Automatic Audio Content Analysis

Silvia Pfeiffer, Stephan Fischer and Wolfgang Effelsberg

University of Mannheim
Praktische Informatik IV
D-68131 Mannheim, Germany
{pfeiffer,fisch,effelsberg}@pi4.informatik.uni-mannheim.de
phone: +49-621-292-5054 fax: +49-621-292-5745

ABSTRACT
This paper describes the theoretic framework and applications of automatic audio content analysis. After explaining the basic properties of audio analysis, we present a toolbox being the basis for the development of audio analysis algorithms. We also describe new applications which can be developed using the toolset, among them music indexing and retrieval as well as violence detection in the sound track of videos.

KEYWORDS *audio content analysis, audio toolbox, audio segmentation, audio indexing and retrieval, 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 [34, 18, 1], genre recognition [8, 33] or automatic creation of digital video libraries [35, 29] 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 [16]. First results have been achieved in automatic genre recognition [8], text recognition in videos [17], video abstracting [22] and audio content analysis.[1]

Research in multimedia content analysis has so far concentrated on the video domain. Few researchers do audio content analysis as well [12, 4, 9, 30]. We are convinced that audio content processing is as important as video content processing. Humans use both eyes and ears to understand contents. Why not do so in processing content by computer?

In this article we describe the tools of our audio content processing toolbox as well as the applications we have developed using the toolbox. Although many audio tools can be found on the Internet, a system as a collection to build new applications has never been reported. As a part of the MoCA workbench, the toolbox is developed as a set of algorithms which can be easily combined to create new applications. In this paper we describe not only the standard algorithms, but also new, more complex algorithms we have developed for automatic audio content processing.

Applications include a music indexing and retrieval system and a violence detection system. Many researchers concentrate on the creation of efficient image retrieval systems [35], very few on the creation of audio retrieval systems [12]. Using our toolbox, we explain in detail a general music retrieval system. We also introduce our computer-assisted violence detection system. It could perhaps one day serve to protect children from rated movies containing violence segments. We explain initial steps towards this challenging goal, among them the automatic recognition of shots, cries and explosions.

This paper is organized as follows. Section 2 describes basic approaches to audio analysis. Section 3 describes our toolset of basic operators for automatic audio content analysis. Section 4 reports on different applications of audio content analysis and Section 5 concludes the paper.

2 BASIC PROPERTIES OF AUDIO

The content of audio must be regarded from two angles: first with regard to measurable properties, from the point of view of physics, e.g. amplitude or waveform, and second, with regard to properties of human cognition such as subjective loudness or harmony.

2.1 Physical Properties

Sound is defined as a change in air pressure which is modelled as a waveform composed of sinusoidal waves of different amplitude, frequency and phase. Experiments with different sounds have shown that the human ear does not differentiate

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

ACM Multimedia 96, Boston MA USA
© 1996 ACM 0-89791-871-1/96/11 ..$3.50

phases, but it is well known that we hear amplitude changes as changes in loudness, and frequency changes as changes in pitch [24]. The phase information is, however, still interesting, e.g. when trying to locate a sound source on the basis of phase differences between both ears. This proves that the human acoustical system analyzes waveforms directly.

More interesting than the waveform itself, however, is often its composition of sinusoidal waves and their amplitudes and frequencies. In physics, this is known as the Fourier Tranform (a fast algorithm to implement the Fourier Transform on computer systems is known as Fast Fourier Transform FFT) [6, 3]. The ear also performs such a transformation via a special reception mechanism in the inner ear [24]. It is the basic step in any kind of detailed audio analysis. Only when we possess 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 can then be classified. We will analyze this in more detail in Subsection 3.3.1.

2.2 Psycho-acoustical Properties

Upon hearing a sound, humans do 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 precise, like "I hear that Jenny is saying that she is hungry". The sound, however, consists only of the physical information. It is surprisingly difficult to automatically 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 do humans accomplish this? Using a computer, we have two methods of simulating 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 occuring in the human auditory system and transfer them into programs. Both methods are rewarding and have been used by researchers, though most prefer one or the other.

The first method 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 3.2).

The second method is better suited to derive higher semantic information. If we do not know how a human identifies as music a sound heard, we must wager a guess. Is it a special frequency pattern that has been learned to be identified as music? How can a computer program model the processes which may occur in a human brain? Psychoacoustics is the science behind this second approach [24, 5, 32]. Researchers in this area have constructed models to derive higher acoustic semantics and have tested them on humans [14, 21, 27, 28]. Some of the theories have also been tested on computers in order to extract higher semantics from digitized sound.

