facing the diversity of variations and background noise. Hence, we use morphology-based techniques to perform the coarse-level classification and employ the model-based technique to identify more specific sounds at the finer level.

4.3. Morphology-based Classification

Morphological operators [Bur97] originate from the field of mathematical morphology and are based around a few simple concepts from set theory. They are used extensively in image analysis [Ser82]. A morphological operation can be thought of as a pattern matching process. A structuring element (a 2D pattern) is moved over each pixel in an image, and a specific logical operation is applied to the matched patterns. The most commonly used morphological operations are dilation and erosion. More sophisticated operations include opening and closing, which are combinations of dilation and erosion.

As we have seen earlier, music is characterized by stable horizontal lines in the spectrogram. To extract these lines, we first apply a lower threshold on the image to remove low level noise. We then apply three opening operations using a \( 15 \times 1 \) structuring element. When applied to a piece of sound (Figure 7(a)) with pure music in the beginning and additional speech at the end, we obtain the spectral patterns originating from music in Figure 7(b).

To extract speech, we use three openings and a \( 2 \times 4 \) structuring element to the thresholded input image. As can be seen in the results in Figure 7(c) chaotic sounds such as explosions would also be classified as “speech”. Unlike explosions, real speech is characterized by distinct gaps between individual frequencies. These can be detected using edge detection algorithms such as the Sobel filter. Figure 7(d) depicts edges in those time steps, where more than 40 pixels were marked as being an edge by the Sobel filter. We use the Sobel results to confirm whether the “speech” detected by the opening is actually speech. Therefore, we compute at each time step how many neighboring time steps contain neither “speech” detected by opening nor Sobel edges. If this value is larger than 3 seconds, the detected “speech” is dismissed. In Figure 7(c), the leftmost vertical line is dismissed, as no Sobel edges are detected within the neighborhood. We use a tolerance range of 3 seconds to cope with strong background signals such as music that might hide the speech patterns in small subsets.

The automatic morphology-based classification works very well for such distinct spectral patterns exhibited by music and speech. For the classification of sounds with more subtle patterns, a model-based approach is more effective, though supervised training is necessary.

4.4. Model-based Classification

Model-based classification is a technique where a statistical model is used to decide if a given sound sample belongs to a particular class of sound or not. Like the morphology-based classification, model-based approaches operate on
the spectral representation of the sound signal. Unlike the morphology-based approach, models are acquired by a training process called model inference. To train a model, one must collect a set of representative sound samples for a given sound class. The spectral features of sound samples are then extracted and used to optimize the parameters of the model statistically. Once the model is trained, it can be used to determine whether or not an arbitrary sound sample belongs to the corresponding sound class.

In this work, we used the model-based classification to detect a collection of sound effects often featured in movies, such as explosions, dog’s barking, the sound of an aircraft, a particular mobile phone ring tone, and the sound of rain. We took advantage of the availability of the open-source Mistral software [d’A09], which is based on the established approach using the Universal Background Model and the Gaussian Mixture Model (UBM/GMM). To evaluate the performance of the classification software, we used data from the following databases: BBC Stimulus Sounds, Aviation-Sounds, PartnersInRhymes, and TSP Lab database. We considered 22 different types of sound effects, and trained each model with 48 to 66 sound samples from different sources. We conducted on average 120 performance tests per model. The threshold used to determine if a sound belongs to a given class or not was set to 2.4 for all models. Sound samples that reach a score higher than 2.4 with a particular model are considered to belong to the given class and those below as outside the class.

Table 1 shows the performance results of a subset of our trained models. The first column states the name of the class of sounds, followed by the number of sample sources used to train the model. For the class male speech, for example, we used sound samples of 60 different male speakers to train the model. We took a five second sound sample from each source resulting in the total training size given in the column Duration. The trained model was tested on average with 120 different sources from 2–4 different classes. For example, the male speech model was tested using samples of male and female speech, dog barking and thunder. The score in each line of the table reflects the number of correct answers. The difference between the score and 100% gives the percentage of false positive and false negative answers. Scores below 80% are mostly due to undertraining of the model. The poor performance of the thunder model is due to the poor quality of the provided samples.

