REIHE INFORMATIK
8/96

Automatic Audio Content Analysis

S. Pfeiffer, S. Fischer und W. Effelsberg
Universität Mannheim
Praktische Informatik IV
L 15, 16
D-68131 Mannheim

Automatic Audio Content Analysis

Silvia Pfeiffer, Stephan Fischer and Wolfgang Effelsberg
Praktische Informatik IV
University of Mannheim
D-68131 Mannheim, Germany
{pfeiffer,fisch,effelsberg}@pi4.informatik.uni-mannheim.de

Abstract

This paper describes the theoretic framework and applications of automatic audio content analysis. After explaining the tools for audio analysis such as analysis of the pitch or the frequency spectrum, we describe new applications which can be developed using the toolset. We discuss content-based segmentation of the audio stream, music analysis and violence detection.

1 Introduction

Looking at multimedia research, the field of automatic content processing of multimedia data becomes more and more important. Automatic cut detection in the video domain [ZKS93, MMZ95, ADHC94], genre recognition [FLE95, ZGST94] or automatic creation of digital video libraries [ZWLS95, SC95] are key topics addressed by researchers.

The MoCA project (Movie Content Analysis) at the University of Mannheim aims at the automatic analysis of streams of video and audio data. We have developed a workbench to support us in this difficult task [LPE96]. First results have been achieved in automatic genre recognition [FLE95]), text recognition [LS96], video abstracting [PLFE96] and audio content analysis.[1]

Humans are well able to recognize the contents of anything seen or heard. Our eyes and ears take in visual and audible stimuli, and our nerves process them. Such processing takes place in different regions of the brain whose exact functions are still not understood in detail.

Research in multimedia content analysis has so far concentrated on the video domain. Few researchers do audio content analysis as well [GLCS95, BC94, Fis94, Smo94]. We demonstrate the strength of automatic audio content analysis. Analogous to the specialized areas that have evolved in the human brain, such analysis merits research in its own right. We therefore explain the algorithms we use, including analysis of amplitude, frequency and pitch, and simulations of human audio perception. We use these algorithms to segment audio data streams into logical units for further processing, and to recognize music as well as sounds indicative of violence like shots, explosions and cries.

This paper is organized as follows. Section 2 describes the basic tools necessary for automatic audio content analysis. Section 3 reports different applications of audio content analysis. Section 4 concludes the paper.

[1] For further information on MoCA see http://www.informatik.uni-mannheim.de/informatik/-pi4/projects/MoCA/

2 Basic properties of audio

2.1 The theoretical model

The content of audio must be regarded from two angles: first there are the measurable properties, from the physics point of view, like amplitude or waveform, and second, the properties of human cognition such as subjective loudness or harmony. These will be presented in the following subsections.

2.2 Physical properties

Sound is defined as an air pressure change which is modelled as a waveform composed of sinus waves of different amplitude, frequency and phase. Experiments with different sounds have shown that the human ear does not differentiate phases, but it is well known that we hear amplitude changes as changes in loudness, and frequency changes as changes in pitch. The phase information is, however, still interesting, when trying to isolate a sound source based on phase differences between both ears. This shows that the human acoustical system analyzes waveforms directly.

More interesting than the waveform itself, however, is often its composition as sinus waves and their amplitudes and frequencies. In physics, this is known as the Fourier Tranformation [Bri74]. The ear also performs such a transformation via a special reception mechanism in the inner ear [Roe79]. It is the basic step in any kind of detailed audio analysis. Only with information on the frequencies can we distinguish between different sounds: every sound we hear is composed of different frequencies and amplitudes whose change pattern is characteristic. The duration of such patterns is the first basic piece of information for partitioning the audio track into "single sounds", which are then classifiable. We will analyze this in more detail in subsection 3.1.

2.3 Psycho-acoustical properties

When a human hears a sound, he/she does not perceive an amplitude and frequencies, but the human auditory system extracts certain desired information from the physical information. The information extracted can be very general like "I hear that somebody is talking" or it can be more accurate like "I hear that Jenny is saying that she is hungry". The sound, however, consists only of the physical information from which it is not easy to derive even general information such as the classification into speech, music, silence or noise, or perceived loudness and dynamics (changes in loudness) from the audio wave.