We claim that given 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 [10, 20, 31]. Each measures a different kind of loudness: dB simply measures amplitude differences, phon compares the loudness of different frequencies that are 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 [37], the frequency differences present in the sound [36], that human's "sound history", his visual perception of the sound source, his sensitivity and his expectations (and probably on 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 as the 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 minutes) and the human's expectations can perhaps be derived from the environment. For example when going to a disco, music of a certain loudness is expected. A kind of "intersubjective" loudness measure will result from such concepts which can surpass those currently available.

3 THE AUDIO OPERATORS TOOLBOX

3.1 General Outline

In Section 2, we have described the basic kinds of approaches to analyze digital audio, mentioning there some common operators of digital audio analysis such as the Fast Fourier Transform (FFT). Having decided to produce a toolbox containing such basic operators, we developed algorithms in C and C++ on a Unix workstation.[2] It is our goal to combine these tools to create new applications. The indicators the toolbox contains so far are

- Volume analysis,
- Frequency analysis,
- Pitch analysis,
- Onset and offset ,

[2] The toolbox is part of the MoCA workbench, via which audio operators are easily combined with picture analysis algorithms.

- Frequency transition maps,
- Audio segmentation,
- Fundamental frequency analysis and
- Beat analysis.

The first three are described in the literature [6], so we omit an explanation here. The other, more sophisticated and uncommon algorithms are described in the following subsections.

3.2 Biological Operators

The major difference between data analysis with and without perception simulation is the use of a special filter. A perception-independent solution directly analyzes frequencies, for example those produced by a Fourier Transform. Therefore, frequencies 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 [7] 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.

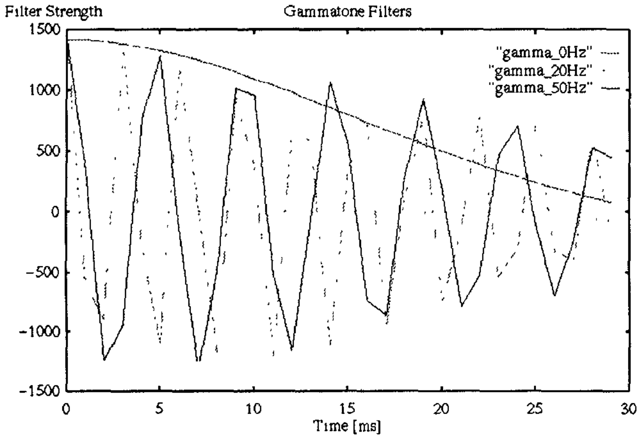

Figure 1: Gammatone Filters

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 [19]. 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.

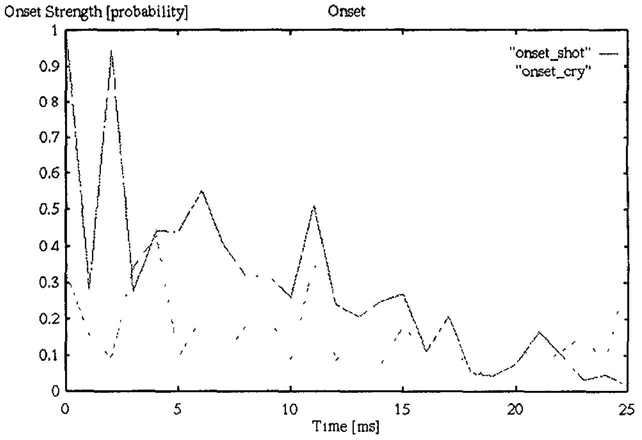

Figure 2: Onset

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.

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 [4].

3.3 Physical Analysis Operators

3.3.1 Content-based Segmentation

In order to recognize the contents of audio, it is necessary to first structure the audio stream. This is similar to determining content in still images: successful object segmentation is the basis for further processing. Our first step in content-based audio segmentation is to distinguish between music, speech, silence and other sound sequences, because handling of content differs fundamentally for each of these. For example, if an arbitrary piece of audio is found to be speech, speech recognition and speaker recognition can be performed on it. If it is found to be music, note, bar or theme boundaries may be extracted, and fundamental frequencies can be determined.

How can the general classification into silence, speech, music and other sounds be achieved? First, we partition the audio stream into similar segments. This is performed both in the temporal and in the frequency domain. In the temporal domain, we produce amplitude (loudness) statistics (similar to [26]) and in the frequency domain, statistics of patterns in frequency bands (tone-color, similar to [2]). These aid us in deciding about the similarity of subsequently analyzed sample groups, which leads to the segments.