5. Audio-visual Mapping

In this section, we present a method for transforming extracted sound features to intuitive and semantically meaningful visual metaphors. We first outline the graphical design of SoundRiver. We then describe our approach to making visual metaphors programmable. This method facilitates a flexible mapping interface between sound features and visual metaphors, and allows visual metaphors to change their appearances dynamically in the SoundRiver in response to the sound features.

5.1. The Graphical Design of the SoundRiver

As overall design guideline we chose the metaphor of a river, because sound can also be thought of as something fluent and changing. Our SoundRiver, as depicted in Figures 11 and 12, expands from the top of the page to the bottom and comprises three principle strands for illustrating the sound track of a movie. The leftmost strand depicts the music contained in the movie. The central strand in gray is dedicated to the illustration of key events such as sound effects and key frames. The third strand illustrates the speech features. In these figures, there is also an optional fourth strand containing annotated key frames of a movie. Time lines are added across the different strands to ease readability and temporally connect the different components of the SoundRiver.

The first three strands have the same temporal scaling, i.e., if structures in the sound effect and speech section are depicted at the same height, they occur at the same instance in time. The images are aligned in order of their appearance and do not temporally coincide with the other elements in the SoundRiver. Labels in the key event section indicate the corresponding time.

For the SoundRiver, we chose a constant distance in time. Alternatively, one could have scaled the distances between key frames equally. With this approach, however, it would be very difficult to judge how long different events take, which is a crucial piece of information. Hence, we kept a fixed temporal distance and used additional icons to indicate key frames.

5.2. Visual Metaphors

Each of the three audio strands in the SoundRiver has a distinct layout to represent the structure of the audio contents.
Music attributes: volume and atmosphere

While the music is represented by an oscillating line which displays the rhythmic character of music, speech is depicted by speech bubble-like icons. For the sound effects we chose a rectangular layout to emphasize their function as reoccurring building blocks. Each component has several parameters that can be modified according to features of the given sound. Hence, we resolve programmable icons that provide more information about the underlying sound than a simple picture or icon. In the following, we will introduce the different parameters and mapping strategies used to transform sound into visual metaphors.

Music: The music strand is represented by the left riverbank of the SoundRiver. The two parameters, volume and atmosphere of music, are used for further manipulation. The width of the riverbank, i.e., expansion to the left, represents the volume of the music. The wider the river, the louder the music in the video. The SoundRivers in Figures 11 and 12 represent ten minutes of video material each. To achieve a smooth flow of the river, we used a spline interpolation of the original data to smooth local variations due to segmentation inaccuracies. Figure 8 (left) illustrates the mapping of the volume. We use a non-continuous jump from zero width for silence to a minimum distance for very silent music to allow for the recognition of the color of silent parts. On the right-hand-side of the same image, the colors for different atmospheres are depicted. All colors have approximately the same luminance. Thus, no color, i.e., atmosphere, sticks out and all of them are perceived equally important. If a particular mood is to be emphasized the luminance can be altered to make the color more striking.

Speech: For the illustration of speech we apply a style similar to speech bubbles. Figure 9 illustrates the mapping of different parameters. A rounded rectangle represents a piece of conversation. A new rectangle is started when the conversation stops for more than 0.75 seconds. Hence, each speech icon represents a continuous piece of speech and the gaps between two icons represent the duration of breaks. The combination of several speech icons gives an impression of the dynamics of the conversation.

Three additional parameters are mapped to size, style and color of the speech icons. Like with music, the width of the icon represents the volume of the conversation. Different shapes of the speech boxes, as shown in Figure 9, indicate the number of speakers involved. A single rectangle on the right-hand-side represents one speaker, two rectangles next to each other indicate a conversation and the presence of more than two speakers is indicated by a large rectangle filling the gap in the middle. The gender of the speakers is mapped to color. Yellow rectangles depict female speakers, blue color denotes male speakers and green symbolizes a group of mixed gender.

Sound effects: The middle strand of the SoundRiver is the depiction of the sound effects. Depending on the movie very different numbers and styles of sounds might occur. While in early Star Wars movies (Figure 12) a lot of sound effects are used to create the futuristic atmosphere, hardly any of them occur in the adventure/fantasy movie “The Golden Compass”. Hence, we decided to use a very versatile approach by depicting the different sound effects as symbolic icons as shown in Figure 10 (left). For each movie an appropriate set of icons can be selected and further manipulated to represent the individual auditory scenery.