How does the human accomplish this? Using a computer, we have two methods of simulating the human auditory perception: either we try to model the human auditory system in every detail that is known, or since we know the input data (physical properties of sound) and the ouput data (audio content), we try to make black box models of the processes that happen in the human auditory system and transfer them into programs. Both methods are rewarding. The first one leads to programs which represent our current biological knowledge of the human auditory system. As our knowledge is incomplete, we can only model the derivation of certain basic information (see subsection 2.4). The second method is better for derivation of higher semantic information. If we do not know how a human identifies the sound he/she hears as music, we must wager a guess. Is it a special frequency pattern which he/she has learned to identify as music? How can a computer program model the processes which may occur in a human brain?

Psychoacoustics is the science behind this approach [Roe79]. Researchers in this area have constructed models to derive higher acoustic semantics and have

tested them on people. Some of the theories have also been tested on computers for extraction of higher semantics from digitized sound. We claim that with a knowledge of biology, psychoacoustics, music and physics, we can set up theories on human auditory perception and transfer them into computer programs for evaluation.

An example is the description of loudness as perceived by a human. Different scales have been invented to judge loudness: for example dB-scale, phon-scale, sone-scale. Each measures a different kind of loudness: dB simply measures amplitude differences, phon compares the loudness of different frequencies but of the same amplitude, and sone compares the loudness of different sounds. But when a human expresses that some sound is "loud", this sensation is also dependent on the duration of that sound, the frequency differences present in the sound, that human's "sound history", his visual perception of the sound source, his sensitivity and his expectations (there are probably even more influences).

How can we approach such a problem with a computer program? dB, phon and sone are implemented easily. The impact of the duration of a sound is explained biologically by adaptation of the auditory nerves - this too can be simulated. Involvement of other parameters has to be discussed because some are very subjective (like that human's sensitivity) or are not extractable from the audio alone (like the visual perception of the sound source). "Sound history" or the human's expectations can perhaps be modelled in more detail. For "sound history" we could use a profile of the loudness the human has perceived in the past (for example during the last 2 min) and the human's expecations can perhaps be derived from the environment, e.g. that when going to a disco, he/she expects music of a certain loudness. A kind of "intersubjective" loudness measure will result from such concepts which can surpass those available so far.

2.4 Biological aspects

Multimedia data can be analyzed in two ways: first, characteristic patterns can be extracted and used for classification. This is done without any regard as to how humans perceive the contents of the data. Second, the extraction can be done by simulating the human perception process. This will be described here.

The major difference between data analysis with and without perception simulation is the use of a special filter. As a perception-independent solution directly analyzes frequencies, for example those produced by a Fourier transformation, freqencies are filtered first in a perception-simulating analysis. The filter hereby computes the response a specific nerve cell of the auditory nerve will produce. This response is frequency-dependent. We use the phase-compensated gammatone filter g_c proposed by [Coo93] to transform the frequency signal.

$$g_c(t) = (t_c + t)^{(n-1)} exp(-2\pi b(t + t_c)) cos(2\pi f_0 t)$$

The filter is a fourth-order filter($n = 4$) where b is related to bandwith, f_0 is the center frequency and t_c is a phase- correction constant. The center frequency is the frequency to which the nerve cell is tuned. We use a filter bank of 256 different filters spaced equally on the frequency scale.

Figure 1 shows three of these filters. The higher the frequency, the more the filter oscillates. Taking the output of a specific filter, the probability of a cell to fire can be calculated using the Meddis hair-cell model [Med86]. The signal transformed into nerve-cell response probabilities can now be used to calculate two important indicators for classifying audio content:

- Onset and offset which are a measure of how fast a cell responds to a signal. These indicators are a measure of how fast a signal changes.

- Frequency transitions which describe glides in frequency over time.

Figure 2 shows an onset plot for a cry and for a shot. The shot's onset is much higher than that of the cry.