The second step is the classification of the segments into speech, music, silence and other sounds. How do we perform this? Humans determine 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 with an adaptive threshold. In that way, silence can be distinguished from other sound classes.[3]

Speech and music are distinguishable simply by the spectrum that they cover: speech frequencies lie in the range of 100 to 7000 Hz, and music frequencies between about 16 and 16000 Hz. Unfortunately, the latter also applies to environmental sounds ("noise"). Therefore, our idea is to distinguish between music and other sounds 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.

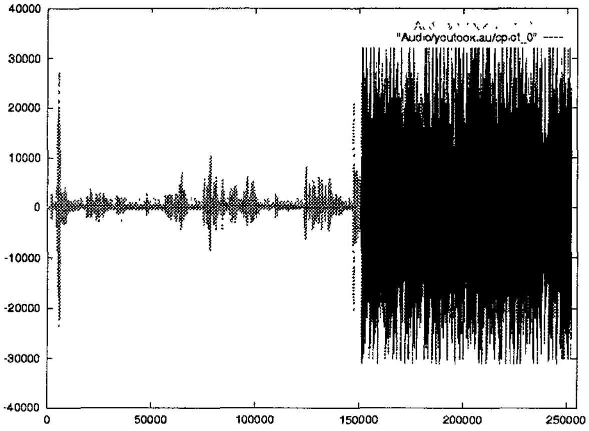

Figure 3: Waveform of file youtook.au

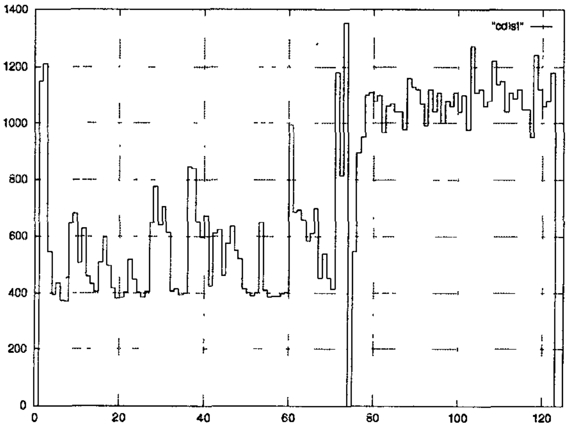

Figure 4: Distance diagram of file youtook.au

We have performed experiments to distinguish speech, music, silence and noise [11]. An example of a distinction 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 values of tone-color where a zero value implies a segmentation point.

Work towards similar aims has been performed before. [26] tried to separate speech and music signals based on amplitude statistics in the time domain alone. [14] presented a psychoacoustic model to distinguish between speech and music based on loudness and pitch characteristics which determined a metric structure. None of them tried, as we did, to classify audio streams completely.

3.3.2 Music Operators

Music is characterized by a temporal structure and a note (pitch and overtones) structure. The analysis of temporal structure is based on amplitude statistics. We have used amplitude statistics to derive the beat in modern disco music [25]. 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 retrieving 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 simultaneously present for a longer time. This approach yields 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 as a first step toward note analysis. The sequence of fuf's in 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 in determining the structure of a piece of music.

Human fuf perception is not trivial [24, 23]. A human is able to hear 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 itself is missing because our auditory system completes this 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 overlaying the higher frequencies. For

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

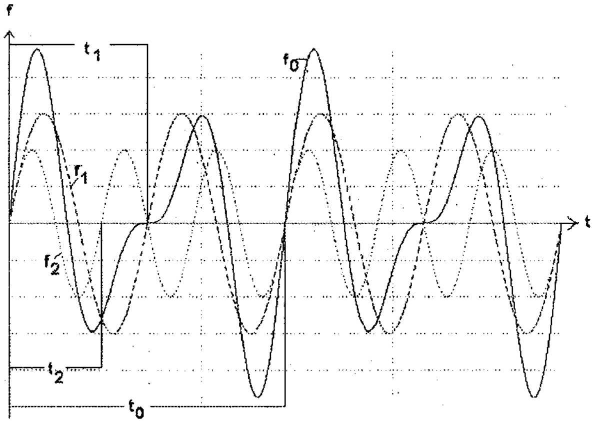

Figure 5: Overlaying frequencies f_1 and f_2

Interval	frequency relation			fundamental frequency	
Fifth	$f_2 = \frac{3}{2}f_1$	$f_2 = \frac{I+1}{I}f_1$	I=2	$f_0 = \frac{1}{2}f_1$	$f_0 = \frac{1}{I}f_1$
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

example, if two frequencies f_1, f_2 are played, which are one 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 one fifth above f_1),

$f_0 = \frac{1}{2}f_1$ (i.e. f_0 is one octave below f_1).