The sound effect strand of the SoundRiver is defined by a light-gray rectangle intersected by the time lines. Additionally we add labels for the key frames in this part to have a central reference to the keyframe strand of the movie. The icons are placed in two columns along the time axes using a simple force directed layout algorithm. If possible, icons of the same type are painted on the same side to ensure visual coherence. We allow for small temporal displacements in the layout, to avoid occlusion of icons. Gray arrows indicate the point in time when a specific sound effect occurs. If the same events occurs several times within a short time interval, e.g., a dog barks repeatedly, several arrows are started from the same icon. A similar approach can be used for sound effects that last for a longer time period.

The parameters of sound icons that can be manipulated are the size and color of the shape in the icon. Figure 10 (right) illustrates the depiction of different bells. A church bell is mapped to a yellow bell, the bell of a bicycle is represented by a white shape and a ringing telephone is depicted in gray. Theoretically, a large variety of colors can be used to
Sound effects

<table>
<thead>
<tr>
<th>Sound effects</th>
<th>Sound effect subtypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>chimneys</td>
<td>quiet</td>
</tr>
<tr>
<td>door being opened or closed</td>
<td>loud phone</td>
</tr>
<tr>
<td>aircraft flying by</td>
<td>lightning</td>
</tr>
<tr>
<td>gun fire</td>
<td>本科 animal</td>
</tr>
<tr>
<td>bell of bicycle</td>
<td></td>
</tr>
<tr>
<td>church bell</td>
<td></td>
</tr>
</tbody>
</table>

Figure 10: Sound effect mapping: (left) Different icons are used in the SoundRiver to represent characteristic sound effects. (right) Subtypes are encoded using changes in color and scale.

In most movies, one or more sounds occur very often, such as gun fire in an action movie or sound of aircraft in StarWars. In such cases or if a certain sound effect shall gain special attention, the system extends the sound effect strand of the SoundRiver by an additional column. This function was applied for the appearance of the sound of footsteps in the movie “The Golden Compass” (Figure 11). As with the standard sound effect strand, the background is colored in light gray. Time periods where footsteps can be heard are colored in dark gray. The position of the footsteps indicates the frequency of the movement. If the icons are painted on the left hand-side, they indicate long intervals between successive steps. If they are located on the right, they indicate frequent footsteps as occur when a person is running.

6. Sound Illustration

Figures 11 and 12 depict the SoundRiver of ten minutes from the movies “The Golden Compass” and “StarWars III: The Revenge of the Sith”. Before going into detail with the resulting visualization, we will summarize the plot of each of the two sequences. From the “Golden Compass”, we selected the first ten minutes with a variety of different sound effects and from the StarWars movie a sequence that represents several different settings and atmospheres.

Plot of the Golden Compass (©New Line Cinema) sequence: After the title sequence (image 1), the narrator introduces the different characters in the movie (image 2). We meet the main character, Lyra Belacqua, while playing tag with her friends (image 3). During the game, the hunted children seek refuge in Jordan College. Lyra’s home, and Lyra offers the opponent party a deal (image 4) to save her friend from being caught: They will get Roger if one of them dares to put on the poisoned cloak Lyra is to “borrow” from Jordan College. Looking for a cloak, Lyra overhears (image 5) a discussion between the Master of Jordan College and Fra Pavel (image 6), a representative of the religious body, who tries to stop Lord Asriel’s, Lyra’s uncle’s, work. Failing in this attempt, he poisons Asriel’s wine (image 7). The plan, however, is put paid by Lyra, who finally lunges out of her hiding place and dashes the glass from Asriel’s hand (image 8). Asriel eventually meets the Jordan scholars and asks for money to fund his research (image 9).

A first glance at the SoundRiver of this movie reveals that the sound is all the time very rich. The atmosphere in the music changes between mystical, tense and jolly scenes, while the volume is heavily modulated. Most of the time, people are speaking, commonly in a conversation. The composition and the plot are from female to mixed to mainly male. The footsteps in the added sound effect column indicate frequent movement.