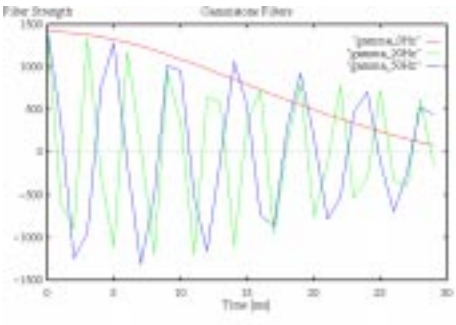

Figure 1: Gammatone Filters

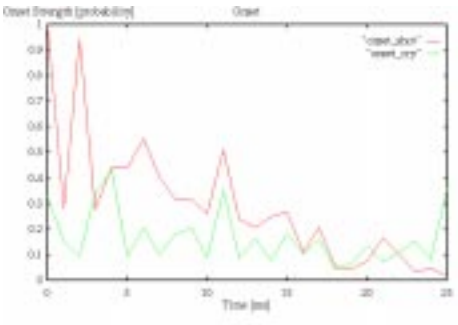

Figure 2: Onset

Frequency-transition maps are calculated using a direction-selective filter, for example the second derivative of a normal distribution rotated by an angle α. This filter is convolved with the response of the Meddis hair-cell model and describes glides in frequency over time as perceived by humans. For further details see [BC94].

We have implemented all theoretical constructs that we explained in this section. We developed algorithms in C and C++ on a Unix Workstation to perform

- a Fourier transform,
- an analysis of waveforms,
- an analysis of the frequency spectrum,
- an analysis of fundamental frequencies,
- a calculation of onset and offset and
- a calculation of frequency transitions.

These algorithms serve us as tools for further audio content analysis. They are part of the MoCA workbench. It is our goal to combine these tools to create new applications. This will be described in the next section.

3 Applications and Experimental Results

3.1 Content-based segmentation

In order to retrieve content of audio, it is necessary to first structure the audio. This is similar to determining content in still images: a decent object segmentation is the basis. The structure of audio can be manifold: a first classification should distinguish music, speech, silence, and other sound sequences, because handling of content is fundamentally different for each of these classes. A second segmentation step could result in determining syllable, word or sentence boundaries for speech, or note, bar or theme boundaries for music. Other sounds, i.e. any kind of environmental sounds that a human may encounter, may be classified, too. In subsections 3.2 and 3.3 we go into more detail on classification of the content of music and of a specific environmental sound class: sounds indicating violence.

How can the general classification into silence, speech, music and other sound be achieved? A human determines silence on a relative scale: a loudness of 0 dB is not very common in any natural environment, let alone in digitized sound. Therefore, an automatic recognition of silence must be based on comparison of loudness levels along a timeline and an adapting threshold. In that way, silence can be distinguished from other sound classes.[2]

Speech and music are distinguishable simply by the spectrum that they cover: speech lies in the area between 100 and 7000 Hz and music between about 16 and 16000 Hz. Unfortunately, the latter also applies to environmental sounds ("noise"). Therefore, a distinction between music and other sounds was made by analyzing the spectrum for orderliness: tones and their characteristic overtone pattern do not appear in environmental sounds, neither is a rhythmic pattern present there.

A segmentation of audio must be performed based on the recognition of acoustical content both in the temporal and the frequency domains. For example, an analysis of amplitude (loudness) statistics belongs to the temporal domain whereas an analysis of pitch or frequency patterns belongs to the frequency domain.

Other work is based on amplitude statistics. One psychoacoustic model presented a segmentation of speech signals based on amplitude statistics [Sch77] and was able to describe speech rhythm and to extract syllables.

The recognition of music and speech has already been a goal in electrical engineering research: Schulze [Sch85] tried to separate them on the basis of amplitude statistics alone. His goal was to determine the signal dynamics in view of the restricted transmission capacity of radio channels. He found out that the spectrally split up amplitude distribution changes over the years because of changing production and listening habits. He therefore used a distribution-density function of the amplitude statistics. This function needed a few seconds to reach the necessary stationariness of the signals, but could then distinguish music and speech.

Köhlmann [Köh84] presented a psychoacoustic model for distinction between music and speech based on a rhythmic segmentation of the sound. He used loudness and pitch characteristics to determine event points (a rhythmic pattern) and found that the metric structure of a sound sequence was already sufficient to determine whether the sound was speech or music.