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 of which it consists. Table 1 shows this result for different intervals.

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

Algorithm fuf:

1. Determine the lowest (significant) frequency appearing in the spectrum. Call it f_1.
2. Check, whether a (significant) frequency one fifth, fourth, major or minor third above f_1 appears in the sound:
   $f_x = \frac{I+1}{I}f_1,\ for\ I = 2,..,5.$
3. If yes, choose $f_0 = \frac{1}{I}f_1$ as fuf.
4. Otherwise, choose f_1 as the fundamental frequency.

4 APPLICATIONS AND EXPERIMENTAL RESULTS

Having explained some of the operators contained in our audio toolbox, we now proceed to the presentation of two applications which have been implemented using the toolbox.

4.1 Music Indexing and Retrieval

The compression of a piece of music into a sequence of fundamental frequencies (fufs) 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.

An example is advertisement analysis (see figure 6): having a multimedia database, we store all TV commercials, including the video and audio tracks in digital format, together with other information such as the respective product name. Most commercials contain an identifying melody on which we perform our fuf-recognition algorithm. The resulting fuf signatures 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 on TV. Provided that all examined commercials contain an identifying melody, we simply record all commercials from all channels, digitize them, extract each single commercial (automatic commercial recognition and segmentation is easily performed on the picture track [15]), and determine the music parts by use of our audio segmentation algorithm (see 3.3.1). On the music parts, we then perform the fuf recognition (see

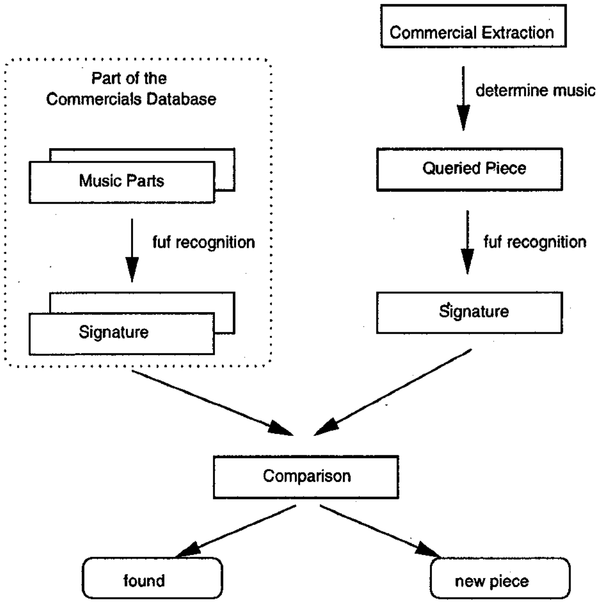

Figure 6: Retrieval of commercials

3.3.2), resulting in the characteristic fuf sequence, which is then compared with the fuf sequences stored in the database. If the queried commercial is already in the database, the respective entry has a significantly higher correlation to the queried piece and we can automatically count its appearances. If the queried commercial is new, i.e. one not yet part of the database, there is no such title with a significantly higher correlation.

We have experimented with the retrieval of music titles based on the fuf recognition and compared it to retrieval based on frequency characteristics [13]. 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. We developed 10 FFT analysis indicators without use of windows, 10 FFT analysis indicators using a Hanning window and 10 fuf analysis indicators. These three classes of indicators are evaluated separately. Stored in the database are 30 indicators per entry. These 30 indicators are also calculated for a queried piece and then compared to each respective indicator of the pieces in the database, resulting in a similarity percentage. The highest similarity percentage determines the entry that is identified by a single indicator. Then, the indicator results are accumulated within the four classes. The piece with the highest similarity percentage and the highest hit rate is determined to be the retrieved piece of the class.

Our current prototype database consists of 100 pieces of digitized commercials. The results of an experiment with 27 newly recorded and digitized pieces, 17 of which were already in the database (different digitization), can be seen in Figures 7 and 8. The 17 pieces were always retrieved correctly by the three classes, i.e. the hit rate was 100%. For new pieces, either all three classes retrieved different pieces or the average similarity percentages of the retrieved pieces were very low. Therefore, a good retrieval decision with this system is based on the three classes and the two following conditions:

1. All three classes must determine the same piece as "winner".
2. The average similarity percentage must be above 85% in all three classes.