In this first example, we can already see very well how important music is to support a certain atmosphere. The different scenes are clearly marked by different music. Whenever the plot is concerned with witches, fighting polar bears or Dust it changes to a mystical style. Arguments are underlined by slow tense music and playing scenes by happy, jolly tunes. Moreover, a strong modulation in the loudness is present, featuring for example a change from slightly tense to dangerous scenes, when Asriel is to be poisoned. Sound effects are used throughout the movie, with footsteps being the most frequent one. The door and footstep signs indicate time intervals when people move around a lot. The same is true for speech, where most of the time conversations between two people takes place.

Plot of the StarWars (©Lucasfilm) sequence: After rescuing the captive Chancellor Palpatine, Obi Wan Kenobi and Anakin Skywalker find themselves stuck on the severely damaged ship, while the crew has fled using all escape capsules. Their only option is to land the damaged ship (images 1-3). After the successful landing, Chancellor Palpatine is welcomed by Jedi Master Mace Windu (image 4). As the party enters the senate, Anakin is finally reunited with his wife Padme Amidala (image 5), who tells him that she is pregnant. Despite Padme’s worries over their secret marriage, Anakin is overjoyed at this news. After fleeing the damaged ship, General Grievous reports to his master, Lord Sidious. At night in their flat, Amidala and Anakin make plans to raise their child (image 7).

The scenery of the StarWars movie features strong changes in volume and atmosphere. Action sequences including a lot of lighting and space shuttle activity alternate with very silent moments between Anakin and Padme. This is clearly represented in the SoundRiver. The music during the action scenes is very loud, many icons represent sound effects and the volume of the speech matches the overall volume. The love sequences on the contrary are much more silent, feature only very few sound effects and whispered conversations. The sound activity during the reception of the
Figure 11: SoundRiver of the beginning of the movie “The Golden Compass”. From left to right, it illustrates the volume and atmosphere of the background music (colored part), audio events and selected key frames (squares on gray background), the occurrence of someone walking or running (footsteps on gray background), volume, gender and number of participants in a conversation (rectangles) and images of selected key frames.

Figure 12: SoundRiver for a sequence from StarWars III - Revenge of the Sith. The different volumes and moods in the music are depicted in the music strand on the left. A futuristic atmosphere is supported by frequent aircraft sounds as illustrated in the sound effect strand in the center. The volume of speech reflects the loudness of the background sound.
Chancellor (23. - 24. minute) is somewhere inbetween, with conversations at a normal level, several sound effects and music that progressively quiets down. The three sceneries are characterized by different moods in music as well: tense and action music for the action scenes, heroic and neutral music for the reception and loving, melancholic and happy music for the scenes with the couple.

7. Evaluation
To evaluate the new technique, we created a video player including the SoundRiver as depicted in Figure 2. Three participants with different degrees of hearing-impairment (mild to deaf) were first shown five minutes of “The Golden Compass” with subtitles and afterwards the same sequence with the additional SoundRiver. The overall feedback was very positive and it was pointed out that the SoundRiver is particularly helpful to understand scenes where audio events cannot be seen, e.g., someone entering through a door that cannot be seen, as this information is often crucial for the plot and not covered in the subtitles. The music strand was rated very helpful by those participants with partial hearing and they could relate SoundRiver to the faintly heard music. For deaf viewers, the music strand is only useful when the conversations at a normal level, several sound effects and music that progressively quiets down. The three sceneries are characterized by different moods in music as well: tense and action music for the action scenes, heroic and neutral music for the reception and loving, melancholic and happy music for the scenes with the couple.

8. Conclusions
We have presented a novel technique, SoundRiver, for illustrating the sound track of a video. In comparison with the traditional sound visualization, SoundRiver is semantically rich, and can thereby provide more meaningful information to the viewers. Its fluidic design, which provides a temporally continuous visual representation, does not require full attention from the viewers, allowing the viewers to pay more attention to moving images in the video. As shown in Figures 11 and 12, SoundRiver provides a noticeable amount of additional information to the key-frame based storyboard, especially brings back the sense of continuousness as one would have when watching a video. It demonstrates an effective application of computer graphics to a real world problem. The future directions of this work include the development of an API to support a larger collection of programmable icons and sound effect models. For the latter, it will be advantageous to carry out the training of a large number of models in an industrial setting.

9. Acknowledgement
We thank Prof. Dr. Daniel Weiskopf, Stuttgart University, for the initial discussion at the early stage of this work.

References

© 2010 The Author(s)
Journal compilation © 2010 The Eurographics Association and Blackwell Publishing Ltd