We have performed experiments in distinguishing silence, speech, music and noise [Ger96]. Our prototype uses characteristic tone-color vectors to define the classes and a comparison of tone-color vectors with an adapting difference threshold to decide upon the classification. Tone-color vectors are defined according to the psychoacoustical literature (see [Ben78]). For our special examples, we have found good characteristic tone-color vectors. An example for a distiction between a speech and a music passage is shown in Figures 3 and 4: the first shows the wave pattern of the analyzed audio piece and the second the difference computation where a zero value implies a segmentation point.

3.2 Music Analysis

Human music cognition is based on the analysis of temporal and frequency patterns, just like any other human sound analysis.

The analysis of temporal structure can be based on amplitude statistics. We have used amplitude statistics to derive the beat in modern music pieces. While an amplitude analysis may be a first step towards the temporal analysis of audio, it does not suffice: spectrum analysis is necessary, too. For example, a segmentation of musical harmony (chords) can be performed by analyzing the spectrum and

[2] Such silence detection is easily exploited for surveillance of rooms. A vault room, for example, may be supervized less noticeably by several microphones than by cameras.

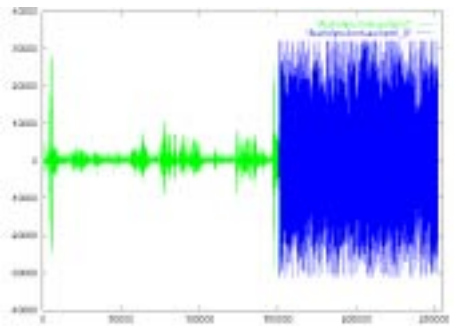

Figure 3: Waveform of file youtook.au

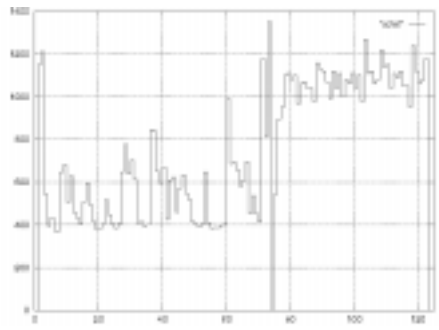

Figure 4: Distance diagram of file youtook.au

retrieving any regularities. Because typical music consists of a series of chords which are frequently changed, the chords are visible in the spectrum as a group of frequencies being simultaneously present for a longer time. In that way, we get a segmentation of music into entities similar to written music. Based on this segmentation, we can perform a fundamental frequency (fuf) determination on the chords.

The sequence of fuf's of a piece of music is very important for the human attribution of content to a piece of music: it determines the perception of melody and is one of the parameters most important to determining the structure of a piece of music.

Human fuf perception is not trivial. A human hears the fuf of a sound even though the fuf itself might not be present. For example, the fuf of an adult male voice lies at about 120 Hz, that of an adult female voice at about 220 Hz. When voice is transmitted via a common telephone line, only the frequencies between 300 and 3400 Hz are transmitted (the lower boundary results from signal-distortion restrictions and the upper boundary from signal resolution). We hear the restricted quality of the speech signal, but we don't realize that the fuf is lacking because our auditory system completes this missing frequency from the rest of the heard frequencies.

The same effect occurs when listening to music on a cheap transistor radio: because of the small loudspeakers, frequencies below 150 Hz are not played. The low frequencies are perceived nevertheless.

The fuf results from overlying the higher frequencies. For example, if two frequencies f_1, f_2 are played, which are a musical fifth apart from each other, the frequency f_0 of the resulting sound is calculated as follows:

$f_2 = \frac{3}{2} f_1$ (i.e. f_2 is a fifth above f_1)

$f_0 = \frac{1}{2}$

Looking at the frequency diagram in figure 5, it can be seen that the period belonging to the fuf is the smallest common multiple of the periods of the frequencies it consists of. Table 1 shows this result for different intervals.

This result can now be used to determine the fuf of a musical chord by a program. It works for intervals, notes with harmonic overtones and harmonic chords.