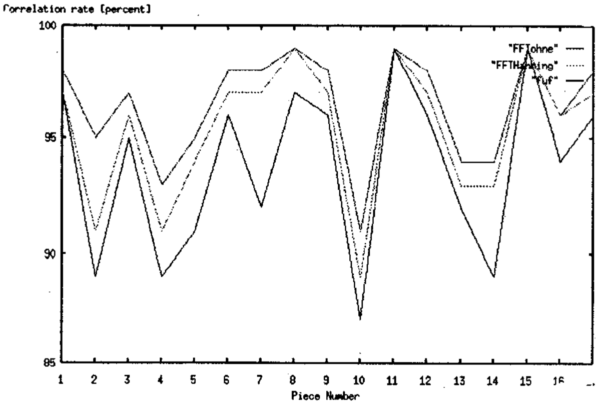

Figure 7: Recognition rates for 17 retrieved pieces

The experimental music indexing and retrieval system shows that we cannot yet produce a characteristic signature based

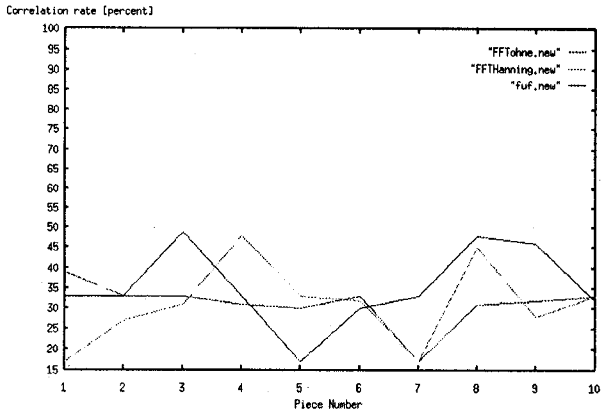

Figure 8: Recognition rates for 10 new pieces

solely on fuf indicators. A combination with FFT indicators is more reliable at the moment. We are investigating further into more reliable fuf signatures with better frequency resolution.

4.2 Violence Detection

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 that 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 as an initial step into this field.

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

We extracted shots, explosions and cries out of audio tracks manually and stored the calculated signature of the events. We then tried to locate these events in the same tracks. 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 whether 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.

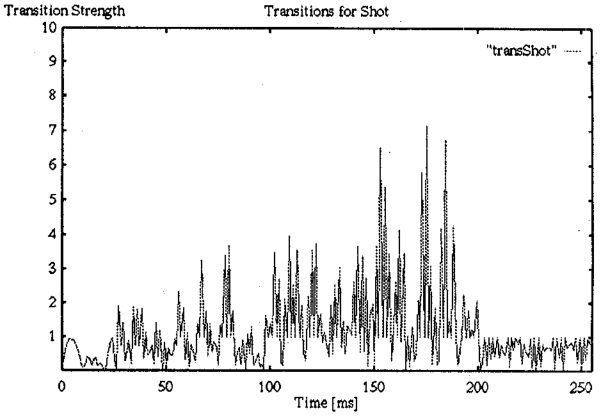

Figure 9: Freqency transition for shot

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.

Statistical Elements					Σ
Maximum	Minimum	Mean	Variance	Median	
33.33	3.33	33.33	20	10	100

Table 2: Weights of statistical instruments

Event	Results in percent			Σ
	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

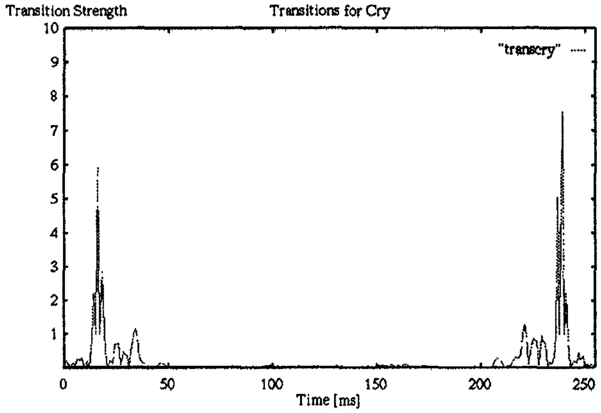

Figure 10: Freqency transition for cry

As the indicators are not equally important for the recognition process we also use different weights. These weights differ from event to event (see Table 3). Using them we are able to calculate a mean correlation between test pattern and stored pattern.