1. Determine the lowest frequency appearing in the spectrum with an amplitude above a certain threshold, called f_1.

2. Check whether a frequency a fifth, fourth, major or minor third above f_1 appears in the sound: $f_x = \frac{I+1}{I} f_1, \ for \ I = 2, .., 5.$

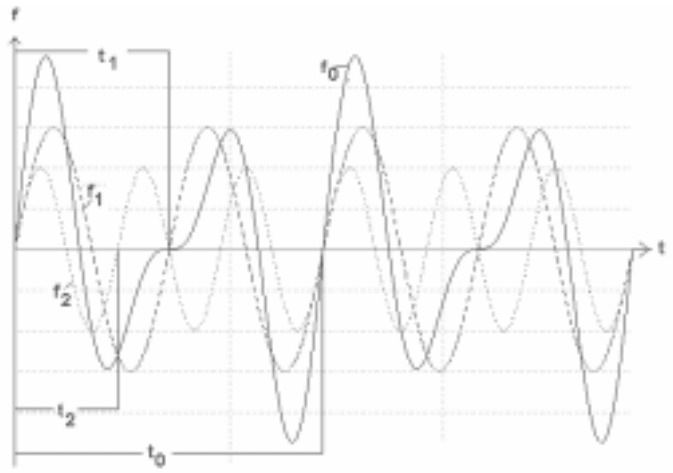

Figure 5: Overlying frequencies f_1 and f_2

Interval	frequency relation		fundamental frequency	
Fifth	$f_2 = \frac{3}{2}f_1$ I=2	$f_2 = \frac{I+1}{I}f_1$	$f_0 = \frac{1}{2}f_1$	$f_0 = \frac{1}{I}f_1$
Fourth	$f_2 = \frac{4}{3}f_1$ I=3		$f_0 = \frac{1}{3}f_1$	
Major Third	$f_2 = \frac{5}{4}f_1$ I=4		$f_0 = \frac{1}{4}f_1$	
Minor Third	$f_2 = \frac{6}{5}f_1$ I=5		$f_0 = \frac{1}{5}f_1$	

Table 1: Correlation between intervals and their perceived fundamental frequency

3. If yes, choose $f_0 = \frac{1}{I}f_1$ as fuf.

4. Otherwise, choose f_1 as the fundamental frequency.

The compression of a music piece into a sequence of fuf's is a means to produce a *characteristic signature* of music pieces. Such a signature can be used for audio retrieval, where music must be recognized and longlasting pattern recognition processes are not acceptable.

We see an example in advertising analysis: having a multimedia database, we store all TV commercials, including the video and audio tracks in digital format, together with the respective product name. Most commercials contain an identifying melody on which we perform our fuf-recognition algorithm. These results are also stored in the database. Now, we are interested to know, how often a specific commercial is run in a certain time period on all channels. Provided that all our commercials contain the identifying melody, we simply record all commercials from all channels (commercial recognition and segmentation is easily performed on the picture track [LS96]), digitize them and perform the fuf recognition on the audio tracks. Then, we compare the resulting fuf sequences with the fuf sequences stored in the database. One title would have a significantly higher correlation to the queried piece such that we could automatically decide on the corresponding product name. If there is no such title, we have run across a "new"commercial, i.e. one which is not yet part of the database, and will add it (see figure 6).

We have experimented with the retrieval of music titles based on the fuf recognition and compared it to retrieval based on amplitude or frequency characteristics. Our prototype database consisted of only 17 pieces of digitized music, but included different kinds of music, like classical and pop music. We tested the retrieval against

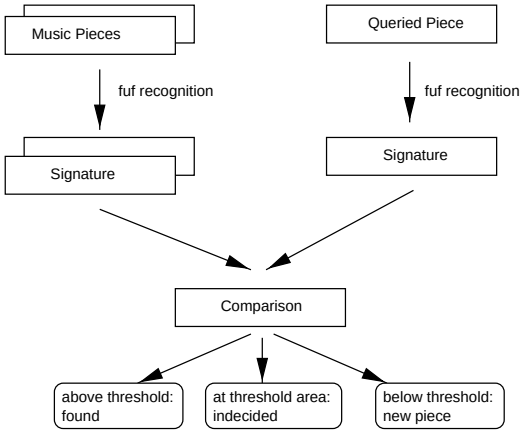

Figure 6: Retrieval of commercials

different digitization qualities, different music lengths and different musicians playing the same piece. Results for different music length can be seen in Figures 7 and 8 for two music pieces [Höf96]. As can be seen, retrieval based on frequency or fuf statsitics gives much better results than retrieval based on amplitude statistics. As we only worked on 8000 Hz sampled audio pieces, the frequency resolution resulting from Fourier Transform is not very detailed and therefore the fuf recognition not very good. This will be changed in the future.