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

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 possibly contain violence.

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
Σ	100	100	100

Table 3: Weights of indicators

5 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 the detection of violence and the indexing and retrieval of music.

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

1. F. Arman, R. Depommier, A. Hsu, and M.-Y. Chiu. Content-based browsing of video sequences. In *Proceedings of Second ACM International Conference on*

Multimedia, pages 97–103, Anaheim, CA, October 1994.

2. Kurt Benedini. *Psychoacoustic Measurements 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, Technical University of Munich, 1978. (in German).

3. E. O. Brigham. *The Fast Fourier Transform.* Prentice-Hall Inc., 1974.

4. G.J. Brown and M. Cooke. Computational auditory scene analysis. *Computer Speech and Language*, (8):297–336, August 1994.

5. T.H. Bullock. *Recognition of complex acoustic signals.* Report of Dahlem Workshop on Recognition of Complex Acoustic Signals. Abakon Verlagsgesellschaft, Berlin, 1977.

6. Gordon E. Carlson. *Signal and Linear System Analysis.* Houghton Mifflin Company, Boston Toronto, 1992.

7. M.P. Cooke. *Modelling Auditory Processing and Organisation.* Cambridge University Press, 1993.

8. S. Fischer, R. Lienhart, and W. Effelsberg. Automatic recognition of film genres. In *Proceedings of Third ACM International Conference on Multimedia*, pages 295–304, Anaheim, CA, November 1995.

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

10. H. Fletcher and W. A. Munson. Loudness, its definition, measurement and calculation. *J. Acoustical Society of America*, 5(82), 1993.

11. Christoph Gerum. Automatic recognition of audiocuts. Master's thesis, University of Mannheim, Germany, January 1996. (in German).

12. 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.

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

14. Michael Köhlmann. *Rhythmic Segmentation of Sound Signals and their Application to the Analysis of Speech and Music.* PhD thesis, Technical University of Munich, 1984. (in German).

15. Christoph Kuhmünch. Automatic recognition of commercials on tv. Master's thesis, University of Mannheim, Germany, July 1996. (in German).

16. 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.

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

18. 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.

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

20. J. A. Molino. Pure-tone equal-loudness contours for standard tones of different frequencies. *Percept. Psychophys.*, 14(1), 1973.

21. Richard Parncutt. *Harmony: A Psychoacoustical Approach*, volume 19 of *Springer Series in Information Sciences*. Springer-Verlag, Berlin Heidelberg, 1989.

22. S. Pfeiffer, R. Lienhart, S. Fischer, and W. Effelsberg. Abstracting digital movies automatically. *Visual Communication and Image Representation*, to appear.

23. R. Plomp. Pitch of complex tones. *J. Acoustical Society of America*, 41(1526), 1967.

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

25. Robert Schulz. Automatic recognition of beat in music, 1995. University of Mannheim. (in German).

26. Klaus Schulze. *Contribution to the Problem of One-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, Dsseldorf, 1985. (in German).

27. Hermann Schütte. *Determination of the Subjective Event Times of Subsequent Sound Impulses via Psychoacoustic Measurements.* PhD thesis, Technical University of Munich, 1977. (in German).

28. Ulrich Sieben. *Binaural Signal Processing: Psychoacoustic Investigation of Central Excitation Patterns.* PhD thesis, Georg-August-University of Göttingen, 1985. (in German).

29. 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.

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

31. S. S. Stevens. Measurement of loudness. *J. Acoustical Society of America*, 27(815), 1955.

32. Christoph von Campenhausen. *The senses of man - an introduction to the psychophysics of perception.* Georg Thieme Verlag, Stuttgart, New York, 1993. (in German).

33. H.J. Zhang, Y. Gong, S.W. Smoliar, and S.Y. Tan. Automatic Parsing of News Video. In *Proceedings of IEEE Conf. on Multimedia Computing and Systems*. IEEE, May 1994.

34. H.J. Zhang, A. Kankanhalli, and S.W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1(1):10–28, January 1993.

35. H.J. 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.

36. E. Zwicker, G. Flottorp, and S. S. Stevens. Critical bandwidth in loudness summation. *J. Acoustical Society of America*, 29(548), 1957.

37. J. J. Zwislocki. Temporal summation of loudness: An analysis. *J.Acoustical Society of America*, 46(431), 1969.