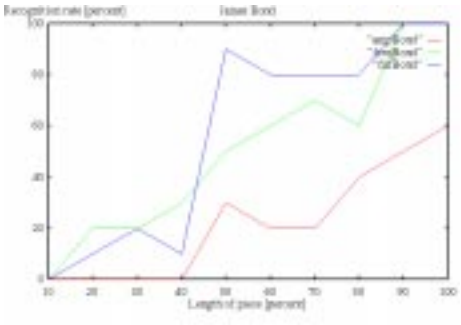

Figure 7: Comparison of recognition rates for James Bond title music

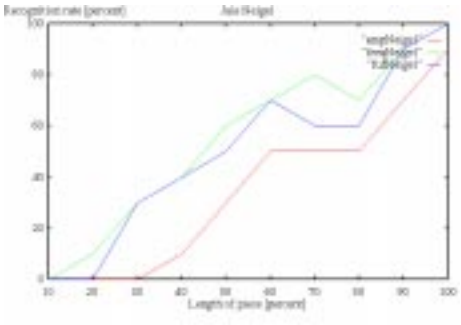

Figure 8: Comparison of recognition rates for a piece by Jule Neigel

3.3 Violence Detection

Automatic violence detection will be described next. Violence in movies can have a bad influence on children which is why movies are rated. Although a computer system will never be able to rate movies in a fully automated fashion, it can assist in the process. Movie sequences which contain violence could be cut out via such a computer-aided film-rating system.

As violence itself contains many aspects and is strongly dependent on the cultural environment, a computer system cannot recognize violence in all its forms. It is most unlikely that a computer would be able to recognize mental violence. It is not our goal to recognize every form of violence, we concentrate on the recognition of a few forms of violence to start to explore this field.

A variety of sounds exist which indicate violence and which are independent of the cultural environment of the user: among them shots, explosions and cries.

The algorithm we propose for their recognition is the following:

1. Compute for each ms amplitude, frequency, pitch, onset, offset and frequency-transition maps statistics of a window of 30 ms of the audio file to be tested.
2. Compare these statistics with signatures of explosions, cries and shots calculated earlier and stored on disk. The comparison can be made either by using the correlation of the two patterns or the Euclidean distance of both patterns.
3. If a similarity between test pattern and stored pattern is found the event is recognized.

Statistics represent only the mean values of the time period examined. To be able to examine changes of the test pattern in time we compare the test pattern with several stored patterns. We store the mean statistics for the entire event: the beginning, the end and that time window which contains the greatest change. The amount of change is hereby determined by the variance. The correlation between 30-ms test patterns and stored patterns of a few seconds length but of the same event type is still very good.

In our experiments we extracted shots, explosions and cries out of audio tracks manually and stored the calculated signature of the whole event on disk.
We then tried to locate these events in the same tracks. Therefore a 30-ms audio track test pattern was calculated and compared with the stored pattern, the time window was incremented by 2 ms and the process repeated until the end of the audio track. The question was if the correlation between the test patterns and the much longer stored pattern was high enough to be able to recognize the event. The correlation between the 30-ms test patterns and the stored pattern in all of the 20 tests exceeded 90 percent. Our test-data set therefore contains four test sets for each event and several sets of the same event. The database currently contains data on 20 cries, 18 shots and 15 explosions.

For every indicator (loudness, frequency, pitch, onset, offset, frequency transitions), we compute minimum, maximum, mean, variance and median statistics. In our experience a linear combination of minimum, maximum, mean, variance and median yields the best results. The weights for such a combination cannot be equal as the correlation is different. Obviously in most cases the correlation between mean and variance is higher than that between mean and maximum. The weights we determined heuristically are shown in Table 2.

Statistical Elements					
Maximum	Minimum	Mean	Variance	Median	$\sum$
33.33	3.33	33.33	20	10	100

Table 2: Weights of statistical instruments

Figures 9 and 10 show plots of frequency transitions for a cry and for a shot. It is evident that these two events can already be distinguished on the basis of this indicator alone.

As the indicators do not have the same importance for the recognition process we also use different weights to outline their importance. These weights differ from event to event (see table 3). Using these weights we are able to calculate a mean correlation between test pattern and stored pattern.

To be able to recognize an event we defined three decision areas. If the correlation of the two patterns is below 60 percent, we reject, if it is beween 60 and 85

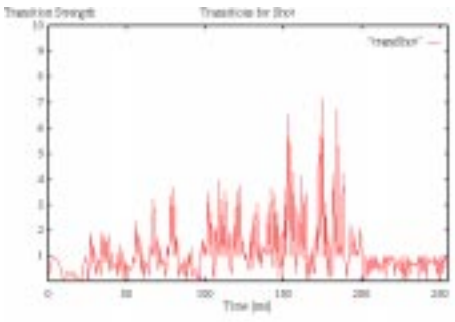

Figure 9: Freqency transition for shot

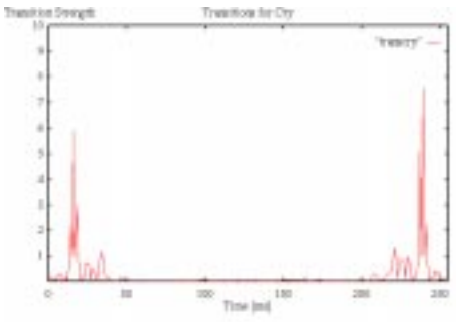

Figure 10: Freqency transition for cry

Indicator	Event		
	Shot	Cry	Explosion
Loudness	10	5	11
Frequency	30	42	27
Pitch	12	21	17
Onset	27	8	26
Offset	9	11	2
Frequency Transition Map	12	13	17
$\sum$	100	100	100

Table 3: Weights of indicators

percent we are undecided, and if the correlation is above 85 percent we accept that the test pattern and the stored pattern are identical.

Our experiment series contained a total of 80 tests. The series contained 27 files which did not contain cries, shots or explosions. Test results are shown in Table 4.

Event	Results in percent			$\sum$
	correctly classified	no recognition possible	falsely classified	
Shot	81	10	9	100
Cry	51	32	17	100
Explosion	93	7	0	100

Table 4: Classification Result

The percentage of correctly classified events is not very high for cries. An important detail of the classification is the very low percentage of falsely classified events. A possibility to avoid uncertain decisions is either to ask the user if the movie part should be shown or not to show at all a part which might contain violence.

4 Conclusion

In this paper, we have described algorithms to analyze the contents of audio automatically. Information on amplitude, frequency, pitch, onset, offset and frequency transitions can be used to classify the contents of audio. We distinguish between algorithms simulating the human perception process and those seeking direct relations between the physical properties of an audio signal and its content.

Further we showed exemplary applications we have developed to classify audio content. These include segmentation of audio into logical units, detection of violence and analysis of music.

We strive to develop more new algorithms to extract information from audio-data streams. These include harmony analysis as well as instruments for tone analysis.

Our efforts in the field of music analysis focus on the distinction of different music styles like pop music and classical music.

References

[ADHC94] Farshid Arman, R. Depommier, Arding Hsu, and Ming-Yee Chiu. Content-based browsing of video sequences. In *Proceedings of Second ACM International Conference on Multimedia*, pages 97–103, Anaheim, CA, October 1994.

[BC94] Guy J Brown and Martin Cooke. Computational auditory scene analysis. *Computer Speech and Language*, (8):297–336, August 1994.

[Ben78] Kurt Benedini. *Psychoacoustic Measurments of the Similarity of Tone Colors of Harmonic Sounds and Description of the Connection between Amplitude Spectrum and Tone Color in a Model.* PhD thesis, Technische Universität München, 1978. (in German).

[Bri74] E. O. Brigham. *The Fast Fourier Transform.* Prentice-Hall Inc., 1974.

[Coo93] M.P. Cooke. *Modelling Auditory Processing and Organisation.* Cambridge University Press, 1993.

[Fis94] Alon Fishbach. Primary segmentation of auditory scenes. In *Intl. Conf. on Pattern Recognition ICPR*, pages 113–117, 1994.

[FLE95] Stephan Fischer, Rainer Lienhart, and Wolfgang Effelsberg. Automatic recognition of film genres. In *Proceedings of Third ACM International Conference on Multimedia*, pages 295–304, Anaheim, CA, November 1995.

[Ger96] Christoph Gerum. Automatic recognition of audio-cuts. Master's thesis, University of Mannheim, Germany, January 1996. (in German).

[GLCS95] A. Ghias, J. Logan, D. Chamberlain, and B.C. Smith. Query by humming: Musical information retrieval in an audio database. In *Proceedings of Third ACM International Conference on Multimedia*, pages 231–236, Anaheim, CA, November 1995.

[Höf96] Alice Höffl. Automatic indexing of digital audio. Master's thesis, University of Mannheim, January 1996. (in German).

[Köh84] Michael Köhlmann. *Rhythmic Segmentation of Sound Signals and their Application to the Analysis of Speech and Music.* PhD thesis, Technische Universität München, 1984. (in German).

[LPE96] R. Lienhart, S. Pfeiffer, and W. Effelsberg. The MoCA workbench: Support for creativity in movie content analysis. In *Conference on Multimedia Computing & Systems*, Hieroshima, Japan, June 1996. IEEE. (to appear).

[LS96] R. Lienhart and F. Stuber. Automatic text recognition in digital videos. In *Image and Video Processing IV, Proc. SPIE 2666-20*, 1996.

[Med86] R. Meddis. Simulation of mechanical to neural transduction in the auditory receptor. *Journal of the Acoustical Society of America*, (34):702–711, 1986.

[MMZ95] K. Mai, J. Miller, and R. Zabih. A feature-based algorithm for detecting and classifying scene breaks. In *Proceedings of Third ACM International Conference on Multimedia*, pages 189–200, Anaheim, CA, November 1995.

[PLFE96] S. Pfeiffer, R. Lienhart, S. Fischer, and W. Effelsberg. Abstracting digital movies automatically. Technical Report TR-96-005, University of Mannheim, April 1996.

[Roe79] J.G. Roederer. *Introduction to the Physics and Psychophysics of Music.* Springer, New York, 1979.

[SC95] M.A. Smith and M. Christel. Automating the creation of a digital video library. In *Proceedings of Third ACM International Conference on Multimedia*, pages 357–358, Anaheim, CA, November 1995.

[Sch77] Hermann Schütte. *Determination of the subjective event times of subsequent sound impulses via psychoacoustic measurements.* PhD thesis, Technische Universität München, 1977. (in German).

[Sch85] Klaus Schulze. *Contribution to the Problem of Eon-Dimensional Amplitude Statistics of Tone Signals with the Attempt to produce a Model and to separate Speech from Music based on Statistic Parameters*, volume 11 of *Fortschritt-Berichte VDI.* VDI-Verlag GmbH, Düsseldorf, 1985. (in German).

[Smo94] Stephen W. Smoliar. In search of musical events. In *Intl. Conf. on Pattern Recognition*, pages 118–122, 1994.

[ZGST94] HongJiang Zhang, Yihong Gong, Stephen W. Smoliar, and Shuang Yeo Tan. Automatic Parsing of News Video. In *Proceedings of IEEE Conf. on Multimedia Computing and Systems.* IEEE, May 1994.

[ZKS93] HongJiang Zhang, Atreyi Kankanhalli, and Stephen W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1(1):10–28, January 1993.

[ZWLS95] HongJiang Zhang, J.H. Wu, C.Y. Low, and S.W. Smoliar. A video parsing, indexing and retrieval system. In *Proceedings of Third ACM International Conference on Multimedia*, pages 359–360, Anaheim, CA, November 1995.